Development of Computer-aided Concepts for the Optimization of Single-Molecules and their Integration for High-Throughput Screenings

Entwicklung Computergestützter Konzepte im Hinblick auf die Optimierung von Einzelmolekülen sowie deren Integration für Hochdurchsatzverfahren

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Dem Fachbereich Biologie der Technischen Universität Darmstadt zur Erlangung des akademischen Grades eines Doctor rerum naturalium vorgelegte Dissertation von

M.Sc. Sven Frederik Jager aus Berlin

Darmstadt 2019
D 17

1. Gutachten: Prof. Dr. Kay Hamacher
2. Gutachten: Prof. Dr. Beatrix Süß

Jager, Sven Frederik: *Development of Computer-aided Concepts for the Optimization of Single-Molecules and their integration*
for High-Throughput Screenings
Darmstadt,Technische Universität Darmstadt
Jahr der Veröffentlichung der Dissertation auf TUprints: 2019
Tag der mündlichen Prüfung: 13.03.2018
Veröffentlicht unter CC BY-SA 4.0 International:
`https://creativecommons.org/licenses/`

Contents

1 Zusammenfassung **5**

2 Abstract **9**

3 Introduction **13**

4 Structure and Dynamics of Biomolecules **21**
4.1 Important Biomolecules: Proteins and Nucleic Acids 21
4.2 On the Structure of RNA and Proteins 23
4.2.1 Protein Structure 23
4.2.2 Nucleic Acid Structure 24
4.2.3 Secondary Structure Prediction of RNA 27
4.3 Molecular Dynamics Simulations 31
4.3.1 An Introduction to Molecular Dynamics Simulations . . 31
4.3.2 Technical Introduction 32
4.4 Graph-based Analysis and Representation of Biomolecules . . . 37
4.4.1 Motifs, Graphs and Dynamic Graphs 37
4.4.2 Conversion of MD Simulations to Dynamic Graphs . . . 38
4.4.3 Comparison of Protein Structures 39
4.4.4 Advanced Representation and Comparison of RNA Secondary Structures 40

5 Protein Engineering **43**
5.1 Cleavage Product Accumulation Decreases the Activity of Cutinase 43
5.2 Statistical Evaluation of HTS Assays for Enzymatic Hydrolysis of β-Keto Esters 100

6 Graph-based Analysis of MD Simulations **121**
6.1 *StreaM* - a Stream-based Algorithm for Counting Motifs in Dynamic Graphs 121
6.2 Motif Based Analysis of MD Simulations 139
6.2.1 Results 144
6.2.2 Dynamics of Water in Confinement of Minerals 147
6.2.3 Structural Dynamics of a Protein Complex 155
6.3 StreAM-Tg: algorithms for analyzing coarse grained RNA dynamics 158
6.4 StreaMD: Advanced analysis of Molecular Dynamics using R . . 176

7 Graph-based Analysis of HTS-SELEX **199**
7.1 Riboswitching with ciprofloxacin 199
7.2 SICOR: Subgraph Isomorphism Comparison of RNA Secondary Structures 228
7.3 Motif Based Analysis of NGS Data 246

8 Contributions **251**
8.1 Puplications 251
8.2 Conference Poster 252
8.3 Invited Talks 253
8.4 Supervised Work 253

9 Summary and Discussion **255**

10 Appendix **III**
10.1 Material to Gromacs GPU Benchmarks III
10.2 Material to Graph-based Analysis of MD Simulations III
10.3 k-vertex Motif Visualization VII
10.4 Abbreviations XV
List of Tables XVII
List of Figures XIX

References for Introduction, Discussion and Summary **XXI**

Curriculum Vitae **XLI**

Ehrenwörtliche Erklärung **XLIII**

Danksagung **XLV**

1 Zusammenfassung

Im Fachgebiet der synthetischen Biologie haben sich in den letzten Jahrzehnten interdisziplinäre Herangehensweisen für das Design und die Modellierung funktioneller Moleküle durch computergestützte Methoden etabliert. Diese computergestützten Methoden finden vor allem Anwendung, wenn Experimentelle-Ansätze an ihre Grenzen stoßen, da Computermodelle in der Lage sind beispielsweise durch Einzelmolekül-Simulationen das zeitliche Verhalten von Nukleinsäurepolymeren oder Proteinen aufzuklären, sowie die funktionelle Beziehung der Aminosäurereste oder Nukleotide zueinander darzustellen. Das mittels Computermodellierung erhobene Wissen kann fortführend genutzt werden, um den weiteren experimentellen Verlauf (z.B. *Screening*), sowie die Gestalt beziehungsweise die Funktion (*Rational Design*) des betrachteten Moleküls zu beeinflussen.

Eine solche vom Menschen durchgeführte Optimierung der Biomoleküle ist oftmals notwendig, da die betrachteten Substrate für die Biokatalysatoren, beziehungsweise die Enzyme meist synthetisch sind („man-made materials“ wie z.B. PET) und die Evolution noch keine Zeit hatte, effiziente Biokatalysatoren dafür bereit zu stellen.

In Bezug auf das computerbasierte Design von Molekülen, teilen sich zwei fundamentale Paradigmen die Vorherrschaft im Fachgebiet der synthetischen Biologie. Die in dieser Arbeit gewonnenen wissenschaftlichen Erkenntnisse lassen sich dementsprechend in diese zwei Bereiche unterteilen. Auf der einen Seite finden probabilistische experimentelle Methoden (z.B. evolutionäre Designprozesse wie z.B. die gelenkte Evolution) in Kombination mit *Hochdurchsatz-Screenings* Anwendung, auf der anderen Seite werden meistens rationale, computergestützte Einzelmolekül-Designmethoden verwendet.

Für beide Themenbereiche wurden Computermodelle/Verfahren entwickelt, evaluiert und veröffentlicht.

Der erste Beitrag in dieser Arbeit beschreibt einen computergestützten Designansatz der *Fusarium Solanie Cutinase* (FsC). Hier wurde im Detail (molekular) der Aktivitätsverlust des Enzyms bei längerer Inkubationszeit mit PET untersucht. Dafür wurden Molekular Dynamik (MD) Simulationen von der räumlichen Struktur der FsC und einem wasserlöslichen Abbauprodukt des synthetischen Substrates PET (Ethylenglycol) berechnet. Das bestehende Modell wurde zusätzlich durch die Kombination mit Reduzierten-Modellen er-

weitert. Durch diese Simulations-Studie konnten bestimmte Bereiche der FsC identifiziert werden, welche sehr stark mit PET (Ethylenglycol) wechselwirken, und dadurch einen signifikanten Einfluss auf die Flexibilität sowie Struktur des Enzyms nehmen.

Die darauffolgende Originalpublikation etabliert ein neues Verfahren zur Auswahl von Hochdurchsatz-Assays für den Einsatz in der Proteinchemie. Die Auswahl geschieht über eine Meta-Optimierung, der zu analysierenden Assays. Hierfür werden Kontrollreaktionen für den jeweiligen Assay durchgeführt. Die Distanz der Kontrollverteilungen wird unter zu Hilfenahme von klassischen statistischen Methoden wie z.B. dem Kolmogorov-Smirnov-Test evaluiert. Anschließend wird jedem Assay eine Performance zugewiesen. Die beschriebenen Kontroll-Experimente werden vor dem eigentlichen Experiment (Screening) durchgeführt und der Assay mit der höchsten Performance wird für das weitere Screening genutzt. Durch Anwendung dieses generischen Verfahrens können hohe Erfolgsraten bei einem solchen Screenings erzielt werden. Dies konnten wir experimentell am Beispiel von Lipasen und Esterasen zeigen.

Im Rahmen der grünen Chemie helfen die oben genannten verfahren, Enzyme für den Abbau von z.B. synthetischen Materialien schneller zu finden oder natürlich vorkommende Enzyme dahingehend zu verändern, sodass diese Enzyme nach erfolgreicher Optimierung synthetische Substrate effizient umsetzten können. Hierfür wird bei der praktischen Durchführung der experimentelle Aufwand (Verbrauch an Materialien) möglichst geringgehalten. Insbesondere bei groß angelegten Screening kann eine vorherige Betrachtung oder Einschränkung des möglichen Lösungsraum (i.e. Sequenzraums) einen entscheidenden Beitrag liefern, die Erfolgsquote zu maximieren, sowie den gesamten Zeitaufwand des Screenings zu minimieren.

Neben der Durchführung klassischer Methoden wie MD Simulationen in Kombination mit Reduzierten-Modellen wurden auch neue Graphen-basierte Methoden für die Darstellung sowie Analyse von MD-Simulationen entwickelt. Hierfür wurden Simulationen in distanzabhängige dynamische Graphen konvertiert. Ausgehend von dieser reduzierten Darstellung wurden effiziente Algorithmen zur Analyse entwickelt und getestet. Dabei wurden insbesondere Netzwerk-Motive untersucht, um festzustellen, ob diese spezielle Art der Semantik geeignet ist molekulare Strukturen und Wechselwirkungen, innerhalb von MD Simulationen, besser zu beschreiben als räumliche Koordinaten. Dieses Konzept wurde für die verschiedensten MD Simulationen von Molekülen wie zum Beispiel Wasser, synthetische Poren, Proteine, Peptide sowie RNA Strukturen evaluiert. Es konnte gezeigt werden, dass sich diese neuartige Form der Semantik ausgezeichnet eignet, (bio)molekulare Strukturen sowie deren Dynamik zu beschreiben. Des Weiteren wurde ein Algorithmus (StreAM-T_g) für das Erstellen von Motiv basierten Markov-Modellen, speziell für die Analyse von Einzelmolekül-Simulationen von Nukleinsäuren, entwickelt. Dieser Algorithmus findet seinen Einsatz im RNA-Design. Die aus der Analyse mit StreAM-T_g gewonnenen Erkenntnisse (Markov-Modelle) können hilfreiche Vorschläge für das (Re)Design von funktioneller RNA liefern.

In diesem Zusammenhang wurde eine neue Methode entwickelt, um die Umgebung (i.e. Wasser; Lösungsmittel-kontext) und deren Einfluss auf verschiedene Moleküle in MD Simulationen zu quantifizieren. Hierfür wurden drei-Vertex-Motive verwendet, um die Struktur der einzelnen Wassermoleküle zu beschreiben. Diese neue Methode bietet viele Vorteile. Mittels dieser Methode kann die Struktur sowie die Dynamik von Wasser akkurat beschrieben werden. Beispielsweise konnten wir die thermodynamische Entropie von Wasser in der Flüssig- und Dampfphase entlang der Dampf-Flüssig-Gleichgewichtskurve vom Tripelpunkt bis zum kritischen Punkt reproduzieren.

Ein weiteres großes Themengebiet, welches im Rahmen dieser Arbeit behandelt wurde, ist die Entwicklung von neuen computergestützten Ansätzen für ein Hochdurchsatzverfahren für das Design von funktioneller RNA. Für die Herstellung von funktioneller RNA wird in der Regel ein experimentelles, runden-basiertes Hochdurchsatzverfahren (SELEX) verwendet. Durch Anwendung von *Next Generation Sequencing* (NGS) in der Kombination mit dem SELEX-Verfahren kann dieser Designprozess erstmals auf Nukleotidebene sowie auf Sekundärstrukturebene verstanden werden. Die Besonderheit bei kleinen RNA-Molekülen im Vergleich zu Proteinen ist, dass die Sekundärstruktur (Topologie), welche die minimale freie Energie aufweist, direkt aus der Nukleotidsequenz, mit hoher Sicherheit, ermittelt werden kann.

Somit gelang es mittels der Kombination von M. Zukers und P. Stieglers Algorithmus, NGS und dem SELEX-Verfahren die strukturelle Diversität einzelner RNA-Moleküle unter Berücksichtigung des genetischen Kontextes zu quantifizieren. Diese Kombination der Methoden, erlaubten die Rundenvorhersagen, in denen unteranderem der erste *Ciprofloxacin-Riboswitch* hervorging.

In diesem Beispiel wurde lediglich ein einfacher, struktureller Abgleich für die Quantifizierung (Levenshtein-Distanz; LD) der Diversität jeder einzelnen Runde vorgenommen. Um dies zu verbessern wurde eine neue Darstellung der RNA-Struktur als gerichteter Graph modelliert, welche anschließend mit einem probabilistischen Subgraph-Isomorphismus verglichen wurde.

Zuletzt wurde der NGS-Datensatz (*Ciprofloxacin-Riboswitch*) als dynamischer Graph modelliert und nach dem Auftreten definierter Sieben-Vertex-Motiven analysiert. Es wurde die motiv-basierte Semantik erstmals für die Anwendung in Hochdurchsatz-Screenings für RNA Moleküle integriert. Die dadurch identifizierten Motive konnten Sekundärstrukturelementen (RNA), die in R10k6 (*Ciprofloxacin-Aptamer*) experimentell bestimmt wurden, zugeordnet werden.

Abschließend wurden alle vorgestellten Algorithmen in einer `R` Bibliothek integriert, veröffentlicht und WissenschaftlerInnen aus der ganzen Welt zur Verfügung gestellt.

2 Abstract

In the field of synthetic biology, highly interdisciplinary approaches for the design and modelling of functional molecules using computer-assisted methods have become established in recent decades. These computer-assisted methods are mainly used when experimental approaches reach their limits, as computer models are able to e.g., elucidate the temporal behaviour of nucleic acid polymers or proteins by single-molecule simulations, as well as to illustrate the functional relationship of amino acid residues or nucleotides to each other. The knowledge raised by computer modelling can be used continuously to influence the further experimental process (screening), and also shape or function (*rational design*) of the considered molecule. Such an optimization of the biomolecules carried out by humans is often necessary, since the observed substrates for the biocatalysts and enzymes are usually synthetic ("man-made materials", such as PET) and the evolution had no time to provide efficient biocatalysts.

With regard to the computer-aided design of single-molecules, two fundamental paradigms share the supremacy in the field of synthetic biology. On the one hand, probabilistic experimental methods (e.g., evolutionary design processes such as directed evolution) are used in combination with *High-Throughput Screening* (HTS), on the other hand, rational, computer-aided single-molecule design methods are applied. For both topics, computer models/concepts were developed, evaluated and published.

The first contribution in this thesis describes a computer-aided design approach of the *Fusarium Solanie Cutinase* (FsC). The activity loss of the enzyme during a longer incubation period was investigated in detail (molecular) with PET. For this purpose, *Molecular Dynamics* (MD) simulations of the spatial structure of FsC and a water-soluble degradation product of the synthetic substrate PET (ethylene glycol) were computed. The existing model was extended by combining it with Reduced Models. This simulation study has identified certain areas of FsC which interact very strongly with PET (ethylene glycol) and thus have a significant influence on the flexibility and structure of the enzyme.

The subsequent original publication establishes a new method for the selection of *High-Throughput assays* for the use in protein chemistry. The selection is made via a meta-optimization of the assays to be analyzed. For this purpose, control reactions are carried out for the respective assay. The distance of the control distributions is evaluated using classical static methods such as the Kolmogorov-Smirnov test. A performance is then assigned to each assay. The described control experiments are performed before the actual experiment (screening), and the assay with the highest performance is used for further screening. By applying this generic method, high success rates can be achieved. We were able to demonstrate this experimentally using lipases and esterases as an example.

In the area of green chemistry, the above-mentioned processes can be useful for finding enzymes for the degradation of synthetic materials more quickly or modifying enzymes that occur naturally in such a way that these enzymes can efficiently convert synthetic substrates after successful optimization. For this purpose, the experimental effort (consumption of materials) is kept to a minimum during the practical implementation. Especially for large-scale screenings, a prior consideration or restriction of the possible sequence-space can contribute significantly to maximizing the success rate of screenings and minimizing the total time they require.

In addition to classical methods such as MD simulations in combination with reduced models, new graph-based methods for the presentation and analysis of MD simulations have been developed. For this purpose, simulations were converted into distance-dependent dynamic graphs. Based on this reduced representation, efficient algorithms for analysis were developed and tested. In particular, network motifs were investigated to determine whether this type of semantics is more suitable for describing molecular structures and interactions within MD simulations than spatial coordinates. This concept was evaluated for various MD simulations of molecules, such as water, synthetic pores, proteins, peptides and RNA structures. It has been shown that this novel form of semantics is an excellent way to describe (bio)molecular structures and their dynamics. Furthermore, an algorithm (StreAM-T_g) has been developed for the creation of motif-based Markov models, especially for the analysis of single molecule simulations of nucleic acids. This algorithm is used for the design of RNAs. The insights obtained from the analysis with StreAM-T_g (Markov models) can provide useful design recommendations for the (re)design of functional RNA.

In this context, a new method was developed to quantify the environment (i.e. water; solvent context) and its influence on biomolecules in MD simulations. For this purpose, three vertex motifs were used to describe the structure of the individual water molecules. This new method offers many advantages. With this method, the structure and dynamics of water can be accurately described. For example, we were able to reproduce the thermodynamic entropy of water in the liquid and vapor phase along the vapor-liquid equilibrium curve from the triple point to the critical point.

Another major field covered in this thesis is the development of new computer-aided approaches for HTS for the design of functional RNA. For the production of functional RNA (e.g., aptamers and riboswitches), an experimental, round-based HTS (like SELEX) is typically used. By using *Next Generation Sequencing* (NGS) in combination with the SELEX process, this design process can be studied at the nucleotide and secondary structure levels for the first time. The special feature of small RNA molecules compared to proteins is that the secondary structure (topology), with a minimum free energy, can be determined directly from the nucleotide sequence, with a high degree of certainty.

Using the combination of M. Zuker's algorithm, NGS and the SELEX method, it was possible to quantify the structural diversity of individual RNA molecules under consideration of the genetic context. This combination of methods allowed the prediction of rounds in which the first *ciprofloxacin-riboswitch* emerged.

In this example, only a simple structural comparison was made for the quantification (Levenshtein distance) of the diversity of each round. To improve this, a new representation of the RNA structure as a directed graph was modeled, which was then compared with a probabilistic subgraph isomorphism.

Finally, the NGS dataset (*ciprofloxacin-riboswitch*) was modeled as a dynamic graph and analyzed after the occurrence of defined seven-vertex motifs. For this purpose, motif-based semantics were integrated into HTS for RNA molecules for the first time. The identified motifs could be assigned to secondary structural elements that were identified experimentally in the ciprofloxacin aptamer R10k6.

Finally, all the algorithms presented were integrated into an `R` library, published and made available to scientists from all over the world.

3 Introduction

> *It's okay not to know all the answers. It's better to admit our ignorance than to believe answers that might be wrong. Pretending to know everything, closes the door to finding out what's really there.*
>
> –Neil deGrasse Tyson

The design of functional molecules by computational means have become a major paradigm to support synthetic biology; still the methodological approaches are anything but complete (*1*). The quest to close these gaps in available computational approaches is the guiding principle of this thesis.

Functional optimization (e.g., increase substrate turnover, thermo-stability, binding small molecules), as well as the *de novo* (re-)design of biomolecules are dominated by two main paradigms: on the one hand, probabilistic methods (e.g., evolutionary design processes or "Irrational Design" (*2*)) in combination with *High-Throughput Screening* (HTS), and on the other hand, computer-aided, *rational in silico* molecular design methods are used (*3*). However, boundaries between these two principles become increasingly blurred. For structure prediction, such combination of the two algorithmic approaches are already successfully applied (*4–8*), while parameters typical are derived from experimental data (*9–11*). In turn, computer-aided modeling is then used to optimize directed evolution. In HTS all data (such as sequences, readouts, parameters) can be used for training of *Machine Learning Models* or *Markov State Models* (MSM) (*12–14*). Experimental parameters are increasingly being extended by simulation results on the molecular level, such as in *Molecular Dynamics* (MD) (*15*). Furthermore, *Next-Generation Sequencing* (NGS) can support this endeavour (*16*). These predictive models can in concert evaluate existing designs (*12*, *17*).

This work is placed on the interface of the two design paradigms described above. In the following, the ideas, their implementation and validation, as well as their application are laid out in detail.

High-Throughput Screenings and directed Evolution

HTS is the process of testing a large library of molecules for a desired purpose to identify 'Hits' that fulfill certain characteristics, e.g., for industrial bio-catalytic processes (*18*). Biomolecular libraries can, e.g., contain purified enzymes, microorganisms from the environment or protein variants from directed evolution or randomization at the gene level. Here, the limiting factor is the fast and reliable identification of the best suited molecule for the given purpose. HTS typically enables simultaneous screenings of samples in 96- to 1536-well plates so that 10^5 to 10^7 samples can be screened per day (*19–21*) – even more so on fully automatized robotic platforms (*22*).

An unique and efficient type of directed evolution in combination with HTS is the *Systematic Evolution of Ligands by exponential enrichment* (SELEX) (*23*). This method was developed to create *de novo* RNA or DNA devices (aptamers) and in some rare cases Riboswitches. To discover the latter, it is practical to apply a system that involves selection, library screening and in addition rational design to yield the desired structures (*24*). SELEX leads to an enrichment of aptamers, starting from a synthetic, combinatorial library of up to 10^{16} individual sequences. SELEX has made it possible to create *de novo* RNA devices capable of recognizing almost any ligand of choice.

Due to this progress, aptamers were found and engineered for a high affinity against different small organic molecules like Tetracycline, Neomycin and Isoleucin, etc. (*25–28*). Still, the discovery of small molecule-binding RNA aptamers with high affinity binding and specificity remains challenging (*29*). In particular, there is the difficulty that the aptamers found should also work *in vivo*. So far, *in vivo* active structures rely only on a small set of ligands (*30*). The combination of directed evolution and NGS makes it possible to extract and mine all randomized sequences from experiments (*12*, *16*). A curse and a blessing, however, are the amounts of data generated by such experiments. Additionally, it is difficult to identify models and data structures that are best suited for this task.

Despite the power and success of HTS, it comes along with many disadvantages. First, biological libraries cannot contain every possible variant compound, simply due to the combinatorial complexity of the solution-space. Second, the choice of the perfect assay in regards to finding the (bio)molecule with the desired function is crucial for the success of HTS. For example, if a substrate needs to be modified for the screening process, the difference between the properties of the original substrate to the substrate used for screening can lead to false positives (*31–34*). At last, HTS is preparative expensive and requires highly skilled experimental researchers.

Computational and Rational Design

The computational ("rational") design paradigm is the biochemical equivalent to computer-aided design concepts in modern mechanical engineering. While HTS methods allowed for great breakthroughs over the last years, a rational design success is still rare in the field (*35*).

Frequently, mainly numerical simulation such as MD methods are used to understand the structure as well as the dynamics of the system under scrutiny (*3*, *36–39*). In order to perform simulations, like in mechanical engineering (e.g., Finite-Elements (*40*) simulations), structural knowledge in the form of three-dimensional coordinates are required to model the system. Unfortunately, experiments typically reveal for proteins and RNAs only static structures in most cases as stored in the *Protein Data Bank* (PDB) (*41*) with some 150 000 entries. In addition, few NMR structures are deposited there – frequently, however, only giving some dozen of snapshots of the dynamics which is typically to few data to understand the full dynamics. Alternatively, structure prediction directly from the sequence could be a feasible alternative. The structure prediction problem for proteins is still not fully solved. The success of modeling via homology or by evolutionary constrains strongly depends on the available templates (*42*, *43*) and includes three-body contacts as well (*44*), while *ab inito* from physical principles alone gained traction in recent years (*45*, *46*). While, here, the computational burden is many times greater than the amount used for homology modeling (*47*, *48*). In contrast to proteins, nucleic acids have a much smaller variety of building blocks (e.g., nucleotides). Additionally, RNA has a modular structure and the nucleotides have more explicit interaction rules like base pairing (*49*). This chemical simplicity ultimately suggests a structural simplicity. However, this is clearly not the case in nature.

Single-strand RNA show a remarkable spectrum of structural diversity, while established algorithms for accurate structure prediction (*50*) were early on reported. For example, the secondary structure of RNA – namely, the base pairing – can be modeled with a high accuracy in shorter time directly from the sequence. Among the first approaches is Zuker's algorithm (*51*, *52*). However, for the determination of dynamic properties of RNA an advanced simulation approach is still required to gain detailed knowledge about the structure and function of the respective domains (*53*).

While all these simulations and computations (*6*, *54*, *55*) produce data for rational molecular design (*56*), the data analysis afterwards requires extensive knowledge of biophysics, computer science, and structural biology is also required to perform and analyses the results of MD simulations.

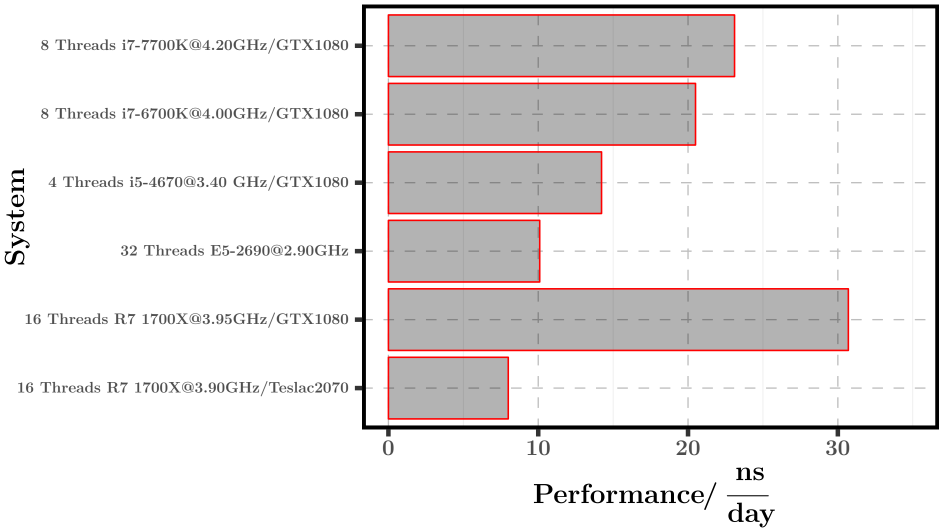


Figure 3.1: Benchmarks of an MD simulation for a large system using standard desktop computers with GPU and *Xeon* boards. Plotted is the simulation performance for the different models. The system simulated was the always the same

A major contribution the wide-spread use of MD is its acceleration by *Graphics Processing Units* (GPU) and moreover, algorithmic improvements (*57–60*). Take, for example, the *Gromacs* project, which involves new algorithms targeting SIMD/streaming architectures as well as new parallelization schemes for inhomogeneous hardware of both CPUs and GPUs (*57*). Thus, we can assess larger biomolecular complexes like channel proteins in cell membranes or micelles on an atomistic scale (*4*, *5*, *61*). In Figure 3.1 I illustrate the achieved performance of MD simulations. Clearly, new low-cost consumer GPU cards can outperform well designed cards for HPC[1] just one generation from the past. As the molecular processes have time scales ranging from femtoseconds to hours, we sometimes might still encounter a "lack of resources" situation. In this theses, however, short to medium timed simulations turned out to be sufficiently long enough to estimate essential dynamics of a system (*6*, *62*).

For predicting movements on larger time scales reduced *Coarse-Grained* (CG) models can be used (*6*). The trade-off in these models is that accuracy and resolution are exchanged to assess longer time scales. CG models combine groups of atoms to form pseudo atoms or beads (*63–65*). The connections of these beads reflect the underlying molecular interactions (*66*). By such a representation the model naturally loses resolution, but also degrees of freedom and thus computational complexity. In combination with simple potentials, even larg assembly processes such as the one of the ribosome can be understood (*67*).

The major challenge in order to use CG models, is to find an adequate representation of the interactions between these particles, which reflects the underlying physical principles. The modeling to static interaction graphs offers an efficient approach to solving this problem, since half a century of established methods, discrete algorithms and fast heuristics can be used. Especially for the modeling of static RNA structures a large number of graph-based representations have been successfully applied (*6*). However, when it comes to integrating dynamic aspects into graph modeling or using dynamic graphs as a data structure, there exist very few examples so far. Reduced models are much faster to compute and offer the possibility of a time-resolved behavior (e.g., Martini or RedMD) in addition to the classical multi-scale analysis using simple harmonic potentials (*68*, *69*). Furthermore, the quality of CG models can be improved if the harmonic potential is parameterized with the help of nuclear resonance spectroscopy (e.g., sdENM) or MD simulations (e.g., Reach) (*62*, *70*). Thus CG has proven to be a powerful tool to probe the spatial and temporal evolution of systems on the micro scale, beyond what is feasible with traditional *all-atom* (AA) models. Considering this technical

---

[1] =High-Performance-Computing

and methodological innovation, simulations and reduced models will become an integral part of synthetic biology in the future.

Readers Digest

Chapter 4 will give a theoretical introduction of (bio)molecules and the structures they adopt. Thereby, three key concepts, namely Molecular Dynamics, structure prediction of RNA and graph-based representations of molecules are introduced.

Section 4.2 introduces basic concepts of nucleic acids and proteins structures followed by a short motivation of Zuker's algorithm for the RNA structure prediction. Section 4.3 gives a brief introduction to MD simulations in combination with technical aspects. Afterwards, Section 4.4 gives an introduction as well as a introduction to graph-based representations of molecules. The Chapter presents two manuscripts Chapter 5 presents two manuscripts of modern computer-assisted protein engineering approaches for HTS as well as single molecule simulations. The first manuscript shows an approach were classical MD simulations were used to address rational design opportunities of the *Fusarium Solani Cutinase* (FsC). The second paper deals with the development of a pre-screening procedure for suitable Esterase/Cutinase Assays for HTS.

Chapter 6 deals with manuscripts and research regarding the concept of representing MD trajectories as dynamic graphs and its promising application to RNA, proteins, protein complexes, water and confined minerals. As a result of this, a new motif-based semantic is introduced. Here, Section 6.1 depicts a 4-vertex motif-based approach which aims to describe secondary structure dynamics of proteins. Section 6.2 generalizes this concept regarding vertex size and extended the application to a wide range of molecules (e.g., on a *toy model*, mineral confinement) as well as thermodynamics of water. As an application scenario for this manuscript, the motif-based semantics concept is also transferred to RNA as well. Here, a novel concept and algorithm for motif-based representations to derive RNA based MSM's is introduced and applied in Section 6.3. At last, in Section 6.4, all the above introduced and motivated algorithms were combined in a software paper.

Chapter 7 depicts theory and applications of advanced statistics and graph theory to improve screening procedures for HTS-SELEX. Section two displays a successful combination of SELEX with NGS and structure prediction that supported the discovery of the first *cirpofloxacin* (CFX) riboswitch.

Section 7.2 describes a novel (sub)graph-based approach for RNA structure comparison with regard to the application in HTS-SELEX. At last, Section 7.3 illustrates a novel methodology of motif-based semantics to RNA structures in order to improve selection for the NGS data set of the experiments. At last, the results above will be summarized and discussed in Chapter 9.

4 Structure and Dynamics of Biomolecules

This chapter gives a brief conceptual introduction into proteins, RNA, and DNA. Section 4.2 explains the structural composition of proteins and RNA. In addition, the secondary structure prediction of RNA by Zuker's algorithm is described in detail. Section 4.3 gives a short introduction of the methodology of MD simulations followed by technical details. At last, an introduction to graph-based analysis and molecular representation can be found in Section 4.4.

4.1 Important Biomolecules: Proteins and Nucleic Acids

Biomolecules are the foundation of life because they are responsible for almost every biochemical function in our body as well as everywhere in nature. Among (bio)molecules proteins and nucleic acids play the key role. Nucleic acids like *Deoxyribonucleic* (DNA) and *Ribonucleic acid* (RNA) act as the blueprints of proteins.

DNA is processed to RNA by a DNA-dependent RNA-Polymerase and afterwards translated into amino acid polymers (proteins) by the Ribosome (*71*). The genetic code is generated by the sequence of bases in the DNA. Here, 64 base triplets coding for 20 proteinogenic amino acids. The genetic code thus looks degenerated at first glance because the 64 possible codons eventually code for less information than would maximally be possible. Proteins are involved in a large number of regulatory processes in the cell. Enzymes are proteins that are capable of catalyzing chemical reactions. In this way, enzymes also work highly specifically in the aqueous phase at room temperature.

The chemical industry relies on enzymatic processes in the sense of sustainable chemistry and the recycling industry (e.g., in Germany) (*72*). It is desirable that enzymatic reactions completely replace chemical processes. Particularly due to the lower energy consumption as well as high yields and selectivity involved, there are many advantages in comparison with classic chemical reactions (*72*). For this reason, proteins are of great interest to the chemical industry. In addition to the economic benefits just mentioned, proteins are macro molecules that perform and control various metabolic functions within every cell. Their biological functions include catalysis (i.e. enzymes), muscle contraction (e.g., titin), the transport of ions (e.g., hemoglobin), the transmission of information between specific cells and organs (e.g., hormones), activities in the immune system (e.g., antibodies), the passage of molecules (e.g., ions) across cell membranes, etc. (*71*). For a long time, it was assumed that the only function of RNA was passing on the genetic information stored in the DNA so that it can be translated into proteins. In this picture, the RNA, therefore, represents only one step on the way from DNA to protein. The assumption that this is the only task of RNA was refuted in the mid-1980s when *Thomas Cech* found out that RNA molecules can also act as enzymes and form stable spatial structures, just like proteins (*73*). Some of these structures have impressive properties such as the highly specific recognition and binding of low-molecular chemicals. These so-called aptamers are short, single-stranded DNA or RNA oligonucleotides that can bind a specific molecule via structural motifs. Aptamers or their binding sensory motifs are often found in riboswitches or ribozymes. The latter molecules can additionally change their conformation by binding a ligand or even catalyzing chemical reactions such as proteins(*74*). The fact that RNA is not only an information carrier but also one of the few molecules that can catalyze chemical reactions even led to the idea of an RNA-world hypothesis (*73*). In this hypothesis, it is assumed to be that the origin of life is in self-replicating RNA molecules, from which molecular Darwinian evolution, complex, and more complex systems have been evolved. Due to the possibility of producing RNA *de novo* as well as a large number of accurate algorithms for structural prediction, RNA became molecules with a high potential regarding industrial applications. During the past years, customized gene network design has become of interest among various disciplines in life science (*13*, *75*). Here, RNA/DNA aptamers serve as highly specific detectors in sensors or therapeutic drugs and are used for bio-computing devices (*13*, *24*, *53*). Proteins and RNA hold a broad range of applications in our society, and the potential is far from being exhausted.

4.2 On the Structure of RNA and Proteins

4.2.1 Protein Structure

Proteins [1] are chemical polymers consisting of 20 different types of amino acids (*76*). For each of these 20 amino acids, several triple sets of nucleic acid bases are encoded on DNA level (i.e. codons). Proteins occur in a wide variety of different forms and structural levels in nature. The computation of the most favorable molecular energy conformation is not feasible due to the many possibilities.

Hence, the structure prediction of proteins faces two major problems. One of them is a prediction of a native conformation for a given sequence of amino acids. This is referred to as the protein folding. The latter one is an inverse folding problem, where a target conformation is given, and one has to find which sequence(s) would fold into this particular conformation.

In biochemistry, four hierarchically arranged structural levels are differentiated among proteins (*71*). First the Primary structure - the amino acid sequence (sequence of amino acids) of the peptide chain. Secondary structure - the spatial structure of a local area in the protein (e. g. helix, loops, sheet).

The next level is the tertiary structure, the spatial (3D) structure of a subunit. Figure 4.1 illustrates elements of the secondary structure such as a helix or beta sheet and allow the tertiary structure to be displayed.

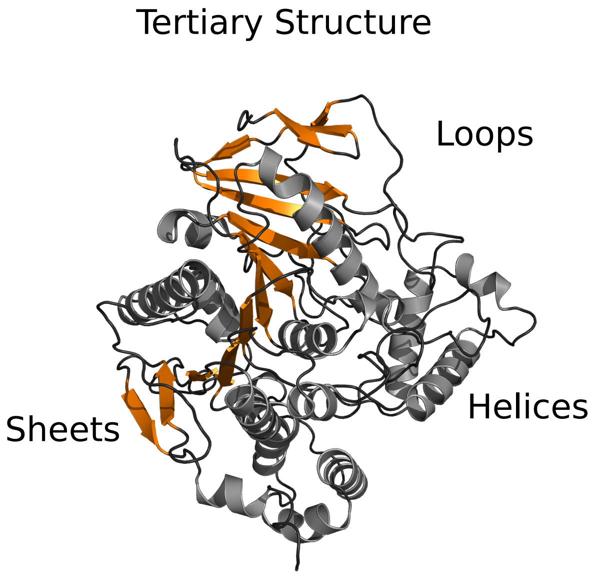


Figure 4.1: Proposed tertiary structure of the pNB-Est13 represented in `pymol` (*77*)

The structural elements mentioned stabilize the structure due to intramolecular interactions and unique spatial arrangements of the participating amino acids. At last, the Quaternary structure - the spatial structure of

---

[1] Small Proteins (molecular weight < 10 000 Da) are referred as peptides

the entire protein complex with all subunits. Secondary structural elements can be determined by advanced experimental spectroscopic methods such as *circular dichroism spectroscopy* (CD) (*71*). On the opposite, 3D structures can only be solved experimentally by NMR or X-ray crystallography (*71*). Both techniques are experimentally complex and limited by the size of proteins. In addition to the two established methods single-particle electron cryomicroscopy methods like *CryoEM* are also used (*78*). Unfortunately, the resolution range with 0.6-2.4 nm of these methods is still too low to serve as a template for modelling (*78*).

The prediction of spatial protein structures yields good results if proteins with similar sequence and known structure already exist (*79*).

This enables homology modeling, whereby the new sequence is mapped to the target sequence whose structure is known, and thus "fitted" into the structure (*80*). This technique is similar to sequence alignment. The prediction is more difficult if no structures of similar amino acid sequences are known.

4.2.2 Nucleic Acid Structure

In nature, DNA is almost always present in the form of double-stranded DNA, whereas the majority of RNAs are present as single-stranded DNA. Hybrid double-stranded molecules are only intermediate products in transcription.
RNA is a single-stranded, long-chain nucleic acid molecule. It is composed of four building blocks, the nucleotides. Each of these nucleotides consists of a ribose molecule (i. e. a sugar with $5'$ carbon atoms), a phosphate residue and one of the four organic bases *adenine* (A), *guanine*(G), *cytosine* (C) and *uracil* (U) (DNA: *thymidin* (T)). An example of this is given in Figure 4.2. The phosphate residue serves as a linker between the sugar molecules of two susceptible nucleotides. It combines the 3' carbon atom of the ribose molecule of a nucleotide with the 5' carbon atom of the ribose molecule of the other nucleotide. This basic structure is also known as the backbone of RNA.

The nucleotides of the double-stranded DNA or single-stranded RNA can form stable spatial structures through base pairing. This kind of interaction is defined as the specific bond between guanine and cytosine or adenine and thymine (DNA) or adenine and uracil (RNA) fixed by hydrogen bonds.

The particularbase pairing (Watson-Crick binding) is a necessary prerequisite for the formation of the double helix structure of DNA from two comple-

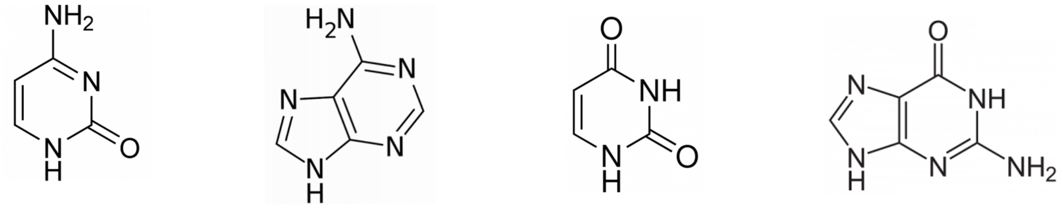


Figure 4.2: Chemical structures of the four nucleotides of DNA/RNA. The bases of *cytosine* (C), *adenine* (A), *guanine*(G) *uracil* (U) (from left to right). *thymidin* (T) has methyl group at the marked position of (U).

mentary single strands. However, there are also exceptions such as the $A-T$ pairing, in which the N_1 atom of the purine ring system does not act as a H-bond acceptor in the sense of the Watson-Crick geometry, but rather the N_7 atom performs this function. This structural type of interaction is also called *Hoogsteen* base pairing and is more stable from a chemical point of view than a Watson-Crick pairing (*81*). However, the latter occur much more frequently in nature. However, if an interaction scheme deviates from Watson-Crick, it is often referred to as *wobble* pairing (cf. Figure 4.3) (*82*).

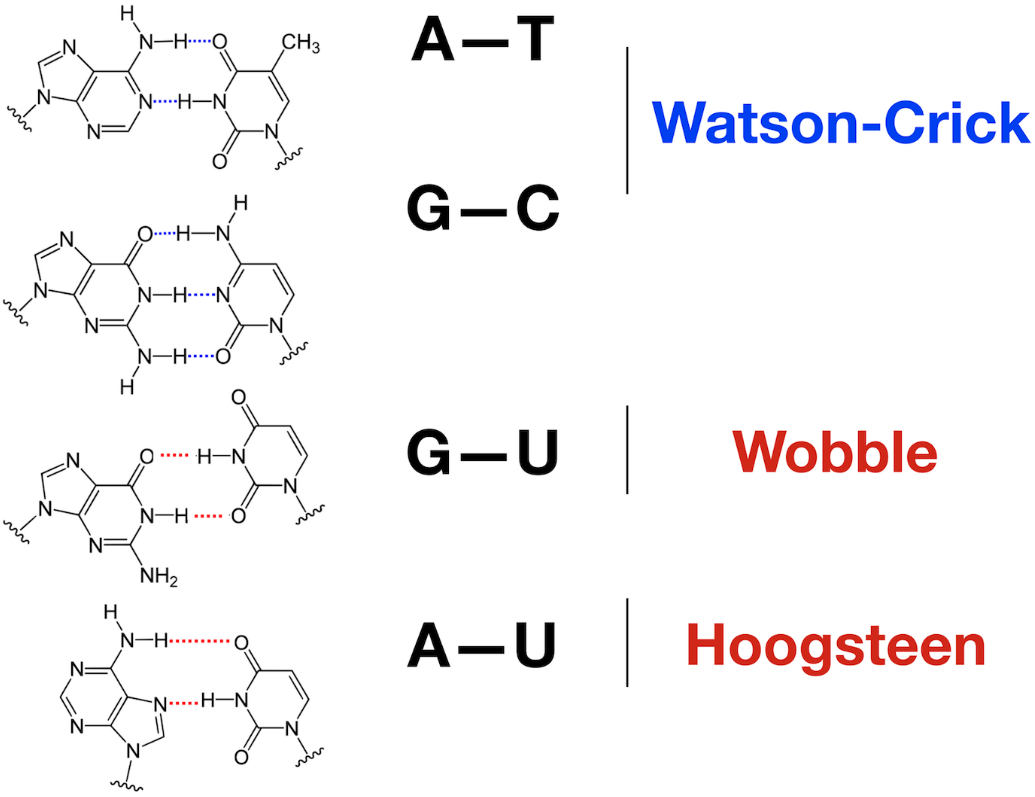


Figure 4.3: Chemical structures of different interaction schemes.

Aptamers and Riboswitches and SELEX

Aptamers (*aptus*, fit and large *meros*, region) are single-stranded DNA or RNA molecules that, due to their three-dimensional structure and the ability to form stable interactions, such as hydrogen bonds, bind target molecules with high affinity and specificity. Riboswitches often have an aptamer or its binding motif as the main component (*83*). Aptamers can be used, similar to monoclonal antibodies, in various fields, from diagnostics to affinity chromatography up to therapeutic applications (*23*). These molecules can be discovered by an iterative version of directed evolution (e.g., SELEX) (*26*, *28*, *84*).

During this process, a library of randomized RNA molecules is combined with a target structure. The target structure is usually immobilized on a matrix from which the unbound RNA species can be washed off after incubation. Species that bind the target structure remain in the matrix. As a next step, non-binding RNA is eluted and binding RNA molecules are enriched and amplified in the later process. The last step is the most challenging one.

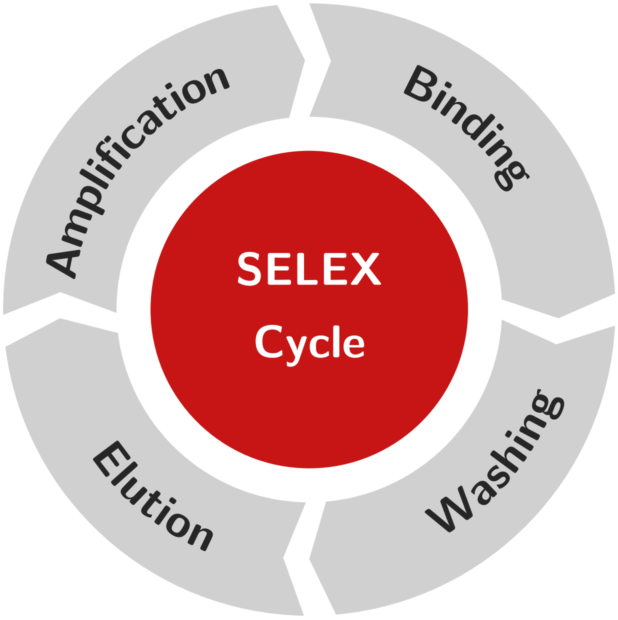


Figure 4.4: Schematic representation of the in vitro selection, SELEX

After eluting, this fraction is converted into the RNA pool of the next selection round by reverse transcription and *polymerase chain reaction* (PCR). It is obvious that in the course of a multi cyclic SELEX experiment the evolution pressure shifts from the side of high affinity sequences to sequences which are just better amplifying. The most potent sequences are not necessarily those with high frequencies. Underrepresented sequences might be the aptamers with the most preferred properties.

There exists many different strategies for performing the selection (*30*, *85–87*). In addition, there are also many methods that are advanced like Counter SELEX. Compared to traditional SELEX, counter SELEX adds an additional step using structurally-similar targets to incubate with aptamers to effectively discriminate non-specific oligonucleotides (*88*). Furthermore there are SELEX protocols for *in vivo* and *in vitro* screening. For *Synthetic Biology*, it is especially important, that newly developed aptamers function also in experiment, *in vivo* (*29*, *88*). This is not always the case due to cellular context dependencies. Nowadays, the SELEX process is still a "black box" at the molecular level. Nevertheless, the application of NGS offers a possibility to reveal some of its mysteries (*16*).

To conclude, the SELEX process is time consuming, and the success rates remain low and most of the current aptamers are obtained *in vitro*, and whether they can function *in vivo* needs to be elucidated – up to now – experimentally (*88*).

4.2.3 Secondary Structure Prediction of RNA

Derivation of the RNA secondary structure supports assessing its function. This renders secondary structure prediction from RNA-sequences as one of the important problems in the filed of Bioinformatics (*89*). To address this problem, various solutions have been proposed so far. These solutions include a variety of theoretical concepts such as e.g.: machine learning (*90*), dynamic programming (*51*), base pair maximization (*91*), genetic algorithms (*92*), calculation of the partition function with *minimum free energy* (MFE) (*93*), context free grammar (*51*), algebraic dynamic programming (*94*), evolutionary constrains (*95*) etc.

Zuker and Stiegler Algorithm

One major milestone in the filed of Bioinformatics was the RNA secondary structure prediction algorithm by Michael Zuker and Peter Stiegler (*51*). The following part explains Zuker's example of structure prediction using dynamic programing to compute the MFE [2].

---

[2] based on Hofacker *et al.* (*96*)

Technical Details The idea behind this algorithm is that each RNA structure with n nucleotides can only be modeled in two different ways from shorter structures:

1. The first nucleotide i is unpaired with a second nucleotide k
2. The first nucleotide i is paired with a second nucleotide k

   with $k - 2 > i$.

However, this is only valid under the assumption that the nucleotides form independent secondary structures (e.g., base pairs does not cross). If the pairs or arcs are crossed in this representation, one speaks of a Pseudo-node (e.g.,*pseudoknot*) in the structure. The term pseudoknots thus covers many essential interactions of nucleic acids. They occur rarely, but are important for the spatial configuration of the RNA. For this reason, the knowledge about pseudoknots is essential for predicting the structure of an RNA molecule in space (*97*). However, these prediction models are not accurate and the underlying models are far too complex [3]. A well-known example is the algorithm of Rivas and Eddy with a run-time complexity of $\mathcal{O}(n^6)$ and a memory requirement of $\mathcal{O}(n^4)$ (*99*). The algorithm of Zuker, on the other hand, has a run-time complexity of $\mathcal{O}(n^3)$. This is still very expensive, but acceptable for the integration in high throughput applications.

In this model, the free energy of each substructure $\boldsymbol{F}_{i,j}$ (cmp. Equation (4.1)) between base i and j is composed of the states when i and j are not paired or paired. The recursions for computing the MFE of an RNA molecule in the loop-based energy model can be summarized as follows:

$$\boldsymbol{F}_{i,j} = \min\Big\{\boldsymbol{F}_{i+1,j},\ \min_{i<k\leq j}\{\boldsymbol{C}_{i,k} + \boldsymbol{F}_{k+1,j}\}\Big\} \tag{4.1}$$

$\boldsymbol{F}_{i,j}$, denotes for the free energy of the optimal substructure on the sub sequence from i to j.

$$\boldsymbol{C}_{i,j} = \min\Big\{\boldsymbol{H}_{i,j},\ \min_{i<k<j<l}\{\boldsymbol{C}_{k,l} + \boldsymbol{I}_{ij,kl}\},\ \min_{i<m<j}\{\boldsymbol{M}_{i+1,m} + \boldsymbol{D}_{m+1,j-1}\}\Big\} \tag{4.2}$$

$\boldsymbol{C}_{i,j}$, is the free energy of the optimal substructure on the subsequence subject to the constraint that i and j form a base pair. $\boldsymbol{C}$ consists of three different

---

[3] Lyngsø and Pedersen *et al.* proven that the problem is NP-complete (*98*)

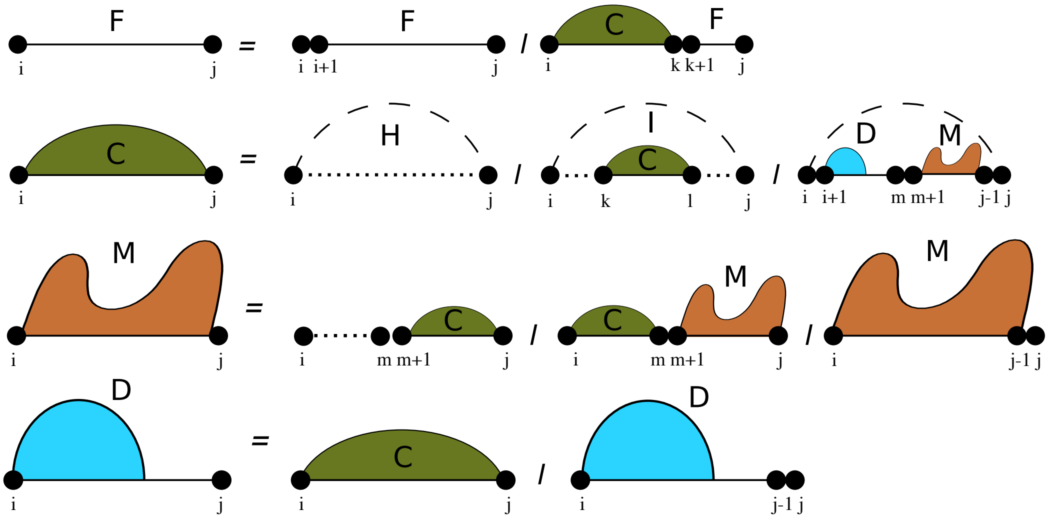


Figure 4.5: Decomposition of RNA secondary structure used for the structure prediction. Here, Feynman Diagrams of the recursion grammar are used for further visualization. Unpaired nucleotides are represented by either an individual unconnected dot or a dashed line. Figure is adapted from Hofacker *et al.* (*96*)

types of complex structures. $\boldsymbol{H}$ describes a hairpin and $\boldsymbol{I}$ an induced complex structure $\boldsymbol{I}$ or a complex structure consisting of a multi-loop $\boldsymbol{M}$ or one with only one component $\boldsymbol{D}$ (cf. Equation (4.2)).

$$\boldsymbol{M}_{i,j} = \min\left\{ \min_{i<m<j}\{(m-i+1)+\boldsymbol{C}_{m+1,j}\},\ \min_{i<m<j}\{\boldsymbol{M}_{i,m}+\boldsymbol{C}_{m+1,j}\},\ \boldsymbol{M}_{i,j-1}\right\} \tag{4.3}$$

$\boldsymbol{M}_{i,j}$ denotes for the free energy of the optimal substructure on the subsequence subject to the constraint that that the structure is part of a multi-loop and has at least one component. In the case of a multi-loop, the calculation of energy becomes somewhat more difficult because the energy depends on the number of substructures it consists of (cf. Equation (4.4)). For this reason, the structure is decomposed and the individual components are listed.

$$\boldsymbol{D}_{i,j} = \min\left\{\boldsymbol{D}_{i,j-1},\ \boldsymbol{C}_{i,j}\right\} \tag{4.4}$$

Here $\boldsymbol{D}_{i,j}$ depicts the free energy of the optimal substructure on the subsequence i, j subject to the constraint that that i, j is part of a multi-loop and has exactly one component, which has the closing pair i, m for some m satisfying $i < m < j$.

For the calculation of MFE or ΔG values for a given (sub)structure, the Zuker algorithm uses experimental enthalpy values for every nearest neighbor combination. These parameter sets exist for DNA and RNA and can be found in the *Nearest Neighbor Data Bank* (NNDB) (*100–102*). These parameters also allow the computation of thermodynamic parameters such as the entropy S or melting temperature T_m of a basepair stack directly from the sequence (*103*).

4.3 Molecular Dynamics Simulations

This section deals with the concept of Molecular Dynamics (MD) simulations and its application followed by some technical details. The first part will give a brief introduction of common simulation methods and the second part technical insights about MD.

4.3.1 An Introduction to Molecular Dynamics Simulations

Single molecule simulation techniques (*75*) can help to understand macroscopic molecular observable's (obtained by an assay, spectroscopy or spectrometry) with microscopic insights into the time dependent behavior of a single molecule.

MD simulations yield in a trajectory of every single atom in the system, and thus allow for very detailed analyses of dynamics and structure. Hence MD is just one out of many simulation methods, it is probably the most frequently used beside Monte Carlo (MC) simulations for molecules in general (*104*). MC Simulations are often applied for the 3D structure prediction of RNA (*105*, *106*), peptides or small proteins (*107*, *108*). This simulation method is based on sampling multiple configurations (random trial steps) in combination with an physical energy/scoring function. One advantage of random sampling of molecular configurations is that MC simulations does not require a continuous energy function likewise in MD (*109*).

MD simulations are an extremely powerful tool when dealing with systems where quantum effects can be well parameterized. There also exist many hybrids methods combining some of the simulation aspects (e.g., QM-MM, Quantum Mechanics in combination with Molecular Mechanics) (*110*).

MD is based on the theory that the *de Broglie* wavelength is very small compared to next-neighbour lengths and it is assumed electrons adjust instantaneously to the motion of nuclear cores (*112–114*). Hence, a major drawback of MD is that chemical reactions or electron tunneling can not be simulated until now. The main advantage of MD is to provide a trajectory of every atom, enabling comparison with the results from experiments like NMR. However, the limit for comparing experiments and MD simulations is the time scale and the size of the given system. Furthermore, the considered systems are very small compared to the number of particles in a typical biochemical asssay. One can imagine that the small system size evokes boundary effects,

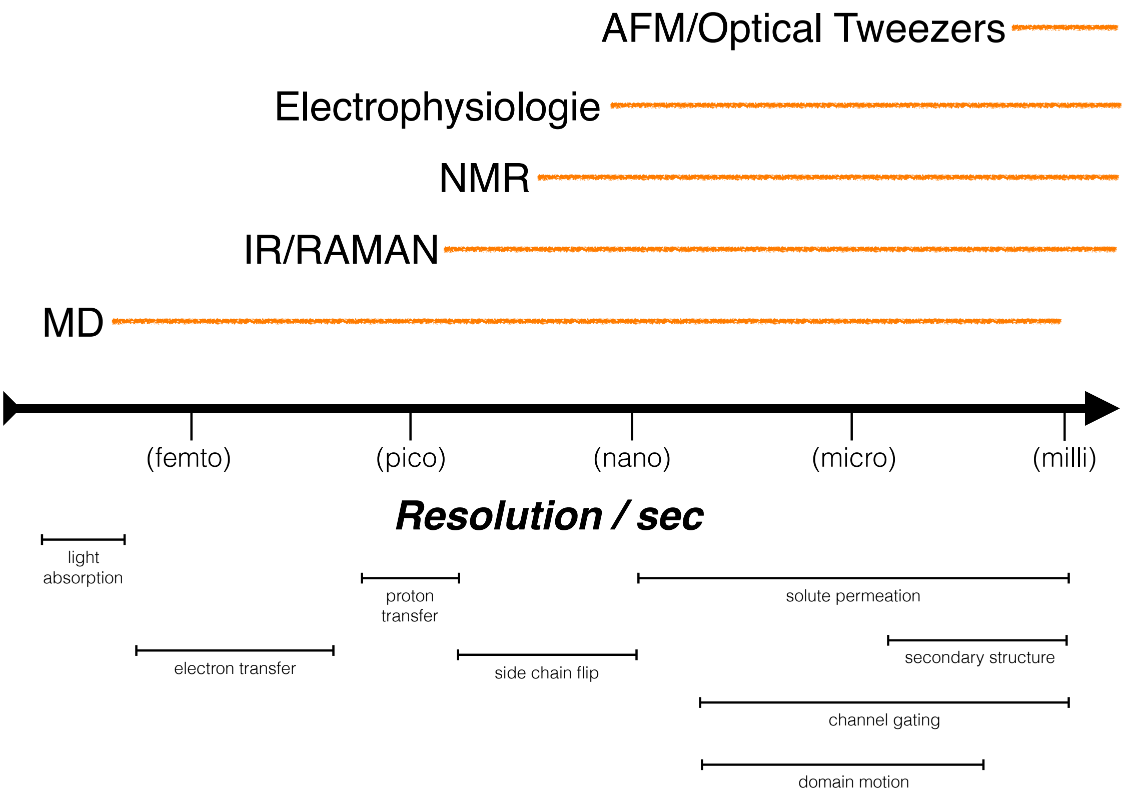


Figure 4.6: Different time-scales for simulations and the according experimental methods. Biomolecular movement is annotated below the black arrow. Figure is adapted from (*111*)

but this problem is solved by applying *periodic boundary conditions* (PBC). Figure 4.6 shows the time scales for selected experiments and for MD simulations. In summary, MD simulations are a powerful tool as they provide access to an amazing variety of dynamical and structural analyses and thus help to bridge the gap between macroscopic observables and microscopic assumptions.

4.3.2 Technical Introduction

MD is one of the frequently used methods applied for single molecule modeling in this work. Several MD simulation suites are available, where *Gromacs* (*115–117*) is one of the most commonly used. The following part gives a short technical introduction as well as a brief description of the methodology and the underlying physics behind MD [4]. In general, MD simulations solve Newton's equations of motion, describing the forces on the i-th out of N particles with m_i as the atomic mass and and H_{total} as the applied potential resulting from the applied *Force Field* (FF) (e.g., potential). ,

$$\vec{F}_i(t) = m_i \frac{d^2\vec{r}_i}{dt^2} = -\frac{\partial H_{total}\{\vec{r}_1, ..., \vec{r}_N\}}{\partial \vec{r}_i}, \tag{4.5}$$

MD simulations aim at solving the latter equation with an numerical approach due to the interaction of too many particles for an analytic treatment. Therefore, the chosen integrator solves Equation (4.5) using discrete time steps Δt for integration. Usually the time-step is in a range from 0.5 to 2 femto seconds. Notice that the potential function H_{total} considers the positions of atomic nuclei only. Thus, electrons are assumed to be in the ground state and [5] instantaneously adjust for center of mass motions of the atom. The potential consists of two parts, the bonded (Equation (4.10)) and the non-bonded (Equation (4.7)) interactions.

$$H_{total} = H_{bonded} + H_{nonbonded} \tag{4.6}$$

The Lennard-Jones potential is one method for describing non-bonded Long-range interactions and the Coulomb for non bonded short range interactions, and reads as follow,

$$H_{nonbonded}\{\vec{r}_1, ..., \vec{r}_N\} = \sum_{j=1}^{N-1} \sum_{i=j+1}^{N} \left\{ \epsilon_{i,j} \left[\left(\frac{C_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{C_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\} \tag{4.7}$$

with the coefficients C_{ij}^{12} and C_{ij}^{6}, determined when developing the FF. This description is only valid if particles of the same atom type interact with each other. The coefficients are modified when calculating the interactions between different atom types (α and β), using the following mixing rules [6]

$$C_{ij,\alpha\beta}^{6} = \left(C_{ii,\alpha}^{6} C_{jj,\beta}^{6} \right)^{1/2} \tag{4.8}$$

$$C_{ij,\alpha\beta}^{12} = \left(C_{ii,\alpha}^{12} C_{jj,\beta}^{12} \right)^{1/2}. \tag{4.9}$$

---

[4] based on the *Gromacs* manual (*114*)

[5] according to the Born-Oppenheimer approximation

[6] *Gromacs* supports further mixing rules or it is possible to specify the parameters directly by way of a matrix.

where α and β are used as indicator functions for the atom type of atoms i and j, respectively.

Long-range electrostatic interactions are calculated with the help of the Ewald summation algorithm. This algorithm is also called Ewald summation and rewrites the interaction potential as a sum of a short range and long range interaction term. This unique feature makes Ewald summation very efficient (e.g., run-time complexity $\mathcal{O}(N\log(N))$). An harmonic oscillator with quantum corrections is used in order to describe bond-angle and bond-stretching motions (see Equation (4.7)). Here, K_r describes spring constant in order to describe bonds and K_θ to describe bond angles (with angle θ). Here, r depicts the distance between atom i and j. Further, constrains are used to handle motion of bonded particles. LINCS is usually used for solving this kind of constraint molecular dynamics (e.g., LINear Constraint Solver; LINCS algorithm) (*118*).

$$\begin{aligned} H_{bonded}\{\vec{r}_1, ..., \vec{r}_N\} = & \sum_{bonds} K_r(r - r_{eq})^2 \\ & + \sum_{angle} K_\theta(\theta - \theta_{eq})^2 \\ & + \sum_{dihedral} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] \end{aligned} \tag{4.10}$$

The standard integration scheme of *Gromacs* is the *leap-frog* algorithm. This algorithm evaluates positions and velocities at different times and has a accuracy up to the third-order in the position for an expansion in a Taylor series. Here, positions are calculated at time t, whereas velocities are evaluated at time $t - \frac{1}{2}\Delta t$, reading

$$\vec{v}(t + \frac{1}{2}\Delta t) = \vec{v}(t - \frac{1}{2}\Delta t) + \frac{\Delta t}{m}\vec{F}(t) \tag{4.11}$$

and

$$\vec{r}(t + \Delta t) = \vec{r}(t) + \Delta t\vec{v}(t + \frac{1}{2}\Delta t). \tag{4.12}$$

One advantage of MD is to provide the opportunity to use different thermodynamic ensembles. In all performed simulations, only two ensembles have been used for this work, isobaric-isothermal (NpT, constant number of particles, pressure, and temperature) and Canonical (NVT, constant number of particles, volume, and temperature). In this work the NpT ensemble is used for equilibration and the NVT ensemble for production runs.

MD simulations temperature is adjusted with an external heat bath. One frequently used example of an heat bath is the *Nosé-Hoover* temperature coupling algorithm (*119*). This algorithm is frequently used in MD in order to study protein as well as water dynamics. *Nosé-Hoover* modifies the equation of motion given by Equation (4.5) by adding a heat-bath term, reading

$$\frac{d^2\vec{r}_i}{dt^2} = \frac{\vec{F}_i}{m_i} - \frac{p_\xi}{Q}\frac{d\vec{r}_i}{dt} \tag{4.13}$$

and

$$\frac{dp_\xi}{dt} = (T - T_0). \tag{4.14}$$

The coupling strength Q[7] and the heat-bath parameter determined by the equation of motion depends on the oscillation period τ_T between the system and the heat bath. Accordingly, the current simulation temperature of the system is given by T whereas T_0 denotes for the desired temperature. Hence, these equations allow the temperature to fluctuate while on average T_0 is obtained.

The drawback of the *Nosé-Hoover* scheme is that phase space is only partially sampled even for infinitely long times. Extensions, which attempt to solve this problem by coupling many heat baths, could not be used, as these so-called *Nosé-Hoover* chains are not supported by the integration algorithm.

Constant-pressure simulations is done by using a barostat (commonly *Parrinello-Rahman*) (*120–122*). This temperature coupling scheme resembles the *Nosé-Hoover* temperature coupling and adds an additional term to Equation (4.13) with the *Nosé-Hoover* contribution included and reads as follows

$$\frac{d^2\vec{r}_i}{dt^2} = \frac{\vec{F}_i}{m_i} - \frac{p_\xi}{Q}\frac{d\vec{r}_i}{dt} - \mathbf{M}\frac{d\vec{r}_i}{dt}, \tag{4.15}$$

The scheme is similar to the previously introduced *Nosé-Hoover* approach, but much more complex due to several matrix operations. *Parrinello-Rahman* takes a single $\beta = 4.5 \cdot 10^{-5}$ 1/bar in case of isotropic pressure coupling, and the pressure time constant τ_p in ps (likewise τ_T in Equation (4.13)) value and physical meaning to τ_T for the temperature coupling. The main advantage of this kind of pressure coupling is that it is flexible and allows for slightly

---

[7] with $Q = \frac{\tau_T^2 T_0}{4\pi^2}$

changing the shape of the box during simulations.

The output of an MD simulation is an *MD trajectory.* It is commonly represented as a list of frames $X = (\vec{x}_{t_0}, \vec{x}_{t_1}, \ldots)$ that describe the snapshot of the simulated system at consecutive points in time $t_0, t_1, \ldots$, e.g., every pico second. Each frame $\vec{x}_t$ contains the three-dimensional, spatial coordinates $\vec{r}_t(i), i \in [1, n]$ of the simulated system of n representative atoms at the corresponding time t, i.e., $\vec{x}_t = (\vec{r}_t(1), \vec{r}_t(2), \ldots, \vec{r}_t(n))$.

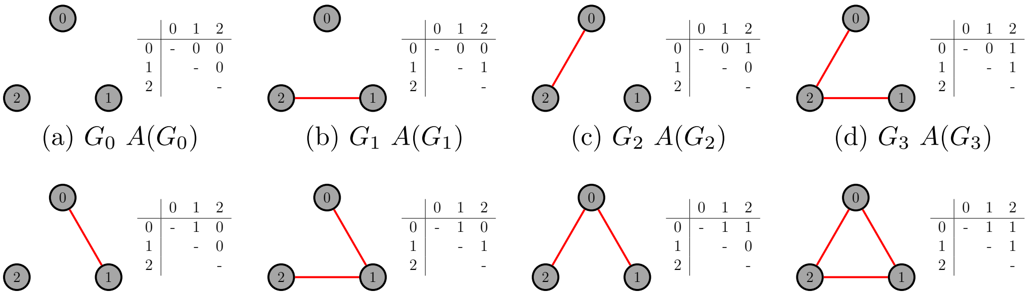
4.4 Graph-based Analysis and Representation of Biomolecules

Graph Theory is a field of discrete mathematics that deals with combinatorial structured of nodes and edges. Extensive application of graphs can be found in electrical and mechanical engineering, transportation and communication systems (*123*). In biology, graphs are used to model biological networks such as protein-protein, DNA regulation and cellular networks (*124*, *125*). However, graphs can also be used to model biomolecular structures (*126*). This methodology is mainly been applied in various contexts to study RNA structure and function (*127*). While the analysis of static graphs is important, time evolving networks recently gained a lot of attention. The following sections will give a short background to graph-based representations of proteins and RNA as well as corresponding notations.

4.4.1 Motifs, Graphs and Dynamic Graphs

A *graph* $G = (V, E)$ is an ordered pair, consisting of a set of *edges* E and a set of *vertices* $V = \{v_1, v_2, \ldots v_{|V|}\}$. In this thesis, we only consider *undirected graphs without self-loops*, i.e., $E \subseteq \{\{v, w\} : v, w \in V, v \neq w\}$. For a graph G, its *adjacency matrix* $\boldsymbol{A}(G)$ is a $|V| \times |V|$ matrix. A graph is called *connected* in case any two vertices $v, w \in V$ are connected. Accordingly ,a *dynamic graph* $G_t(V_t, E_t)$ can be defined as a list of graphs (e.g., $G_{t0}, G_{t1}, G_{t2}, \ldots$). Two graphs $G = (V, E)$ and $G' = (V', E')$ are called *isomorphic* if they contain the same number of vertices, i.e., $|V| = |V'|$, and there exists a so-called edge-preserving bijection $f : V \to V'$ such that $\{v, w\} \in V \iff \{f(v), f(w)\} \in V'$. Hence, graphs are considered to be isomorphic if they express the same topology. A *motif* is defined as the equivalence classes of isomorphic, connected k-vertex graphs. The *motifs of size k* is also called or *k-motifs* is defined as $\mathcal{M}_k = \{m_1, m_2, \ldots\}$. Accordingly, the *set of all k adjacency matrices* is defined as $\mathcal{A}_k$ with a size of $|\mathcal{A}_k| = 2^{\frac{k \cdot (k-1)}{2}}$. An example of adjacency matrices and graph topology is given for $\mathcal{A}_3 = \{\boldsymbol{A}^0, \boldsymbol{A}^1, \ldots \boldsymbol{A}^7\}$, the set of all eight adjacency matrices of 3-vertex graphs, in Figure 4.7.

While there exist $|\mathcal{A}_3| = 8$ different adjacency matrices, only four of them are connected (cf. Figure 4.7). All three connected adjacency matrices with two edges are isomorphic to each other and can be represented by a motif



(a) G_0 $A(G_0)$ (b) G_1 $A(G_1)$ (c) G_2 $A(G_2)$ (d) G_3 $A(G_3)$

(e) G_4 $A(G_4)$ (f) G_5 $A(G_5)$ (g) G_6 $A(G_6)$ (h) G_7 $A(G_7)$

Figure 4.7: $\mathcal{A}_3$ - all adjacency matrices of 3-vertex graphs ($|\mathcal{A}_3| = 2^{\frac{3\cdot(3-1)}{2}} = 8$). Figure is adapted from Schiller 2016 (*128*)

(equivalence class) m_1, i.e., $\boldsymbol{A}^3 \sim \boldsymbol{A}^5 \sim \boldsymbol{A}^6 \sim m_1$. Here $\boldsymbol{A}^7$, the only adjacency matrix with three edges, forms the second motif m_2, i.e., $\boldsymbol{A}^7 \sim m_2$. Hence, there exist $|\mathcal{A}_3| = 2$ different 3-vertex motifs. We describe the counts of the k-vertex motifs $m \in \mathcal{M}_k$ in a snapshot G_t as a function $F_{\mathcal{M}_k}(m) : \mathcal{M}_k \to \mathbb{N}$, where $F_{\mathcal{M}_k}(m)$ is the number of (k-vertex subgraphs of G that are isomorphic to m). These motif counts in dynamic graphs were efficiently computed using the StreaM$_k$ algorithm. StreaM$_k$ is defined and also motivated in Section 6.1 and details regarding the number of different k-vertex counts can be found in Table 10.2. A $\mathcal{Y}$ is defined as a SELEX set. This set consists of $|\mathcal{Y}|$ different sequences $\mathcal{Y} = \{S^0, S^1, ..., S^{|\mathcal{Y}|}\}$. Figure 4.7 is created with Benjamin Schillers `LaTeX-Graphs` [8].

4.4.2 Conversion of MD Simulations to Dynamic Graphs

Schiller *et al.* showed that MD simulations can be represented as a dynamic graph that specify the connections between the vertices at consecutive points in time (*126*, *129*). The graph-based representations of protein structures have proven to be very useful in computational biology (*130–133*). A point in time can be defined as a *frame*, which contains the positions of all n simulated atoms $\vec{x}_t$. Therein, each representative atom $i \in [1, n]$ is represented as a vertex $v_i \in V_t$.

These components are assumed to interact with each other if their L_2 norm is below a spatial distance cutoff d. In that case, an undirected edge is created between the corresponding vertices $\{v_i, v_j\}$. Using this unit-sphere

---

[8] `https://github.com/BenjaminSchiller/LaTeX-Graphs`

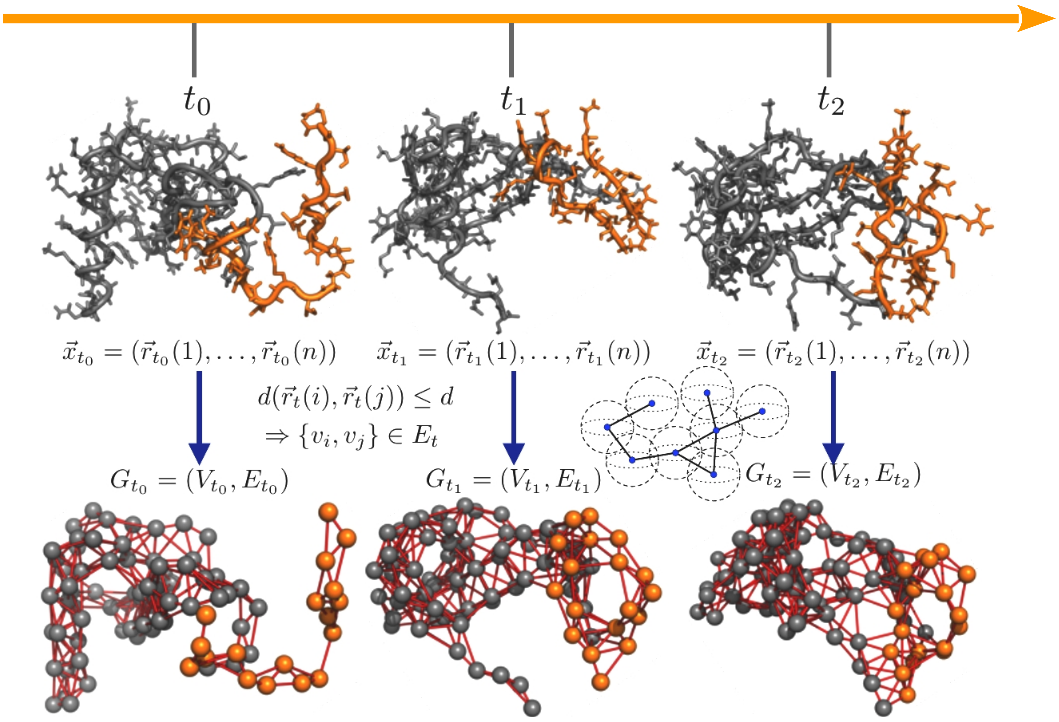

t_0
t_1
t_2
$\vec{x}_{t_0} = (\vec{r}_{t_0}(1), \ldots, \vec{r}_{t_0}(n))$
$\vec{x}_{t_1} = (\vec{r}_{t_1}(1), \ldots, \vec{r}_{t_1}(n))$
$\vec{x}_{t_2} = (\vec{r}_{t_2}(1), \ldots, \vec{r}_{t_2}(n))$
$d(\vec{r}_t(i), \vec{r}_t(j)) \leq d$
$\Rightarrow \{v_i, v_j\} \in E_t$
$G_{t_0} = (V_{t_0}, E_{t_0})$
$G_{t_1} = (V_{t_1}, E_{t_1})$
$G_{t_2} = (V_{t_2}, E_{t_2})$

Figure 4.8: Representation as a dynamic graph. The graphic shows a small synthetic peptide in complex with its receptor. The first row shows the side chain representations whereas the second row the dynamic graph representation. The unit sphere graph is obtained with a distance threshold of $d = 0.8nm$. Figure is created together with Benjamin Schiller.

model, we can model each frame $\vec{x}_t,\ t \in T$ as a unit-sphere graph G_t for a given distance threshold d. Thereby, we obtain a dynamic graph $G_{t_0}, G_{t_1}, \ldots$ that describes the interaction of the simulated atoms over time. An example of a graph obtained using the unit sphere model is given in Figure 4.8.

4.4.3 Comparison of Protein Structures

Structure comparison is also an important aspect regarding the analysis of protein simulations. Hence, this knowledge serves as a classifier in order to identify different configurations during the course of a simulation. Commonly, the configuration of a set of atoms is expressed using the *Root-Mean-Square Deviation* (RMSD, e.g., Equation (4.16)) (*134–136*). It expresses the Euclidean distance (L_2 norm) of each atom's position $\vec{r}_t(i)$ in a frame $\vec{x}_t$ to its initial position $\vec{r}_{t_0}(i)$ or an time-averaged position $\bar{\vec{r}}(i)$. Thus, the RMSD of an MD trajectory X at a point in time t is defined as follows:

$$RMSD(X,t) = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} d(\vec{r}_t(i), \vec{r}_{t_0}(i))^2} \tag{4.16}$$

Here, $d(\ldots)$ is a distance function using the L_2 norm. The RMSD is also frequently used to classify transition states of a protein and is extensively applied for the creation of MSM form simulations (*14*, *137*).

Unfortunately, it has been shown that RMSD is not even accurate enough for an intuitive determination of proteins equilibrium configuration (in MD simulations) (*138*).

4.4.4 Advanced Representation and Comparison of RNA Secondary Structures

RNAs highly modular and hierarchical structure makes it perfect for advanced representations as undirected, directed or tree graphs (*1*, *139*). Graph-based concepts (e.g., frequent subgraph mining) are promising and highly relevant in order to identify RNA patterns from NGS experiments (*74*, *140–142*). The first approaches of connecting graph theory dated back in the early 1970 by Waterman, Nussinov and Shapiro *et al.*, followed by the first comparison schemes for RNA tree graphs (*143*, *144*). Over time, various additional graph-based representations have been proposed so far. However, these are almost always reduced models (CG). These models include all tree representations except for the *full tree* representation from Lorenz and Hofacker (*145*, *146*). Yet, this does not represent the entire RNA structure, as nucleotides are also combined here. For example, one internal node of the tree corresponds to a base pair, a leaf node corresponds to a single unpaired nucleotide, and the root node is a virtual parent to the external structure elements (*96*).

A more condensed representation of the secondary structure is proposed by Fontana *et al.*, the *homomorphically irreducible tree* (HIT) representation (*147*, *148*). Some of the algorithms even combine structural elements or motifs. Shapiro as well as Schlicks *coarse grained tree* representations use multi-loops and other complex secondary structures as leafs and internal nodes (*143*, *144*). The frequently used way of representing RNA secondary structure is the *dot-bracket* (DB), where matching brackets symbolize base pairs and unpaired bases are represented as dots. Hence, this encoding converts an RNA structure

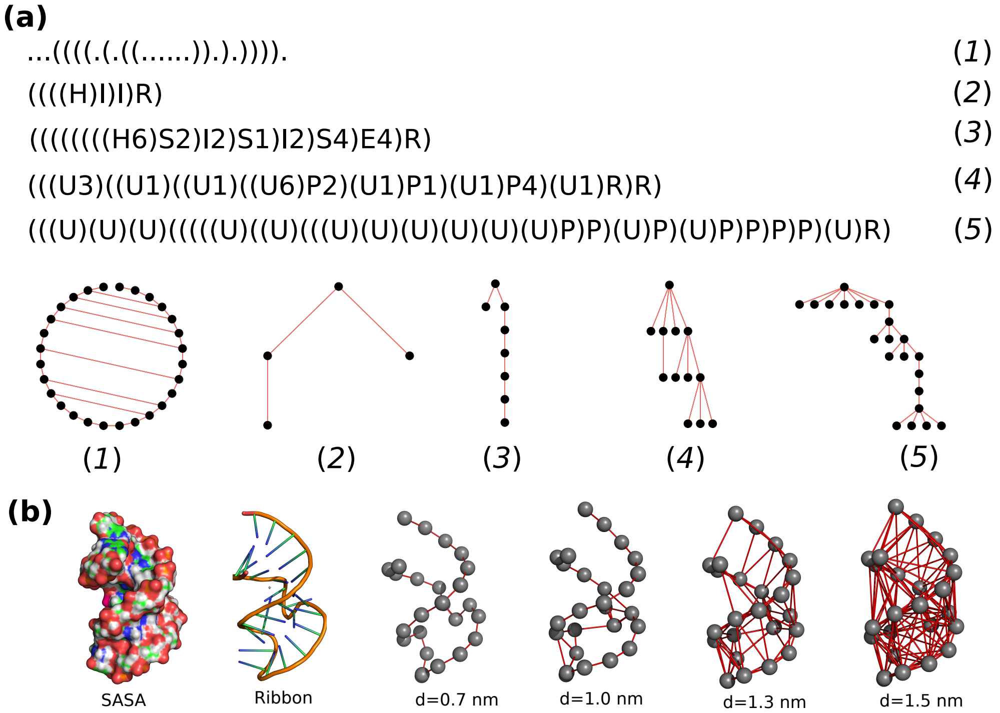

(a)
...((((.(.((......)).).)))).
(1)
((((H)I)I)R)
(2)
((((((((H6)S2)I2)S1)I2)S4)E4)R)
(3)
(((U3)((U1)((U1)((U6)P2)(U1)P1)(U1)P4)(U1)R)R)
(4)
(((U)(U)(U)(((((U)((U)(((U)(U)(U)(U)(U)(U)P)P)(U)P)(U)P)P)P)P)(U)R)
(5)
(1)
(2)
(3)
(4)
(5)
(b)
SASA
Ribbon
d=0.7 nm
d=1.0 nm
d=1.3 nm
d=1.5 nm

Figure 4.9: Different types of RNA structures representations. Graph representations as strings and illustrated using Reingold and Tilfords layout algorithm (*149*). **(a)**(1) Hofackers Dot Bracket and undirected Graph representation. **(a)**(2) Schlicks coarse grained tree representation (*1*, *139*). **(a)**(3) Shapiros coarse grained tree representation (*143*, *144*). **(a)**(4) Fontanas HIT trees (*147*, *148*). **(a)**(5) Hofackers and Lorenz full tree representation (*145*, *146*). **(b)** 3D NMR structure of *2KXM* (*27*). The first illustration represents the molecular surface and the second image the *ribbon* representation. From the third molecule on, the representation of a distance-dependent graph, which is derived from the RNA, begins. For this modeling different d are used and in addition the structure is reduced to its $C3'$ atoms. Connections are shown in red and $C3'$ atoms in gray.

to a vector $\vec{S}$ consisting of three characters ($\vec{S} \in \{., (,)\}$). Due to its simplicity, the DB encoding is one of the mostly used schemes for the representation of RNA molecules (*96*, *145*, *146*, *150*). DB can be converted to an adjacency matrix $\boldsymbol{A}$ by adding trivial edges (i.e. backbone contacts and nearest neighbor) and matching brackets as edges (hydrogen bonds). Figure 4.9 shows all the different representations of a single RNA molecule introduced in the previous paragraph.

Here, the structure of one of the smallest synthetic aptamers ever discovered was used as an example (e.g., *2KXM* (*27*); cf. Figure 4.9 **(b)**. Accordingly, Figure 4.9 **(a)** depicts the graphs in their nested string representation. In the second part, the topology of the considered graphs is additionally drawn. Figure 4.9 **(b)** depicts the unit-sphere CG scheme form Jager *et al.*. An additional very interesting concept offers the transformation of RNA structures into dual graphs from Schlick *et al.* (*89*, *151*). This particular representation is mainly used for motif mining (*142*).

5 Protein Engineering

5.1 Cleavage Product Accumulation Decreases the Activity of Cutinase during PET Hydrolysis.

Fusarium Solanie Cutinase (FsC) loses activity during prolonged incubation with PET. The reason for this behavior is widely unknown. Therefore, computational methods were carried out to describe the interactions between degradation product and the molecular mechanics of the enzyme. The following paper:

- Gross, C*., Hamacher, K., Schmitz, K., & **Jager, S.** (2017) Cleavage Product Accumulation Decreases the Activity of Cutinase during PET Hydrolysis. Journal of Chemical Information and Modeling, 57(2),243-255.

describes the use of MD simulations in combination with Reduced Models (i.e. CG), to show accumulations (PET degradation products and water) and their molecular impact on the enzyme. The simulations in combination with CG thus formed a good model, which explained at the molecular level why the loss of activity occurred. After the evaluation of the simulations and the reduced models, design proposals for the FsC were derived. In the course of this, a software package was also created. This package includes an extension of the method: Linear Response Theory by Ikeguchi *et al.* (*152*).

Contributions In order to find the reason for decreasing activity of FsC during the process of PET degradation, the initial concept of studying the influences of degradation products on the enzyme activity was given by me.

Together with Christine Groß, the concept of this study was further specified and reasonable evaluation methods were defined. Within the context of this work, I have performed and evaluated eight MD simulations. For the evaluation of these simulation trajectories, I used a variety methods (e.g. DSSP, RMSD, MSD, RMSF, SDC, Tetrahedral Order Parameter etc.). Some of the evaluation methods were known methods (e.g. RMSD, RMSF) but I also derived own solutions, which include the SDC and accumulation calculations. This method follows the accumulation of degradation products on the molecular surface (e.g., Protein or RNA SASA) during an MD simulation. In the course of this manuscript, I helped Christine Groß to publish the used scripts of the LRT null model as an R package: `R` library `LRTNullModel`. I was also responsible for creating Figures 2,3,5 and 6, and I created Figures 1 and 4 together with Christine Groß. For the Supporting Information I created Figures: S4, S5, S6, S7 and S8. Furthermore, I helped to motivate the paper and wrote parts of the manuscript. In this article, I am senior author. Kay Hamacher and Katja Schmitz helped to write the manuscript and improved it.

Cleavage Product Accumulation Decreases the Activity of Cutinase during PET Hydrolysis

Christine Groß,*,† Kay Hamacher,† Katja Schmitz,‡ and Sven Jager*,†

†*Department of Biology, Computational Biology & Simulation Group, Technische Universität Darmstadt, Schnittspahnstraße 2, 64287 Darmstadt, Germany*

‡*Department of Chemistry, Biological Chemistry Group, Technische Universität Darmstadt, Alarich-Weiss-Straße 8, 64287 Darmstadt, Germany*

E-mail: c.gross@bio.tu-darmstadt.de; jager@bio.tu-darmstadt.de

Phone: +49 6151 16 20372. Fax: +49 6151 16 72772

Abstract

The *Fusarium solani* cutinase (*Fs*C) is a promising candidate for the enzymatic degradation of the synthetic polyester polyethylene terephthalate (PET), but still suffers from a lack of activity. Using atomic MD simulations with different concentrations of cleavage product ethylene glycol (EG), we show influences of EG on the dynamic of *Fs*C. We observed accumulation of EG in the active site region reducing the local flexibility of *Fs*C. Furthermore, we used a coarse-grained mechanical model to investigate whether substrate binding in the active site causes an induced fit. We observed this supposed induced fit or "breath-like" movement during substrate binding indicating that the active site has to be flexible for substrate conversion. This guides rational design: mutants with an increased flexibility near the active site should be considered to compensate the solvent-mediated reduction in activity.

Introduction

To reduce the worldwide increasing environmental pollution by plastic waste, new methods to convert polymers back into monomers are needed. For the degradation of synthetic polymers, like polyethylene terephthalate (PET) or polyamide (PA), enzymatic degradation is to be favored over chemical or mechanical methods that suffer from the use of environmentally harmful chemicals or high energy costs.[1] In this context, the use of hydrolytic enzymes called cutinases, which are secreted by plant pathogenic fungi or bacteria, is a quite promising approach.[2–5] Due to their ability to degrade the natural high-molecular weight polyester cutin, the main component of the plant cuticle, some cutinases are also able to degrade synthetic polyesters.[2,5] Cutinases have been subject to numerous activity and mutation studies regarding the degradation of several synthetic polymers.[2–5]

PET, the synthetic polymer most commonly used worldwide, is a main target of enzymatic polymer degradation studies. PET waste causes environmental damage worldwide, while the overall PET production steadily increases.[6] The great interest to find a sustainable solution for PET waste treatment is underlined by the growing number of publications regarding this topic during the last decade. In Google Scholar we obtain 965 hits for the combined keywords "polyethylene terephthalate" and "cutinase" for the last decade, 683 of them for the last five years. A number of comprehensive studies have been undertaken for the *Fusarium solani* Cutinase (*Fs*C; EC: 3.1.1.74) by combining experimental studies on enzyme kinetics with experimental and computational studies on structure and dynamics[5] - especially since the molecular structure of the *Fs*C has been solved at a resolution of 1 Å.[7] The good quality of the X-ray structure allowed for the detailed analysis of the time dependent behavior of *Fs*C via molecular dynamics (MD) simulations, at timescales up to 15 ns.[8] As opposed to other cutinases requiring extreme conditions, the catalytic optimum of *Fs*C lies at 40 °C, which makes it the ideal candidate for an environmentally sustainable process.

PET hydrolysis leads to oligomeric fragments as well as to monomeric terephthalic acid (TPA) and ethylene glycol (EG). *Fs*C is able to catalyze this process, but its hydrolysis rate is quite low for the wildtype and converges to zero after a period of 24 to 96 h so that it only achieves a total weight loss of PET film of 5%.[3] Previous mutation studies only considered structure or shape guided design, to enlarge the active site. In these mutants large residues were replaced by smaller ones as reviewed by Chen *et al.*[5] and references cited therein. This resulted in activity enhancement for high-molecular weight polyesters but not for low-molecular monoesters. While structure guided design is based on the structure of the protein, a rational design approach uses additional simulations, modelling, or statistics to predict a promising mutant. Furthermore, the issue of activity loss over time has not been addressed so far. It is important to fully understand the limitations of wildtype *Fs*C and to carve out a clear hypothesis about the requirements to the mutants.

Our study focuses the low and decreasing activity of wildtype *Fs*C during PET hydrolysis. We assume that the increasing amount of the cleavage products plays a key role in this context. The small polar water analogue EG appears to be a reasonable candidate as it increases the viscosity and density of the solvent and may alter the hydration of the protein.[9–11] TPA is unlikely because it is not soluble in water so that its concentration in the solution is negligible. The effect of EG monomers on the structure and dynamics of proteins has not yet been sufficiently studied. Thus, in our study we investigated the effect of increasing concentrations of EG on the structure and the dynamics of *Fs*C at a molecular level by combination of multi-scale simulations.

In the first part of our study, we used all-atom MD simulations to analyze the influence of different concentrations of EG on the enzyme dynamics. We chose MD to study the allosteric effects of solvent molecules as a state-of-the art method to determine protein dynamics and its solvent interactions.[12] In the second part, we used a coarse-grained model to investigate possible conformational changes of *Fs*C upon binding of a high-molecular weight polyester within the active site. Coarse-graining matches experimental data of small proteins or RNA structures (e.g. thermodynamics of the bovine trypsine inhibitor) up to huge biological complexes (e.g. assembly of the ribosomal subunits).[13–15] Furthermore, it overcomes the limitations of MD simulations regarding the required timescale, which means that less computational effort is needed to simulate larger timescales.[12] In particular, we used the linear response theory[16] (LRT) to simulate the substrate binding in the active site by an external force vector to investigate possible structural changes. Based on our findings, we point out changes in the protein structure that may lead to improved enzyme activity.

Material and Methods

Molecular Dynamics (MD) Simulations

MD simulations were performed using the native *Fs*C structure (PDB-Code: 1CEX) with a resolution of 1 Å.[7] The simulation box with dimensions of x = 59.90 Å, y = 57.16 Å, and z = 66.83 Å was filled with TIP3P water molecules and varying amounts of EG up to final concentrations of 0% (0 molecules), 2% (41 molecules), 3% (61 molecules), 5% (103 molecules), 10% (210 molecules), and 20% (420 molecules). To neutralize the simulation box, 9 Na^+ and 12 Cl^- ions were added to a final concentration of 0.9%. The simulations were performed with the Yasara software suite[17] and the AMBER03 force field[18] at constant temperature of 313 K, constant pressure of 1 bar, and constant pH of 7.4.

We used a van der Waals cutoff of 10 Å. Long range Coulomb interactions were calculated using the Particle Mesh Ewald algorithm. Grid points for the PME evaluation were evenly spaced in each dimension (27 grid points). For temperature control we used a velocity rescale thermostat which keeps the time average macroscopic temperature at the requested value by rescaling the atom velocities using a Berendsen thermostat.[19] For pressure control we chose the Manometer barostat in Yasara.[17]

Prior to the simulations, the simulation box including the *Fs*C structure was filled with the defined number of EG molecules, then filled with water, and at the end with counter ions. Possible clashes were removed via energy minimization using the steepest descent algorithm with subsequent simulated annealing until convergence, i.e. energy improvement of less than

$0.01\frac{kg}{J\cdot mol}$ per atom over 200 steps. After 500 ps of solvent equilibration the simulation with 2 fs time-steps was run for an overall simulation time of 100 ns.

During the simulation we did not use rototranslational constraints, but prior to trajectory analysis we preprocessed the trajectory files using `trjconv` function in gromacs[20] in order to correct drift and rotation of the proteins from their initial positions in the simulation boxes.

Parametrization for EG

For the parametrization of EG we used GAFF (General AMBER Force Field)[21] atom types and force field parameters followed by a calculation of semi-empirical AM1 Mulliken point charges[22] and a geometry optimization with the COSMO solvation model.[23] Furthermore we improved the AM1 charges for EG with the 'AM1 Bond Charge Correction'.[24] This parametrization procedure is carried out by Yasara.[17]

Tetrahedral Order Parameter

To account for the ability of water to form hydrogen bonds with adjacent water molecules and thus, to establish a tetrahedral network, the tetrahedral order parameter is defined as follows:[25–27]

$$Q_i = 1 - \frac{3}{8}\sum_{j=1}^{3}\sum_{k=j+1}^{4}\left[\cos(\psi_{jik}) + \frac{1}{3}\right]^2 . \tag{1}$$

The index i denotes the considered oxygen atom and j, k the nearest oxygen neighbours (not necessarily hydrogen bonded to the local atom). The time average $\left\langle \frac{1}{N}\sum_i Q_i \right\rangle$ of a system with N water molecules is 0 for random configurations and 1 for perfect tetrahedral orientation of all molecules. It can range form -3 to 1.

Mean Square Displacement

The mean square displacement (MSD) is a measure for the quantification of the dynamics of molecules.[28]

$$r^2(\tau) = \frac{1}{N}\sum_{i=1}^{N}[\vec{r}_i(\tau) - \vec{r}_i(0)]^2 \tag{2}$$

$\vec{r}_i(\tau)$ is the current position of particle i at timestep τ. $r^2(\tau)$ is the MSD at timestep τ for a system of N particles compared to the inital positions $\vec{r}_i(0)$. We used this measure to indirectly determine the viscosity of the solvent, as high viscosity correlates with reduced dynamics and vice versa. This antiproportional relation is given by the diffusion coefficient in the approximation of a sphere $D = \frac{k_\mathrm{B} \cdot T}{6 \cdot \pi \cdot \eta \cdot R_0}$ according to the Einstein-Stokes equation:

$$r^2(\tau) = 2n \cdot D \cdot \tau = \frac{n \cdot k_\mathrm{B} \cdot T \cdot \tau}{3 \cdot \pi \cdot \eta \cdot R_0} \tag{3}$$

with the Boltzmann constant k_B, temperature T, timestep τ, viscosity η and particle radius R_0 in a n-dimensional system. For our MD simulations we computed the MSD of the waters' oxygen atoms in order to indirectly quantify the solvent viscosity.

Root Mean Square Deviation

In order to quantify structural differences and conformational changes of the overall protein structure we used the root mean square deviation (RMSD). The RMSD can be computed as follows:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i^a - x_i^b\right)^2 + \left(y_i^a - y_i^b\right)^2 + \left(z_i^a - z_i^b\right)^2} \tag{4}$$

where N is the number of atoms and x_i^a is defined as the x coordinate of atom i in conformation a. All coordinates of all atoms in one conformation are compared to all coordinates of all atoms in another conformation. We used all Cα atoms and `trjconv fit rot+trans` as well as pbc corrections using `trjconv` function in gromacs[20] to process the trajectory. For the RMSD computation we compared the initial structure of the production run to all remaining frames. For curve smoothing we used the locally weighted scatterplot smoothing (LOESS) algorithm.[29]

Root Mean Square Fluctuation

The residual flexibility was analyzed by the root mean square fluctuation (RMSF). It represents the average movement per residue during the simulation of T frames by measuring the deviation of the coordinates x_i, y_i, and z_i of a specific residue i from the average coordinates $\tilde{x}_i$, $\tilde{y}_i$, and $\tilde{z}_i$.

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(x_i^t - \tilde{x}_i\right)^2 + \left(y_i^t - \tilde{y}_i\right)^2 + \left(z_i^t - \tilde{z}_i\right)^2} \tag{5}$$

RDF –Radial Distribution Function

The radial distribution function[30] $g(\vec{r})$ gives the probability of finding a particle $\vec{R}_j$ within a spherical shell of radius $\vec{r}$ from another particle $\vec{R}_i$ within an infinitesimal thickness:

$$g(\vec{r}) = \frac{V}{N^2} \cdot \sum_{i \neq j} \delta\big(\vec{r} - (\vec{R}_i - \vec{R}_j)\big) \tag{6}$$

with N particles in a system of volume V. For our analysis, we used the protein Cα atoms and computed the radial distribution of EG O atoms within radii from 2-30 Å and a thickness of 2 Å.

Surface Density Calculations

In our study we used $\mathcal{T} = 40,000$ frames for each simulation. We denoted the number of amino acids as $\mathcal{Q}$ and of EG mass centers as $\mathcal{E}$. We defined $V_{t,i}$ the volume between amino acid i (represented by its Cα atom) and the maximum interaction distance d to the solvent normalized by its Solvent Accessible Surface Area (SASA) at frame t:

$$V_{t,i} = SASA_{t,i} \cdot d \tag{7}$$

The SASA computations were carried out using the parameter optimized surface calculator (POPS).[31] For each simulation frame a SASA, at amino acid level as well as for the whole protein was calculated. This resulted in a time series of SASAs for each amino acid. Local densities of solvent molecules could be quantified as the number of particles ρ_i inside a volume fraction $V_{i,t}$.

To define the number of molecules occupying $V_{i,t}$ in a trajectory, we calculated the number of particles within the volume of every backbone amino acid i for every frame t. This yielded to the accumulation tensor M (Eq. 8) with the axis defined by simulation frame t, the backbone amino acid a_i, and the EG mass center e_j. We denoted $\|\vec{a}_{it} - \vec{e}_{jt}\|$ as the Euclidean norm of the Cα coordinates ($\vec{a}_i$) and the mass center coordinates of an EG molecule ($\vec{e}_j$) at frame t. The entries of an accumulation tensor M are defined as follows:

$$M_{ijt} = \begin{cases} 0 & \text{if } \|\vec{a}_{it} - \vec{e}_{jt}\| > d \\ 1 & \text{if } \|\vec{a}_{it} - \vec{e}_{jt}\| \leq d, \end{cases} \tag{8}$$

with d set to 7 Å, which corresponds to the first coordination shell of an amino acid. Now we could approximate the average local EG density ρ_i at amino acid i over a whole simulation by:

$$\rho_i = \frac{1}{\mathcal{T}} \sum_{t}^{\mathcal{T}} \sum_{j}^{\mathcal{E}} M_{t,i,j} \cdot V_{t,i}^{-1}. \tag{9}$$

Furthermore we defined the time dependent density (ρ_t) for the complete protein by:

$$\rho_t = \frac{1}{\mathcal{Q} \cdot \mathcal{E}} \sum_{i}^{\mathcal{Q}} \sum_{j}^{\mathcal{E}} M_{t,i,j} \cdot V_{t,i}^{-1}. \tag{10}$$

For density comparisons the mean density over a defined period of time was used:

$$\tilde{\rho} = \frac{1}{\mathcal{T}} \sum_{t}^{\mathcal{T}} \rho_t. \tag{11}$$

Linear Response Theory (LRT)

The linear response theory (LRT) introduced by Ikeguchi *et al.*[16] is a model to predict the structural changes of a protein upon ligand-binding. It is based on the normal mode analysis (NMA), which is a well suited method to study the collective motions in proteins.[32] To reduce computational effort, this method can also be applied on coarse-grained structures where proteins are reduced to a network of beads and springs. The beads represent the amino acids of the protein and the springs represent bonded or non-bonded interactions between several amino acids when their spatial distances fulfill a given cutoff criteria. Such a network is called an Elastic Network Model (ENM).[33]

Instead of treating the fluctuations as isoptropic, like a Gaussian Network Model (GNM) does, the LRT is based on an Anisotropic Network Model (ANM) that considers anisotropic fluctuations of amino acids.[34] It could be shown that combinations of low frequency modes correspond to the protein structural changes upon ligand-binding. Hence, using the LRT the direction of a structural change of a protein can be predicted via the formula:[16]

$$\Delta\vec{R}_i \simeq \beta \cdot \sum_j \langle \Delta\vec{R}_i \cdot \Delta\vec{R}_j \rangle_0 \cdot \vec{f}_j \tag{12}$$

where $\Delta\vec{R}_i$ represents the predicted translocation of atom i after the perturbation and $\langle \Delta\vec{R}_i \cdot \Delta\vec{R}_j \rangle$ denotes the covariance matrix of atomic fluctuations in the ligand free state. $\vec{f}_j$ represents the external force vector mimicking ligand-binding and β is $1/k_B T$ with the Boltzmann factor k_B. For the computation of the coordinate changes, the covariance matrix derived from an ANM or extracted from a MD simulation can be used. The covariance matrix can be computed as the Moore-Penrose pseudo-inverse[35,36] of the $3N \times 3N$ Hessian Matrix $\mathscr{H}$ that describes the second derivatives of the harmonic potential of the ANM with N residues:[33]

$$\mathscr{H} = \begin{pmatrix} \mathcal{H}_{11} & \mathcal{H}_{12} & ... & \mathcal{H}_{1N} \\ \mathcal{H}_{21} & \mathcal{H}_{22} & ... & \mathcal{H}_{2N} \\ \vdots & & & \\ . & & & \\ \mathcal{H}_{N1} & \mathcal{H}_{N2} & ... & \mathcal{H}_{NN} \end{pmatrix} \tag{13}$$

with a 3 × 3 super element

$$\mathcal{H}_{ij} = \begin{pmatrix} \frac{\partial^2 V}{\partial x_i \partial x_j} & \frac{\partial^2 V}{\partial x_i \partial y_j} & \frac{\partial^2 V}{\partial x_i \partial z_j} \\ \frac{\partial^2 V}{\partial y_i \partial x_j} & \frac{\partial^2 V}{\partial y_i \partial y_j} & \frac{\partial^2 V}{\partial y_i \partial z_j} \\ \frac{\partial^2 V}{\partial z_i \partial x_j} & \frac{\partial^2 V}{\partial z_i \partial y_j} & \frac{\partial^2 V}{\partial z_i \partial z_j} \end{pmatrix}. \tag{14}$$

Regarding *Fs*C, we used this model to predict the structural changes (induced fit) during substrate binding to the active site serine (S120). We obtained the covariance matrix from a heterogeneously parameterized ANM using the energy minimized structure of *Fs*C with spatial cutoffs of 7Å and 13Å, respectively for connected residues (see Figure 1). We used the matrix for intra-chain interactions between amino acids by Miyazawa and Jernigan[37] as well as the matrix for inter-chain interactions of amino acids by Keskin *et al.*[38] provided by the `R`[39] package `BioPhysConnectoR`.[40]

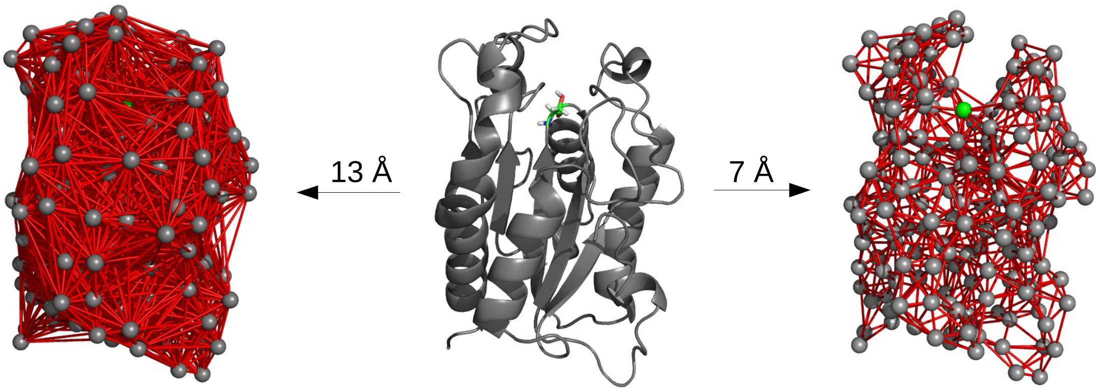


Figure 1: Illustration of the elastic network models with distance cutoffs of 7Å and 13Å for connected residues in comparison with the all-atom model of *Fs*C in cartoon representation. The Cα atoms are reduced to gray spheres and the connections between atoms are shown as red lines. The catalytic S120 is highlighted in green.

To mimic the substrate binding, the external force vector was directed towards S120 from a possible position for the substrate's carbonyl carbon upon formation of the tetrahedral intermediate. This position was randomly chosen from a cluster of accessible positions in the binding pocket. Note that no substrate or solvent is present in the ANM setup.

LRT Null Model

The implementation of a null model has proven beneficial to study the statistical significance of a computational approach.[41] To investigate the influences of the direction of the force vector as well as the significance of the above chosen force direction, we used a reference model of isotropic perturbation with 1,000 force vectors randomly originating from different points on a sphere around S120, similar to a previous study.[42] We chose spherical coordinates ϕ and θ uniformly distributed with $\phi \in [0, 2\pi]$ and $\theta \in [0, \pi]$ to generate 1,000 different force vectors. The force vectors $\vec{f_j} = (x_j, y_j, z_j)$ were parameterized as follows:

$$f_{xj} = f_0 \cdot \sin(\theta) \cdot \cos(\phi) \tag{15}$$

$$f_{yj} = f_0 \cdot \sin(\theta) \cdot \sin(\phi) \tag{16}$$

$$f_{zj} = f_0 \cdot \cos(\theta) \tag{17}$$

with f_0 being an arbitrary scaling factor that eventually has no impact on our subsequent results. The induced fit of the enzyme substrate complex was demonstrated by perturbing S120 with 1,000 different external force vectors $\vec{f_j}$ with repulsive and attractive forces in comparison.

To check for clustering, we clustered the displacement vectors of selected residues after perturbing S120 from each random direction. The selected residues for displacement calculation were residues 80-90, 179-187, and 42-45, as they are reported to participate in the functional behavior of FsC.[43] Force directions were clustered by applying the `kmeans` algorithm from Hartigan and Wong[44] on the 1,000×(3·24) matrix of the x-, y-, and z-displacements of the Cα atoms of the selected residues. The optimal number of clusters was investigated by comparing the log values of maximal within-cluster sum of squares (maximum withinss) from `kmeans` clustering as a function of number of clusters.

Software Contribution

`R`[39] is an environment for statistical analysis of data that offers many additional packages, especially for computational biology. We implemented the method of linear response upon substrate binding[16] in `R` using the `BioPhysConnectoR`[40] and `bio3d`[45] packages and enhanced the model by further statistics in our null model. We included both in the `LRTNullModel` package in `R` to make the applied methods accessible to the community. Software link: `http://www.cbs.tu-darmstadt.de/LRTNullModel.tar.gz`

Results

In order to investigate the influence of increasing concentrations of the cleavage product EG in the reaction solution on the activity of *Fs*C, we performed MD simulations with different EG concentrations (0%, 2%, 3%, 5%, 10%, and 20%) in the solvent. The trajectories were analyzed regarding two different aspects. First, we were interested in the influences of increasing amounts of EG on the overall dynamics of *Fs*C. Second, we analyzed the results with respect to local accumulations of EG on the surface of *Fs*C.

Increasing EG Concentrations Reduce the Overall Dynamics of *Fs*C

As a measure for the overall dynamics of *Fs*C during the simulation, we compared the RMSD values of the different runs (Figure 2 A). It is noticeable that increasing concentrations of EG in the solvent reduce the overall dynamics of *Fs*C. Just the small change from 2% to 3% EG in the solvent causes a remarkable drop of the RMSD values, restricting movements to at least half of the range found in pure water. Furthermore, we analyzed the effect on the residual fluctuations of *Fs*C in terms of RMSF (Figure 2 B). With increasing concentrations of EG the RMSF also declines for all residues.

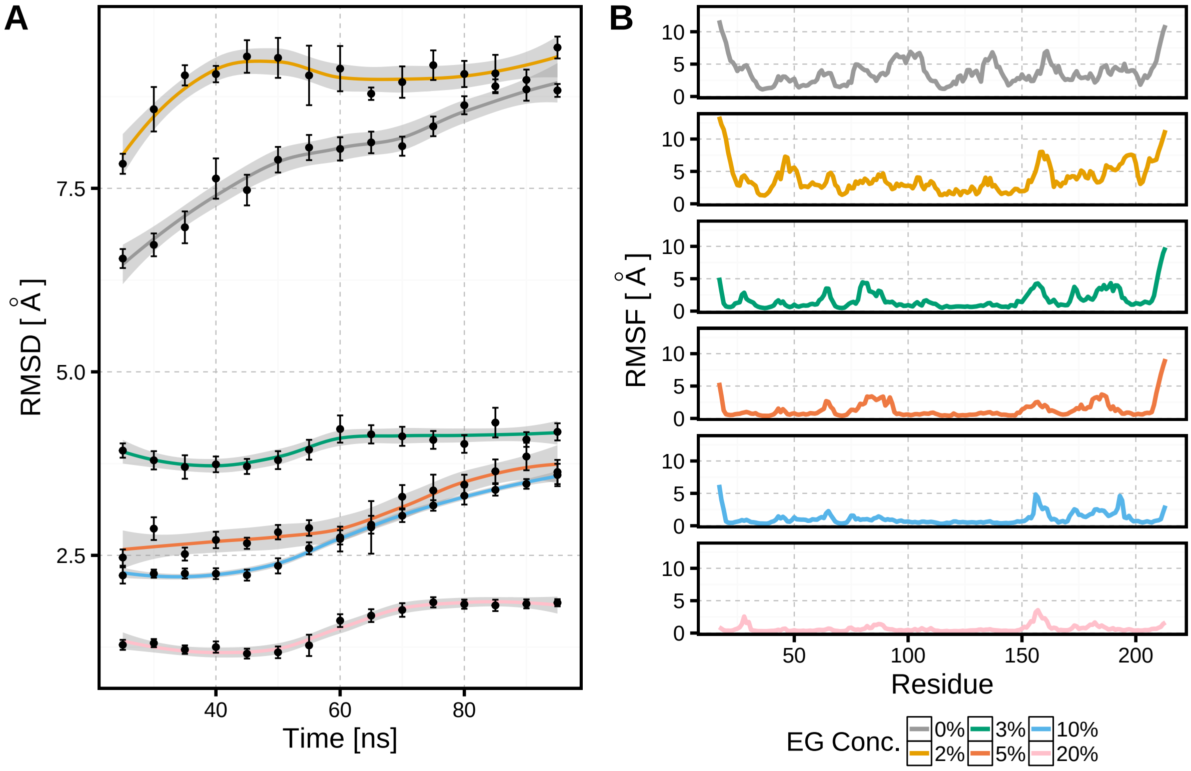


Figure 2: RMSD (in steps of 5 ns, i.e. 2000 frames) and RMSF of the MD simulations with different concentrations of EG in comparison. **A:** The RMSD curves were smoothed using the LOESS[29] algorithm. The grey area is the uncertainty of the fit, while the error bars represent the standard deviation of 2000 frames each. The higher the EG concentration, the lower the average RMSD (except for the simulation with 0%) - with a remarkable drop in overall dynamics of *Fs*C from 2% to 3% of EG. **B:** The residual flexibility also strongly reduces with increasing concentrations of EG.

Accumulation of EG near the Active Site

To find out if and where EG accumulates on the surface of *Fs*C, we performed surface density calculations based on the MD trajectories. Figure 3 A shows the mean occurrence of EG particles on the surface of *Fs*C for the trajectories with different EG concentrations. Furthermore, we distinguished between surface residues in the active site region and remaining surface residues. Residues with a SASA$<$1 Å^2 were not considered as surface residues. A 12.4-fold higher slope of particles per volume with rising EG concentration was observed for the surface residues near the active site than for the remaining surface residues (Figure 3 B). This indicates, that EG accumulates near the active site with increasing concentrations of EG.

For visual analysis, the average EG densities on the *Fs*C surface are depicted in Figure 4 via color-coded surface representations. With increasing EG concentrations, we observed higher densities of EG on the surface of *Fs*C. Furthermore it is noticeable that for increasing concentrations EG accumulates near the active site, which agrees with the results taken of Figure 3 B. These observations give a first indication why the activity of *Fs*C decreases during the hydrolysis of PET as increasing amounts of the cleavage product EG are generated.

EG Accumulations Reduce the local RMSF of catalytic H188

As EG accumulates near the active site, we focus on the catalytic triad and the oxyanion hole of *Fs*C (S120, D175, H188, Q121, S42, N84). We observed increasing EG densities on each catalytic residue when the EG concentrations of the solvents increased (Figure 5 A). The effects on the residual flexibility are pointed out by the local RMSF values (Figure 5 B). The accumulation of EG most effects the flexibility of H188 (drop from 5.5 Å to 0.5 Å), while for low EG concentrations H188 hardly encounters any EG at all.

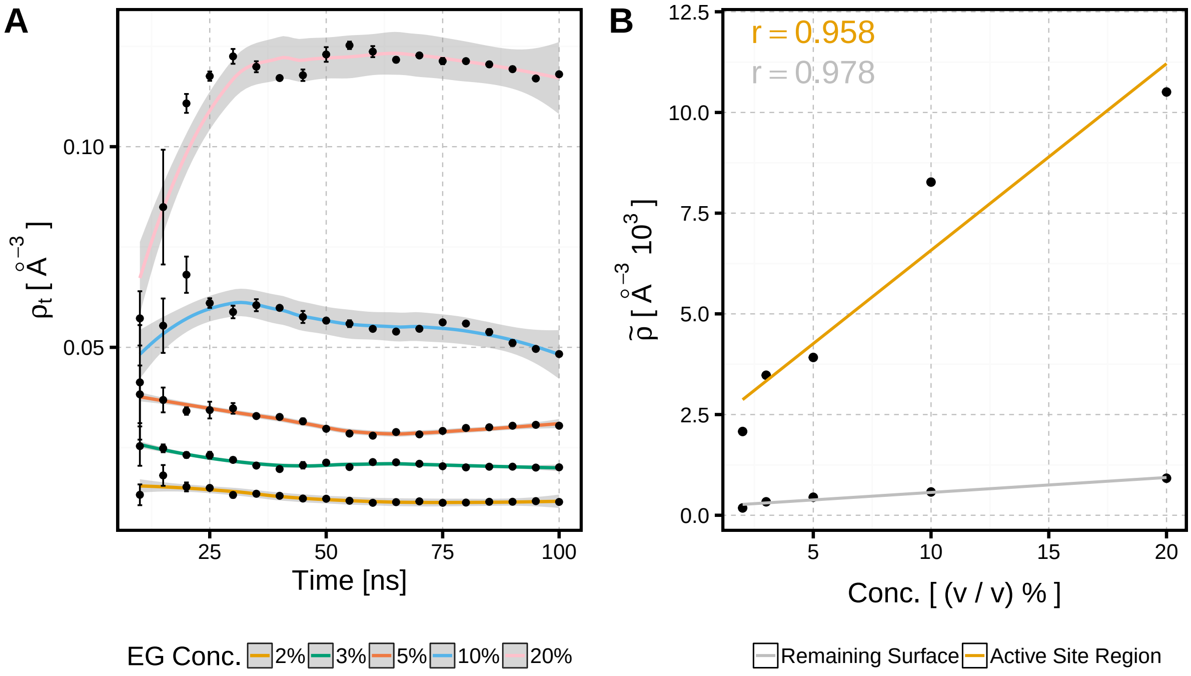


Figure 3: **A**: Time dependent densities of EG particles ρ_t for the simulations with different concentrations of EG in the solvent. The curves were smoothed using the LOESS[29] algorithm. The grey area is the uncertainty of the fit, while the error bars represent the standard deviation of 2000 frames each. **B**: Linear regression of mean densities $\tilde{\rho}$ of EG particles (60-100 ns) as a function of the concentration of EG in the solvent distinguished for surface residues near the active site (slope= 0.460, p-value= 0.0038) and the remaining surface residues (slope= 0.037, p-value= 0.04).

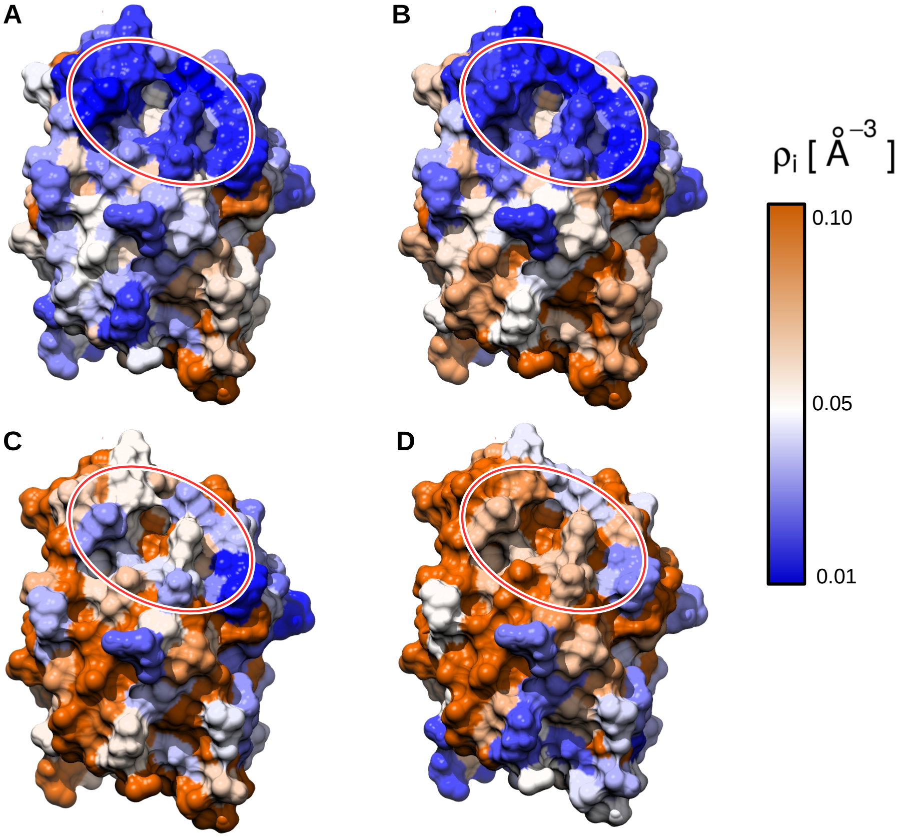


Figure 4: Accumulation of EG on the surface of *Fs*C. Blue denotes low densities of EG whereas red denotes high densities (in Particles/Å^3). From **A** to **D** increasing concentrations (3%, 5%, 10%, and 20%) of EG were used for the density calculations. With increasing concentrations we observe a movement of the regions of EG accumulation towards the active site (red ellipsoid).

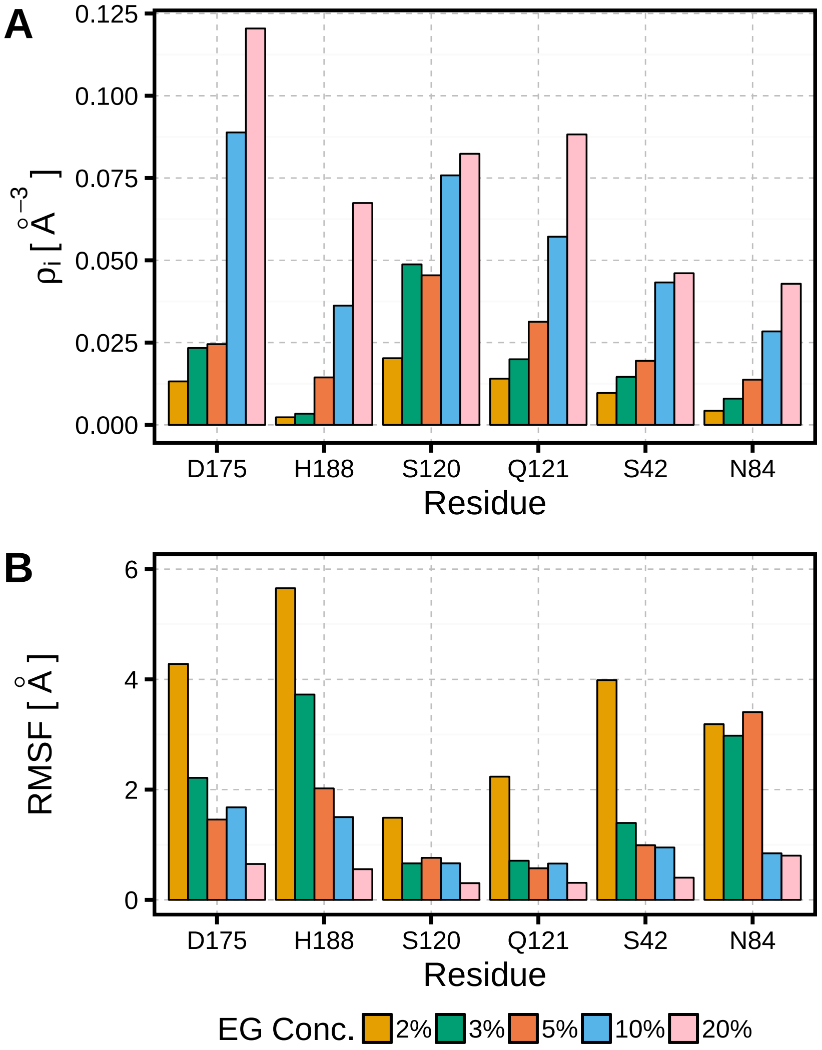


Figure 5: Densities of EG (**A**) and RMSF (**B**) for the residues in the active site and oxyanion hole based on the MD simulations of *Fs*C with different concentrations of EG in water.

Characterization of the ***Fs*****C** environment

To elucidate the influence of EG to the solvent properties, we investigated the mean square displacement (MSD) of the water molecules as an indirect measure of the solvent viscosity as well as the tetrahedral order parameter of water molecules to study the effects on the hydrogen bonding network. The MSD with EG does not significantly differ from the MSD without EG which indicates that the viscosity of the solvent is marginally influenced by EG (Figure 6 A). Interestingly, the distribution of the tetrahedral oder parameter Q_i is right shifted in the simulations with EG compared to the simulation without EG (Figure 6 B). The hydrogen bonding network is more structured, when EG is added to the solvent, no matter at which concentration.

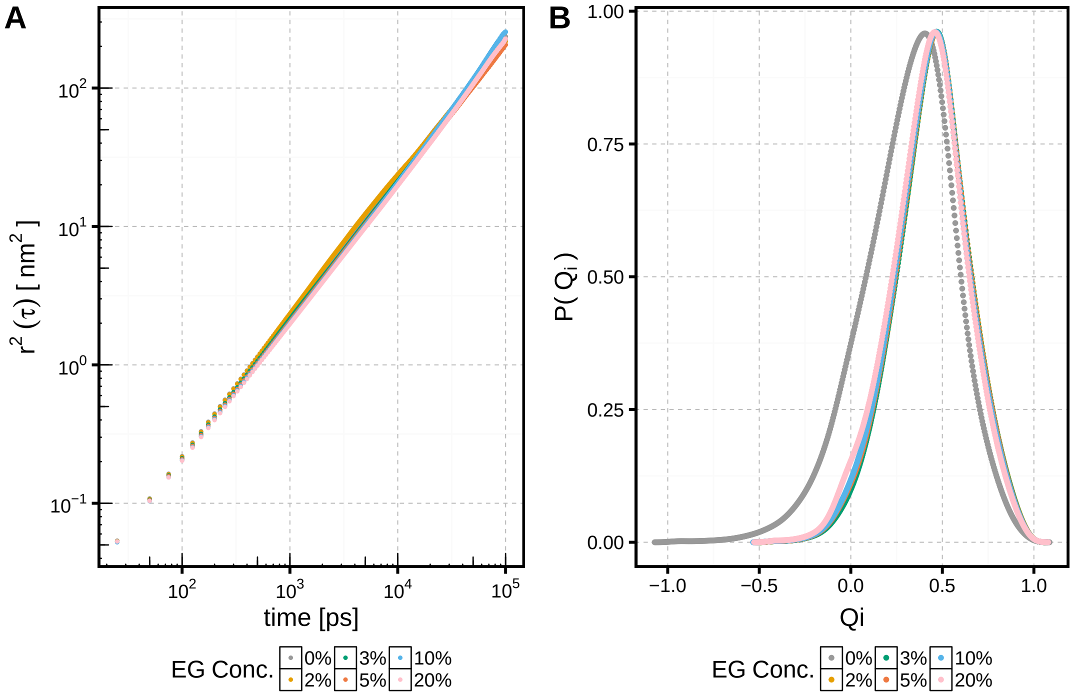


Figure 6: Characterization of the solvent for different concentrations of EG. **A:** Double logarithmic representation of MSD of the water molecules as indirect measure of the viscosity. The constant increase of the MSD during the simulation describes normal diffusion. The MSD for the different EG concentrations does not significantly differ, which means that the viscosity of the solvent is not influenced by EG. **B:** Tetrahedral order parameter Q_i to quantify the ability of water molecules to form hydrogen bonds to adjacent water molecules. In all simulations with EG, the distribution of Q_i values is right shifted compared to the simulation without EG. This indicates that with EG the hydrogen bonding network is in a more orderly state.

LRT Reveals the Need for more Flexibility in the Active Site Region

The results above show that with increasing amounts of EG in the solvent EG accumulates near the active site and decreases the flexibility of the catalytic residue H188 and causes a reduction in overall protein structural dynamics. Whether these observations are sufficient as explanations for the low activity of *Fs*C with progressing cleavage product release, had to be further evaluated. As QM/MM simulations[46] are prohibitively expensive for the simulation of enzymatic reactions in the bulk, a coarse-grained approach based on an ANM[33] was chosen. Using linear response theory (LRT), it is possible to investigate the structural response of proteins to a mechanical stimulus, e.g. due to ligand-binding. We used this method to predict the structural change of *Fs*C upon substrate-binding in the active site (formation of the tetrahedral intermediate) in order to investigate to what extent flexibility in the active site is actually required.

We computed the LRT model for distance cutoffs of 7Å and 13Å, respectively. The model with 7Å only includes the connections with residues in the first coordination shell, whereas larger cutoffs (e.g. 12-15Å) better reproduce experimental observations.[33] For the ENM with a cutoff distance of 13Å for connected residues, Figures 7 A and C show the structural response of *Fs*C upon mechanical perturbation at S120 (gray sphere) with increasing forces. Attractive force vectors with increasing forces were used to simulate the release of the cleavage product (Figure 7 B and D).

It is remarkable that for both cases mainly the loop regions near the active site are affected, while the α/β-core remains stable - although S120 is located at the interface of a β-sheet to an α-helix within the α/β-core. In case of the repulsive forces (substrate binding) the loops around the active site move closer to each other, leading to a closing of the binding pocket. In contrast to that, the attractive forces cause movements of the loops away from each other corresponding to an opening of the binding pocket. These observations confirm the hypothesis of a "breath-like" movement of *Fs*C during the hydrolysis reaction proposed by Longhi *et al.*[7]

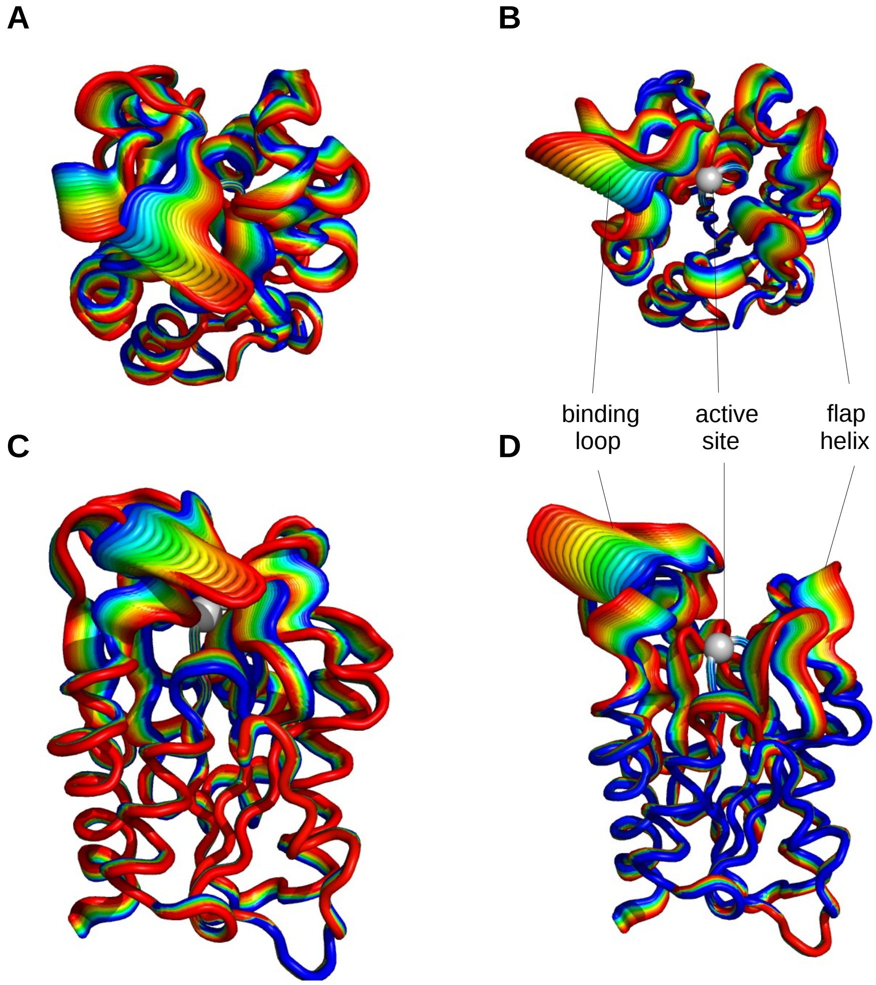


Figure 7: LRT model with repulsive **A** + **C** and attractive **B** + **D** force vectors with different forces between 0 (blue) and 3000 (red) in arbitrary units from top and side view (right angle to each other). The repulsive forces represent substrate binding and formation of the tetrahedral intermediate whereas the attractive forces represent the release of the cleaved substrate. The observed motions strengthen the idea of a "breath-like" movement as reported by Longhi *et al.*[7]

Analysis of LRT Results based on a Null Model

To prove the significance of the above results with a randomly chosen force direction, we used a reference model of isotropic perturbation. We clustered the 1,000 different force directions regarding their resulting displacements of selected residues after perturbation of S120 with the corresponding force vector (for more information see Methods Section). The development of the log values of maximal cluster sum of squares (maximum withinss) after clustering as a function of the number of clusters using the `kmeans` algorithm (Figure S1) convergences within ten iterations and has the most obvious drop from three to four clusters. Therefore, the use of four clusters is plausible.

The result of the clustering is shown in Figure 8. The 1,000 different force directions represented as 1,000 different end points on a sphere around S120 (Cα atom) are colored according to their assignment to a cluster. All force directions of a cluster result in the same conformational change of the functionally relevant regions of FsC.

Interestingly, the force directions belonging to one cluster are located in four different spatial areas. This means that functional clustering is linked with the spatial distribution of the force directions. It is remarkable, that the possible force directions for the substrate to bind to S120 from a sterical side of view, exclusively belong to the green cluster. This proves the insensitivity of the obtained structural changes upon substrate-binding in our LRT model towards small inaccuracies in the structural mechanical model.

To assess the robustness of the clusterings by our LRT reference model, we compared the probability to be in cluster i by the Kullback-Leibler Divergence:[47] $D_{KL}(P||Q) = \sum p_i \cdot ln(\frac{p_i}{q_i})$, where p_i denotes for the probability to be in i in one run of the LRT protocol and q_i is the probability to be assigned to i in an independent repetition. Our 999,000 comparisons showed a mean (maximum) D_{KL} of 0.14 nat (0.03 nat). These rather small values suggest high robustness of the implied clusterings.

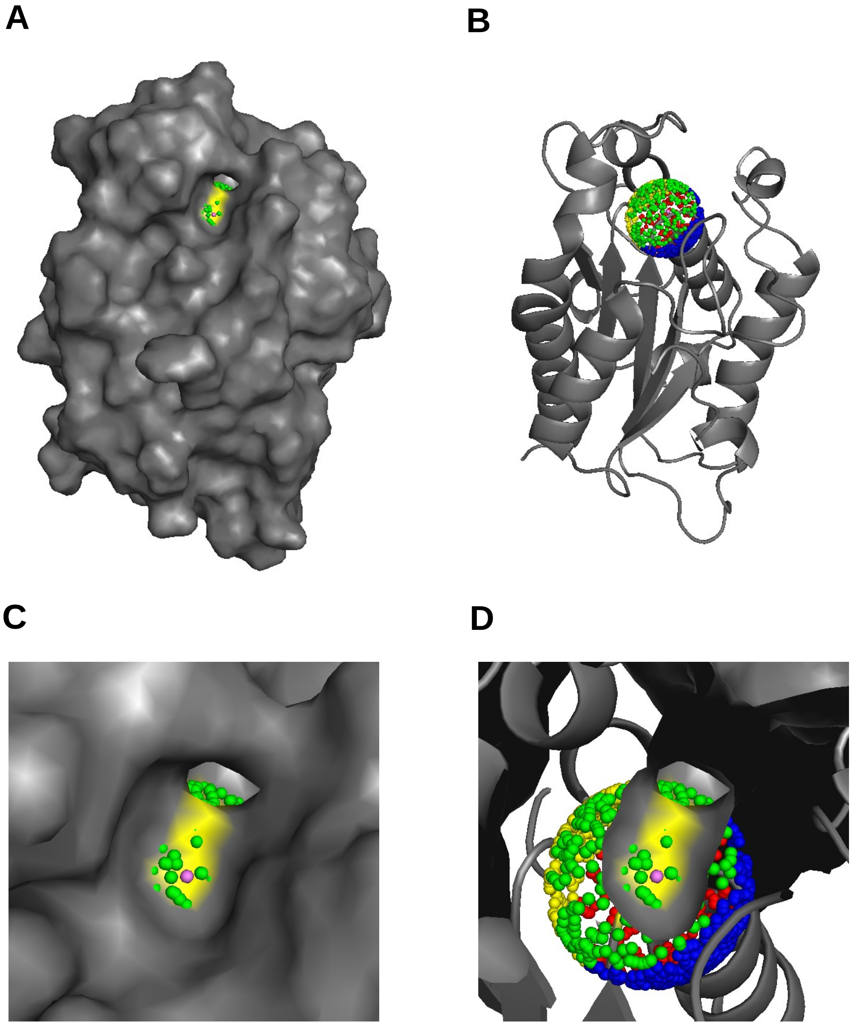
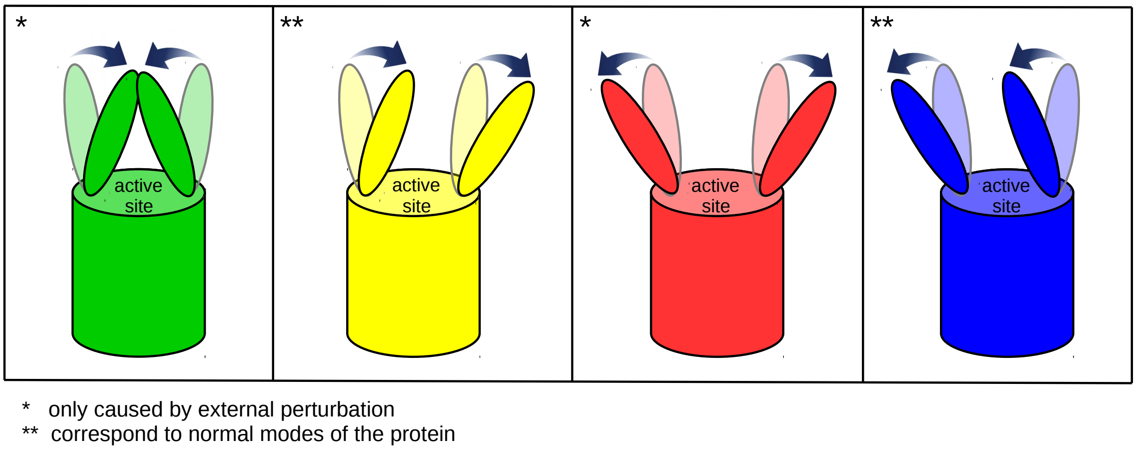


Figure 8: LRT null model with 1,000 force directions for the external force vector represented as small spheres colored according to the cluster. In this illustration, the spheres have a distance of 3.9 Å to the Cα atom of S120 in order to visualize possible positions for the substrate on the surface of the binding site. The realistic force directions from which the substrate can perturb the S120 during the catalytic intermediate, all belong to one cluster (green). The randomly selected force direction within the group of realistic ones, which was used for the LRT model in the previous section, is highlighted in violet. **A** and **C**: surface representation. **B**: cartoon representation with a sphere of 1,000 force directions. **D**: zoom into the surface to see the non-realistic force directions of the other clusters.



* only caused by external perturbation
** correspond to normal modes of the protein

Figure 9: Schematic representation of the loop displacements near the active site after perturbing S120 with repulsive forces from other clusters. The small representations are colored regarding the respective clusters in Figure 8. Note: The parallel loop motions caused by perturbation from directions in the yellow and blue clusters correspond to normal modes of *Fs*C, while the "breath-like" motions caused by perturbation from possible directions in the active site pocket do not correspond to normal modes. They only occur by external perturbation due to substrate binding or cleavage.

Figure 9 shows the representative movements of the two main loops near the active site after perturbing S120 in a schematic manner. Here, the results for repulsive forces are shown which mimic the binding of the substrate and formation of the tetrahedral intermediate. For attractive forces, the opposite displacements occur.

To ensure that the observed "breath-like" motions are actually caused by substrate binding/cleavage, we compared them to the low frequency normal modes of *Fs*C. Low frequency normal modes (eigenvectors of the Hessian matrix) underlie equilibrium dynamics of a protein and represent intrinsically accessible motions (conformational changes) without external forces acting on the protein.[48] In fact, the normal modes only comprise motions of the loop regions near the active site that either move parallel to the left or parallel to the right (Figure S2, Movie M1, Movie M2), similar to those observed by perturbations from the yellow or blue cluster. Normal modes do not comprise "breath-like" motions that are similar to those caused by the perturbation of S120 from possible directions in the active site pocket.

Summary and Discussion

The aim of our study was to investigate the reasons for the low and decreasing activity of wildtype *Fs*C during hydrolysis of PET in order to find means of improving enzyme activity. For this purpose, we focused on the influence of increasing concentrations of the cleavage product EG on (1) the overall dynamics of *Fs*C and (2) the accumulation of EG on the surface of *Fs*C via explicit all-atom MD simulations. We found, that increasing concentrations of EG result in reduced flexibility of *Fs*C caused by EG accumulation near the active site. The local flexibility of the catalytic H188 is most affected by this accumulation. With our simulations we can confirm important residual fluctuations measured in NMR studies of *Fs*C[49–51] (Figure S3). These NMR studies already pointed out the catalytic H188 to be highly flexible, which also goes along with our RMSF and surface density calculations, irrespective of the different timescales. As shown by Prompers *et al.*[50] H188 actually requires this flexibility to enable the enzymatic reactions.

The high RMSD values of the MD simulations with 0% and 2% EG in the solvent could lead to the assumption that unfolding events could have occured. However, structural analysis of our protein (Figure S4) proves that our MD simulations are stable over time.

Based on the characterization of the protein environment during the MD simulations, we sum up that the hydrogen bonding network of the water molecules in the solvent is more orderly with EG than without. Interestingly, the dynamics of water molecules, which we analyzed in terms of MSD, were only marginally reduced. At first glance these observations do not match, but having a look at the radial distribution function of protein Cα to EG O, we see that EG mainly accumulates near *Fs*C (Figure S5 A), which fits well with our surface density computations. This means that there are not many EG molecules left in the bulk to reduce the viscosity significantly. To make sure that the reduced flexibility of *Fs*C is actually caused by the accumulation of EG and not only by solvent mediated effects caused by the hydroxy functionality of an alcohol in general, we additionally made MD simulations with the same simulation setup and force field but with 5% methanol (MeOH) or 5% ethanol (EtOH) respectively, instead of EG (Figures S5-S7). MeOH strongly increases water dynamics and thus strongly reduces the viscosity. Furthermore, it increases the protein RMSD as well as RMSF. Interestingly, it shows the same effect on the hydrogen bonding network of the water molecules as with 5% EG, but MeOH is mostly located in the bulk and does not accumulate near *Fs*C. For EtOH we observe hardly any effect compared to the simulation without any alcohol added to the solvent. It seems that, at least at this concentration, EtOH behaves like water in the solvent. We only see a slight accumulation of EtOH near *Fs*C. This suggests that few EtOH molecules alter the dynamics of the protein, which is evident from Figure S5 B.

To account for the preferred type of interaction between EG and the active site residues, we determined the distribution of densities of EG near hydrophobic and hydrophilic surface residues in the active site and for all remaining surface residues in comparison. We found that EG densities are significantly higher for hydrophilic than for hydrophobic surface residues in the active site (Figure S8). For the remaining surface, EG densities are quite similar for both types with a slight tendency towards hydrophobic residues. This shows that the accumulation of EG monomers within or near the active site is mainly based on hydrophilic interactions, i.e. hydrogen-bonding interactions.

The importance of the flexibility of the region near the active site actually was also shown via the LRT model simulating the induced fit upon substrate-binding. While all-atom simulations of enzymatic reactions are not possible with the available computational resources, our coarse-grained model is a valid method to investigate the required flexibility during the reaction and to demonstrate the induced fit. Contrary to previous X-ray studies[52] predicting a preformed oxyanion hole, Prompers *et al.*[50] observed an induced fit during NMR experiments, which was our motivation for using the LRT method to investigate the mechanical principles behind the induced fit. We were able to demonstrate significant motions of the loop regions near the active site that do not correspond to the intrinsically accessible normal modes of *Fs*C and thus confirm the "breath-like" movement proposed by Longhi *et al.*[7]

The significance of our LRT results was further validated via a LRT null model. We found four main conformational changes of the loop regions near the active site. Interestingly, the random directions belonging to the same clusters are located in discrete spatial areas. By using this null model we were able to demonstrate that (1) our model is realistic and (2) that the presented results of the LRT model are reliable, as all physically possible force directions on the active site surface belong to the same cluster so that small perturbations would yield the same result in our modelling approach.

The observed movements of the loop regions near the active site appear straightforward for the LRT model based on the cutoff distances of 13 Å, as there are direct connections (springs) between S120 and the loop regions (Figure 1, left). However, our model with the smaller distance cutoff of 7Å shows similar structural conformations of the loop regions near the active site (Figure S9), although there is no direct connection (Figure 1, right). This demonstrates that the mechanism of mechanical force transfer is highly complex and forces can be transferred over a wide range of edges within the whole network. The corresponding null model is shown in Figure S10. To account for long-range interactions Figure S11 shows the displacements for 15, 17, and 21 Å cutoff, which do not differ in their direction, only in their magnitude. This demonstrates that the 13 Å modell does not disregard possible deviating long-range interactions.

We suppose that bigger substrates correspond to higher forces acting on S120 during the formation of the tetrahedral intermediate. This means, that in the LRT model, the red conformations (high forces) correspond to the induced fit caused by binding of high-molecular weight polyesters, whereas the conformations near the blue one (low forces) correspond to the induced fit caused by low-molecular weight substrate binding. Therefore, our results underline the necessity for flexibility of the active site regions for lager substrates.

Conclusion and Outlook

In our study we found that (1) with increasing concentrations EG accumulates near the active site and reduces the overall flexibility of *Fs*C and (2) that the loop regions near the active site perform a "breath-like" motion during substrate binding and cleavage. From this, we conclude that increasing accumulation of EG negatively affects the activity of *Fs*C which is based on the following mechanism: The arrangement of residues of the active site and the residues stabilizing the tetrahedral intermediate via H-bonds within the oxyanion hole is crucial for the success of the catalytic mechanism (Figure 10). In order to enable the nucleophilic attack of S120 at the carbonyl carbon of the substrate's ester bond, the nucleophilicity of the hydroxy group of S120 has to be increased. This is achieved by the hydrogen bond network within the catalytic triad (D175, H188, S120).[5] Due to the nucleophilic attack, a covalent bond between S120 and the substrate is formed and the proton is transferred to the adjacent H188. Residues S42 and Q121 stabilize the negatively charged carbonyl oxygen of the tetrahedral intermediate via H-bonds, so that further steps of the catalyic mechanism are facilitated.[5] N84 and Q121 further stabilize the position of S42.[7] During the loop movements, especially upon opening of the active site to accomodate large substrates, the respective residues are pulled appart from the perfect arrangement. A loss of flexibility of these residues impairs their instantaneous self-rearrangement to the positions required for the catalyic triad and oxyanion hole. This distortion is likely to reduce binding and substrate conversion rates. With further accumulation of EG and increased rigidity, even the loop motions may abate so that the binding of high-molecular substrates, like PET, is completely prevented.

One intuitive solution of the problem of EG accumulation would be the improvement of the industrial degradation process towards removing the cleavage products during the process. A further solution is to improve the enzyme properties by means of rational design in order to counteract this cleavage product mediated effect. For this purpose, mutants with

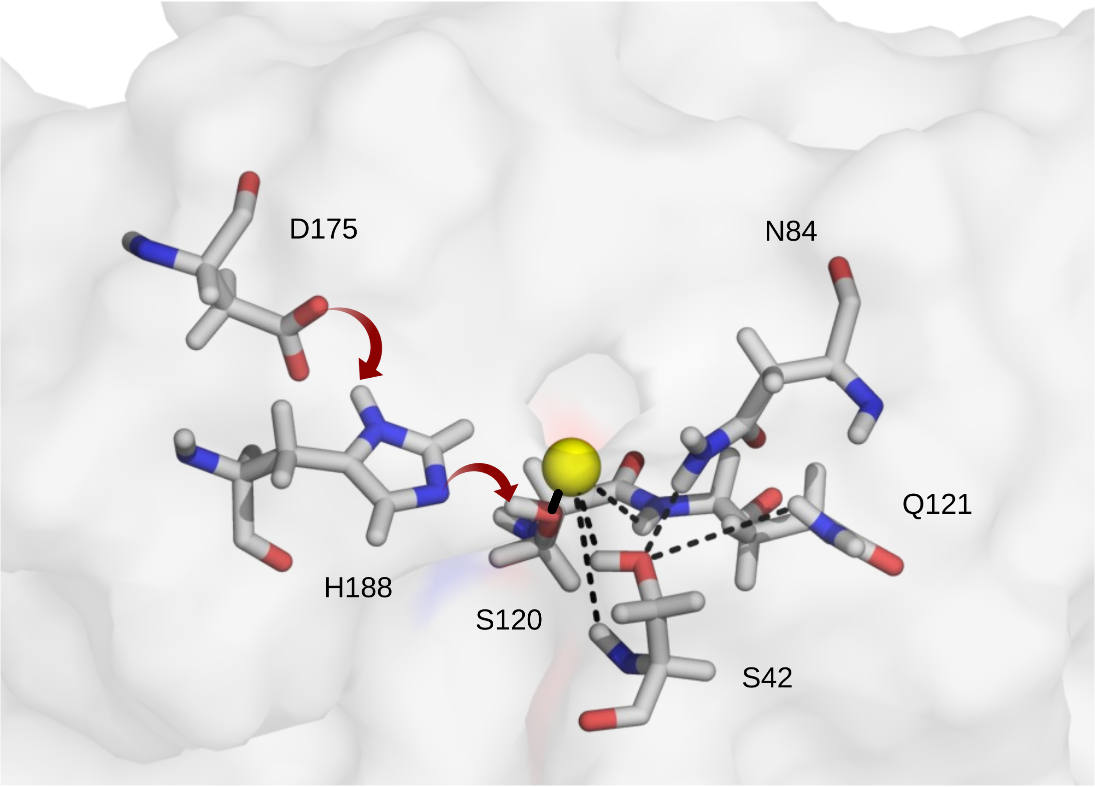


Figure 10: Active site (D175, H188, S120) and oxyanion hole (S42, Q121, N84) of *Fs*C as stick representation. Note that the residues are in close spatial proximity even though they are located at remote positions in the protein sequence. In the catalytic triad, red arrows demonstrate the hydrogen bond network, so that the more nucleophilic S120 can attack the substrate (carbonyl group of the ester bond simplified as yellow sphere). The covalent bond of the tetrahedral intermediate is shown as a black solid line. The stabilizing hydrogen bonds in the oxyanion hole are shown as black dashed lines. To illustrate the orientation within the protein its transparent surface is shown.

increased flexibility near the active site should be favored. Furthermore, mutants with an increased hydrophobicity in the active site might be promising candidates as we found that the accumulation of EG in the active site is mainly based on hydrogen-bonding interactions. The initiation of EG accumulation could be prevented by reducing the number of possible hydrogen-bonding partners. As our LRT model demonstrates that the mechanical correlations are rather complex, more sophisticated methods than structure guided design should be considered.

Our conceptual results have to be complemented by experiments investigating the enzymatic activity in terms of K_m and k_{cat}, e.g. colorimetric assays.[53] In these experiments standard 4-nitrophenyl esters (like pNPA, pNPB, pNPP, etc.) are commonly used to determine the activity of *Fs*C towards the hydrolysis of low-molecular esters but not necessarily towards the hydrolysis of PET. We propose to search for more complex model substrates for 4-nitrophenyl assays, that mimic PET more closely.

During the preparation of the current manuscript the discovery of bacteria that are supposed to be able to digest PET has been published.[54] The responsible enzymes may be quite interesting for enzymatic PET degradation, but a lot of basic research has to be done in order to reach the current state of knowledge that *Fs*C has in its community.

Acknowledgement

Christine Groß und Kay Hamacher thank the LOEWE project iNAPO funded by the Ministry of Higher Education, Research and the Arts (HMWK) of the Hessen state. Sven Jager thanks for support by iGEM@TU-Darmstadt 2012-2015.

Associated Content

Figures S1-S11 depicting `kmeans` clustering, lowest-frequency normal mode, comparison with NMR data, secondary structure analysis, radial distribution functions, RMSD and RMSF using other alcohols, solvent characterization using other alcohols, distributions of EG densities, LRT model with 7Å, LRT null model with 7Å, LRT with long-range interactions (PDF)

Movies M1-M2 depicting simultaneous motions of all ten lowest frequency normal modes using two different force fields (GIF)

This information is available free of charge via the Internet at http://pubs.acs.org

References

(1) Geyer, B.; Lorenz, G.; Kandelbauer, A. Recycling of Poly(ethylene terephthalate)–A Review Focusing on Chemical Methods. *Express Polym Lett* **2016**, *10*.

(2) Longhi, S.; Cambillau, C. Structure-Activity of Cutinase, a Small Lipolytic Enzyme. *Biochim. Biophys. Acta* **1999**, *1441*, 185–196.

(3) Ronkvist, Å. M.; Xie, W.; Lu, W.; Gross, R. A. Cutinase-Catalyzed Hydrolysis of Poly(ethylene terephthalate). *Macromolecules* **2009**, *42*, 5128–5138.

(4) Herrero Acero, E.; Ribitsch, D.; Steinkellner, G.; Gruber, K.; Greimel, K.; Eiteljoerg, I.; Trotscha, E.; Wei, R.; Zimmermann, W.; Zinn, M.; Cavaco-Paulo, A.; Freddi, G.; Schwab, H.; Guebitz, G. Enzymatic Surface Hydrolysis of PET: Effect of Structural Diversity on Kinetic Properties of Cutinases from *Thermobifida*. *Macromolecules* **2011**, *44*, 4632–4640.

(5) Chen, S.; Su, L.; Chen, J.; Wu, J. Cutinase: Characteristics, Preparation, and Application. *Biotechnol. Adv.* **2013**, *31*, 1754–1767.

(6) Merchant Research & Consulting Ltd., Polyethylene Terephthalate (PET): 2016 World Market Outlook and Forecast up to 2020. 2016.

(7) Longhi, S.; Czjzek, M.; Lamzin, V.; Nicolas, A.; Cambillau, C. Atomic Resolution (1.0 Å) Crystal Structure of *Fusarium solani* Cutinase: Stereochemical Analysis. *J. Mol. Biol.* **1997**, *268*, 779–799.

(8) Matak, M. Y.; Moghaddam, M. E. The Role of Short-Range Cys171-Cys178 Disulfide Bond in Maintaining Cutinase Active Site Integrity: A Molecular Dynamics Simulation. *Biochem. Biophys. Res. Commun.* **2009**, *390*, 201–204.

(9) Tsierkezos, N. G.; Molinou, I. E. Thermodynamic Properties of Water+ Ethylene Glycol at 283.15, 293.15, 303.15, and 313.15 K. *J. Chem. Eng. Data* **1998**, *43*, 989–993.

(10) Sun, T.; Teja, A. S. Density, Viscosity, and Thermal Conductivity of Aqueous Ethylene, Diethylene, and Triethylene Glycol Mixtures between 290 K and 450 K. *J. Chem. Eng. Data* **2003**, *48*, 198–202.

(11) Schmitz, R.; Müller, N.; Ullmann, S.; Vogel, M. A Molecular Dynamics Simulations Study on Ethylene Glycol-Water Mixtures in Mesoporous Silica. *J. Chem. Phys.* **2016**, *145*, 104703.

(12) Maximova, T.; Moffatt, R.; Ma, B.; Nussinov, R.; Shehu, A. Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics. *PLoS Comput. Biol.* **2016**, *12*, e1004619.

(13) Hamacher, K. Free Energy of Contact Formation in Proteins: Efficient Computation in the Elastic Network Approximation. *Phys Rev E Stat Nonlin Soft Matter Phys* **2011**, *84*, 1–6.

(14) Jager, S.; Schiller, B.; Strufe, T.; Hamacher, K. StreAM-T_g: Algorithms for Analyzing Coarse Grained RNA Dynamics Based on Markov Models of Connectivity-Graphs. International Workshop on Algorithms in Bioinformatics. 2016; pp 197–209.

(15) Hamacher, K.; Trylska, J.; McCammon, J. A. Dependency Map of Proteins in the Small Ribosomal Subunit. *PLoS Comput. Biol.* **2006**, *2*, 80–87.

(16) Ikeguchi, M.; Ueno, J.; Sato, M.; Kidera, A. Protein Structural Change Upon Ligand Binding: Linear Response Theory. *Phys. Rev. Lett.* **2005**, *94*.

(17) Krieger, E.; Vriend, G. Increasing the Precision of Comparative Models with YASARA NOVA — a Self-Parameterizing Force Field. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 393–402.

(18) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P. A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

(19) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(20) Pronk, S.; Páll, S.; Schulz, R.; Larsson, P.; Bjelkmar, P.; Apostolov, R.; Shirts, M. R.; Smith, J. C.; Kasson, P. M.; van der Spoel, D.; Hess, B.; Lindahl, E. GROMACS 4.5: A High-Throughput and Highly Parallel Open Source Molecular Simulation Toolkit. *Bioinformatics* **2013**, *29*, 845–854.

(21) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(22) Stewart, J. J. MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput. Aided Mol. Des.* **1990**, *4*, 1–103.

(23) Klamt, A. Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem.* **1995**, *99*, 2224–2235.

(24) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

(25) Errington, J. R.; Debenedetti, P. G. Relationship between Structural Order and the Anomalies of Liquid Water. *Nature* **2001**, *409*, 318–321.

(26) Lynden-Bell, R. M.; Debenedetti, P. G. Computational Investigation of Order, Structure, and Dynamics in Modified Water Models. *J. Phys. Chem. B* **2005**, *109*, 6527 – 6534.

(27) Kumara, P.; Buldyrevb, S. V.; Stanleyc, H. E. A Tetrahedral Entropy for Water. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 22130 – 22134.

(28) Klameth, F. From Brownian Motion to Supercooled Water in Confinements - A Molecular Dynamics Simulation Study. Ph.D. thesis, Technische Universität, Darmstadt, 2015.

(29) Cleveland, W. S. Robust Locally Weighted Regression and Smoothing Scatterplots. *J Am Stat Assoc* **1979**, *74*, 829–836.

(30) Yarnell, J.; Katz, M.; Wenzel, R. G.; Koenig, S. Structure Factor and Radial Distribution Function for Liquid Argon at 85 K. *Phys. Rev. A* **1973**, *7*, 2130.

(31) Cavallo, L.; Kleinjung, J.; Fraternali, F. POPS: A Fast Algorithm for Solvent Accessible Surface Areas at Atomic and Residue Level. *Nucleic Acids Res.* **2003**, *31*, 3364–3366.

(32) Tama, F.; Sanejouand, Y. H. Conformational Change of Proteins Arising from Normal Mode Calculations. *Protein Eng.* **2001**, *14*, 1–6.

(33) Atilgan, A. R.; Durell, S. R.; Jernigan, R. L.; Demirel, M. C.; Keskin, O.; Bahar, I. Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophys. J.* **2001**, *80*, 505–515.

(34) Tirion, M. M. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905–1908.

(35) Moore, E. H. On the Reciprocal of the General Algebraic Matrix. *Bull. Am. Math. Soc.* **1920**, *26*, 385–396.

(36) Penrose, R. A Generalized Inverse for Matrices. *Math. Proc. Cambridge Philos. Soc.* **1955**, *51*, 406–413.

(37) Miyazawa, S.; Jernigan, R. L. Residue-Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. *J. Mol. Biol.* **1996**, *256*, 623–644.

(38) Keskin, O.; Bahar, I.; Badretdinov, A. Y.; Ptitsyn, O. B.; Jernigan, R. L. Empirical Solvent-Mediated Potentials Hold for Both Intra-Molecular and Inter-Molecular Inter-Residue Interactions. *Protein Sci.* **1998**, *7*, 2578–2586.

(39) R Development Core Team, R: A Language and Environment for Statistical Computing. 2008.

(40) Hoffgaard, F.; Weil, P.; Hamacher, K. BioPhysConnectoR: Connecting Sequence Information and Biophysical Models. *BMC Bioinform.* **2010**, *11*, 199.

(41) Weil, P.; Hoffgaard, F.; Hamacher, K. Estimating Sufficient Statistics in Co-Evolutionary Analysis by Mutual Information. *Comput. Biol. Chem.* **2009**, *33*, 440–444.

(42) Knorr, S. In Silico Strategies to Modulate DNA Damage Response. Ph.D. thesis, Technische Universität, Darmstadt, 2015.

(43) Creveld, L. D.; Amadei, A.; van Schaik, R. C.; Pepermans, H. A. M.; de Vlieg, J.; Berendsen, H. J. C. Identification of Functional and Unfolding Motions of Cutinase as Obtained from Molecular Dynamics Computer Simulations. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 253–264.

(44) Hartigan, J. A.; Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *J. R. Stat. Soc. Series C (Applied Statistics)* **1979**, *28*, 100–108.

(45) Grant, B. J.; Rodrigues, A. P. C.; ElSawy, K. M.; McCammon, J. A.; Caves, L. S. D. Bio3d: An R Package for the Comparative Analysis of Protein Structures. *Bioinformatics* **2006**, *22*, 2695–2696.

(46) Pople, J. A. Quantum Chemical Models (Nobel Lecture). *Angew. Chem., Int. Ed.* **1999**, *38*, 1894–1902.

(47) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.

(48) Bahar, I.; Lezon, T. R.; Bakan, A.; Shrivastava, I. H. Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins. *Chem. Rev.* **2009**, *110*, 1463–1497.

(49) Prompers, J. J.; Hilbers, C. W.; Groenewegen, A.; Schaik, R. C. V.; Pepermans, H. A. M. 1H, 13C, and 15N Resonance Assignments of *Fusarium solani pisi* Cutinase and Preliminary Features of the Structure in Solution. *Protein Sci.* **1997**, *6*, 2375–2384.

(50) Prompers, J. J.; Groenewegen, A.; Hilbers, C. W.; Pepermans, H. A. M. Backbone Dynamics of *Fusarium solani pisi* Cutinase Probed by Nuclear Magnetic Resonance: The lack of Interfacial Activation Revisited. *Biochemistry* **1999**, *38*, 5315–5327.

(51) Poulsen, K. R.; Sørensen, T. K.; Duroux, L.; Petersen, E. I.; Petersen, S. B.; Wimmer, R. The Interaction of *Fusarium solani pisi* Cutinase with Long Chain Spin Label Esters. *Biochemistry* **2006**, *45*, 9163–9171.

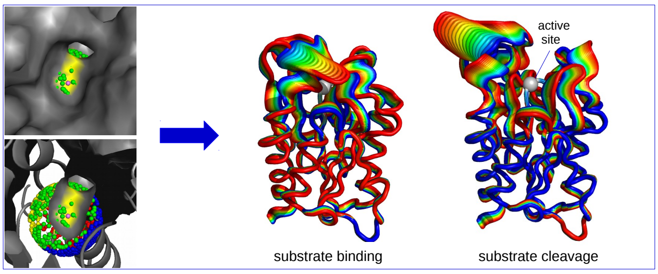
(52) Martinez, C.; De Geus, P.; Lauwereys, M.; Matthyssens, G.; Cambillau, C. *Fusarium solani* Cutinase is a Lipolytic Enzyme with a Catalytic Serine Accessible to Solvent. *Nature* **1992**, *356*, 615–618.

(53) Buß, O.; Jager, S.; Dold, S.-M.; Zimmermann, S.; Hamacher, K.; Schmitz, K.; Rudat, J. Statistical Evaluation of HTS Assays for Enzymatic Hydrolysis of β-Keto Esters. *PloS One* **2016**, *11*, e0146104.

(54) Yoshida, S.; Hiraga, K.; Takehana, T.; Taniguchi, I.; Yamaji, H.; Maeda, Y.; Toy-

ohara, K.; Miyamoto, K.; Kimura, Y.; Oda, K. A Bacterium that Degrades and Assimilates Poly(ethylene terephthalate). *Science* **2016**, *351*, 1196–1199.

For Table of Contents Only



Supporting information for: Cleavage Product Accumulation Decreases the Activity of Cutinase during PET Hydrolysis

Christine Groß,[*,†] Kay Hamacher,[†] Katja Schmitz,[‡] and Sven Jager[*,†]

[†]*Department of Biology, Computational Biology & Simulation Group, Technische Universität Darmstadt, Schnittspahnstraße 2, 64287 Darmstadt, Germany*

[‡]*Department of Chemistry, Biological Chemistry Group, Technische Universität Darmstadt, Alarich-Weiss-Straße 8, 64287 Darmstadt, Germany*

E-mail: c.gross@bio.tu-darmstadt.de; jager@bio.tu-darmstadt.de

Phone: +49 6151 16 20372. Fax: +49 6151 16 72772

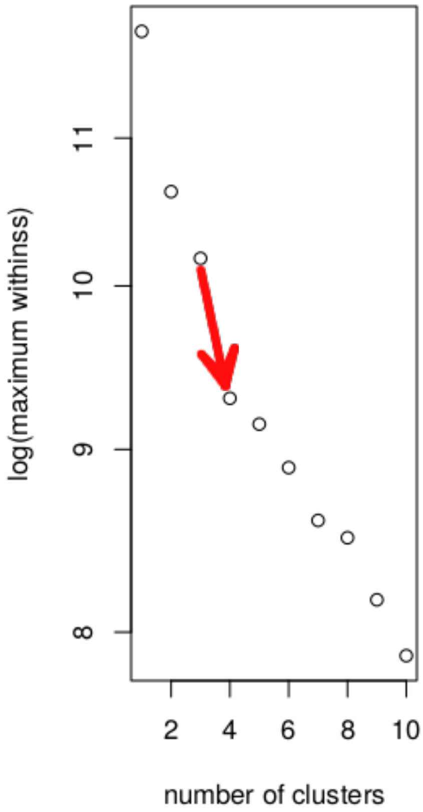


Figure S1: `kmeans` clustering of the 13Å null model with maximal within-cluster sum of squares (withinss) as a function of number of clusters with the red arrow highlighting the drastic drop using four instead of three clusters.

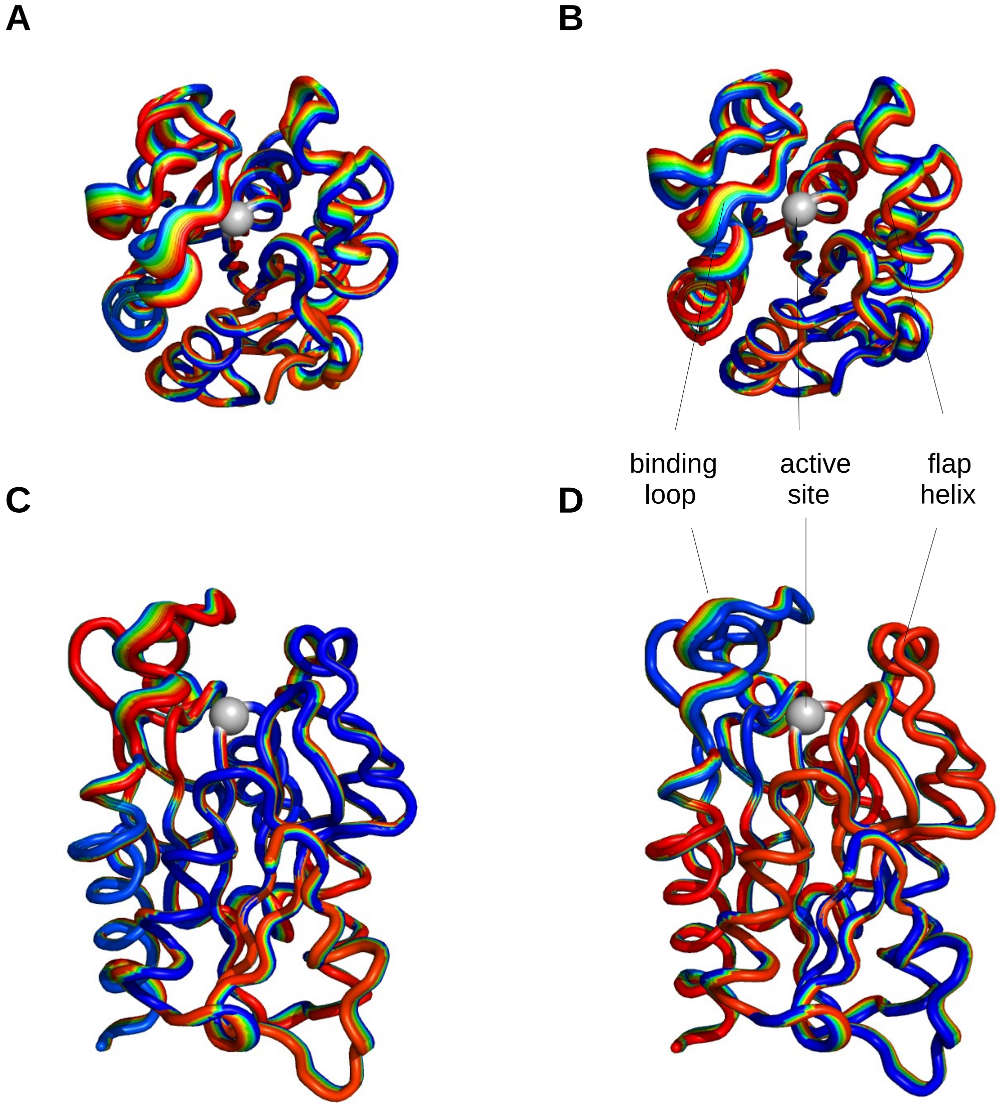


Figure S2: Lowest frequency normal mode of *Fs*C in one direction A + C and in the opposite direction B + D in top and side view (right angle to each other). In contrast to the "breath-like" motions after perturbation of S120 by substrate binding/cleavage, the normal mode loop motions are in a parallel manner (both loops move to the left or both loops move to the right).

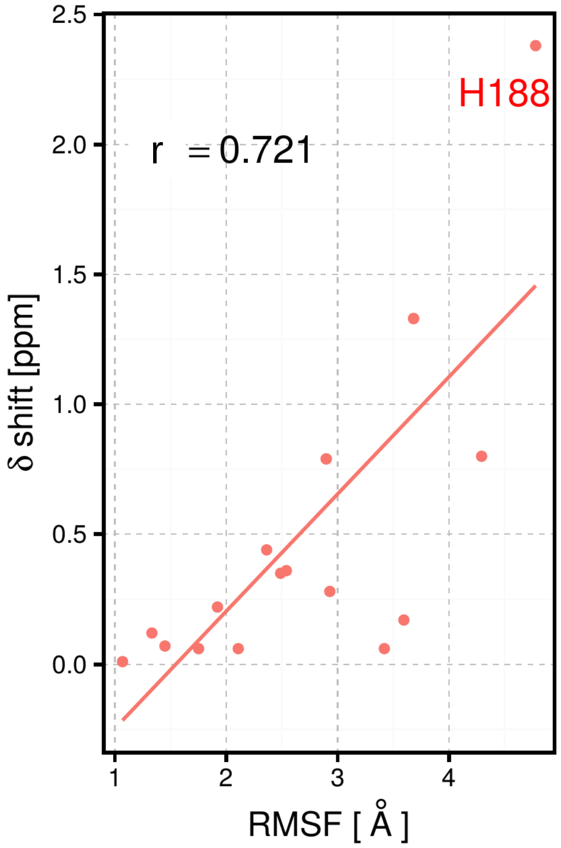


Figure S3: Comparison of NMR data with RMSF values of the wildtype simulation with Pearson's $r = 0.721$, p-value$= 0.00043$. H188 is labeled in red. δ shift values are obtained from Prompers *et al.*[S1]

For secondary structure quantification we used standardized secondary structure assignment, `Define Secondary Structure of Proteins` (short: `DSSP`). `DSSP` begins by identifying the intra-backbone hydrogen bonds of the protein using a purely electrostatic definition.[S2] We computed `DSSP` for each frame and computed the mean occurrence of secondary stucture elements of the complete production run of the simulation using `R`[S3] and `bio3d`.[S4]

We measured the following secondary structural states derived from `DSSP` (Figure S4 D):
B = residue in isolated β-bridge,
E = extended strand, participates in β ladder,
G = 3-helix (3_{10} helix),
H = α-helix,
I = 5 helix (π-helix),
S = bend,
T = hydrogen bonded turn,
U = loop region.

Furthermore, to see that the protein does not unfold during the simulation time, we plotted the percentages of α-helices, residues in isolated β-bridges, and bends over time (Figure S4 A-C).

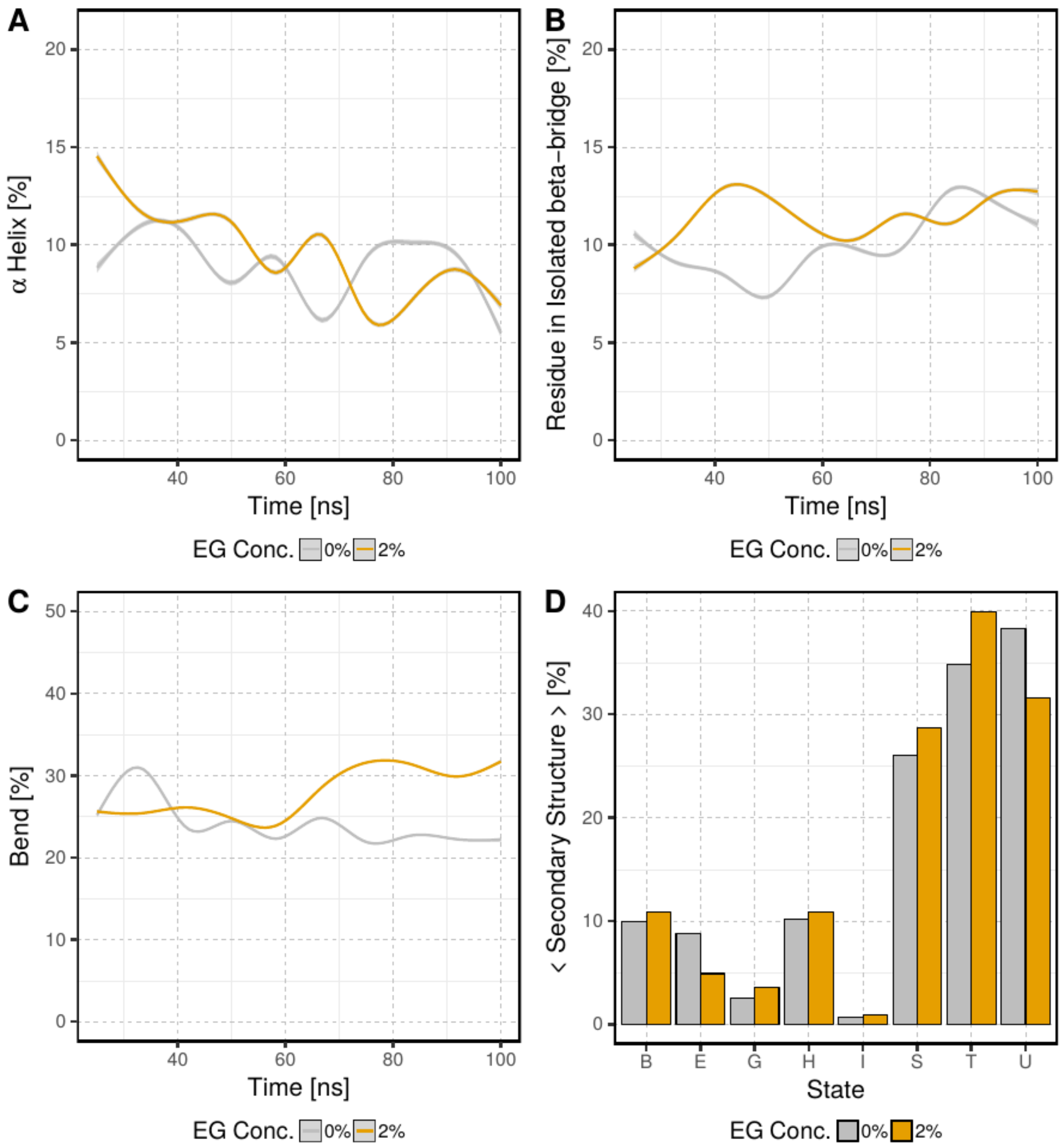


Figure S4: Secondary structure analysis for the MD simulations with 0% and 2% EG in the solvent. (A-C) Main secondary structure elements over time. The curves were smoothed using the LOESS[S5] algorithm. (D) Average percentages of all observed secondary structure elements.

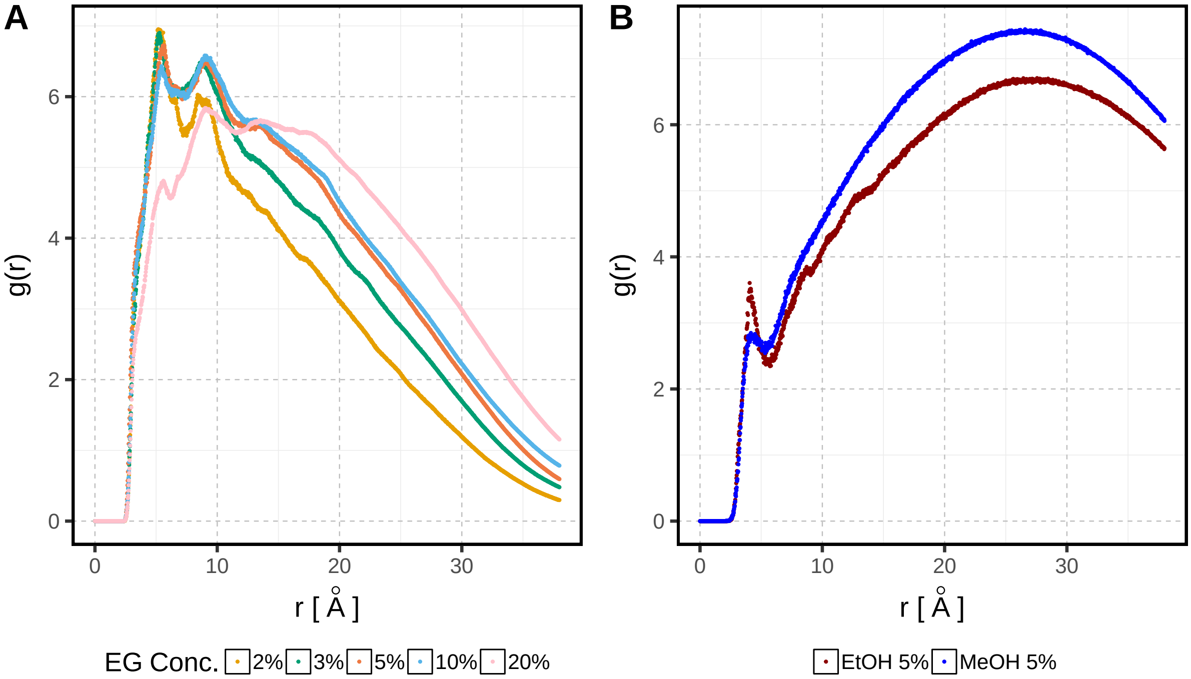


Figure S5: (A) Radial distribution function (RDF) of protein Cα to EG O for the simulations with different concentrations of EG in the solvent. For all EG concentrations we see a peak around 5-7 Å which is the first coordination shell of an amino acid followed by a second peak at around 10 Å. With further distances the radial distribution strongly decreases. These results indicate accumulation of EG near the *FsC* surface while the amount of EG in the bulk is minor. (B) RDF of protein Cα to EtOH O or MeOH O, respectively, for the simulations with other alcohols in the solvent. For both alcohols we see a tiny peak in the first coordination shell but in contrast to EG most of the EtOH and MeOH molecules are located in the bulk far away from the protein surface.

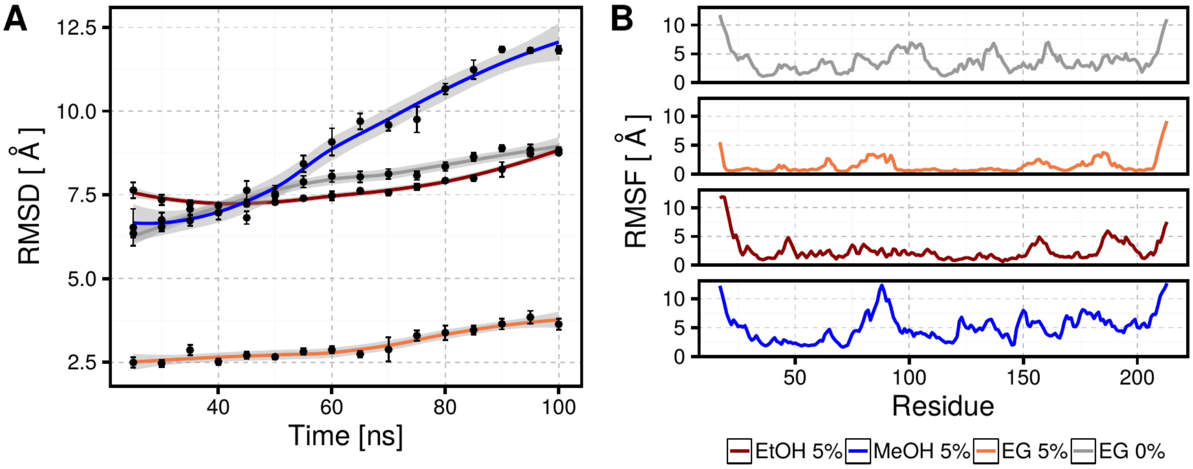


Figure S6: Comparison of RMSD (A) and RMSF (B) for 5% EG, 5% EtOH, and 5% MeOH in the solvent. To see the effect of the different alcohols in the solvent, the corresponding RMSD and RMSF without any alcohol (0% EG) is also plotted. The curves were smoothed using the LOESS[S5] algorithm.

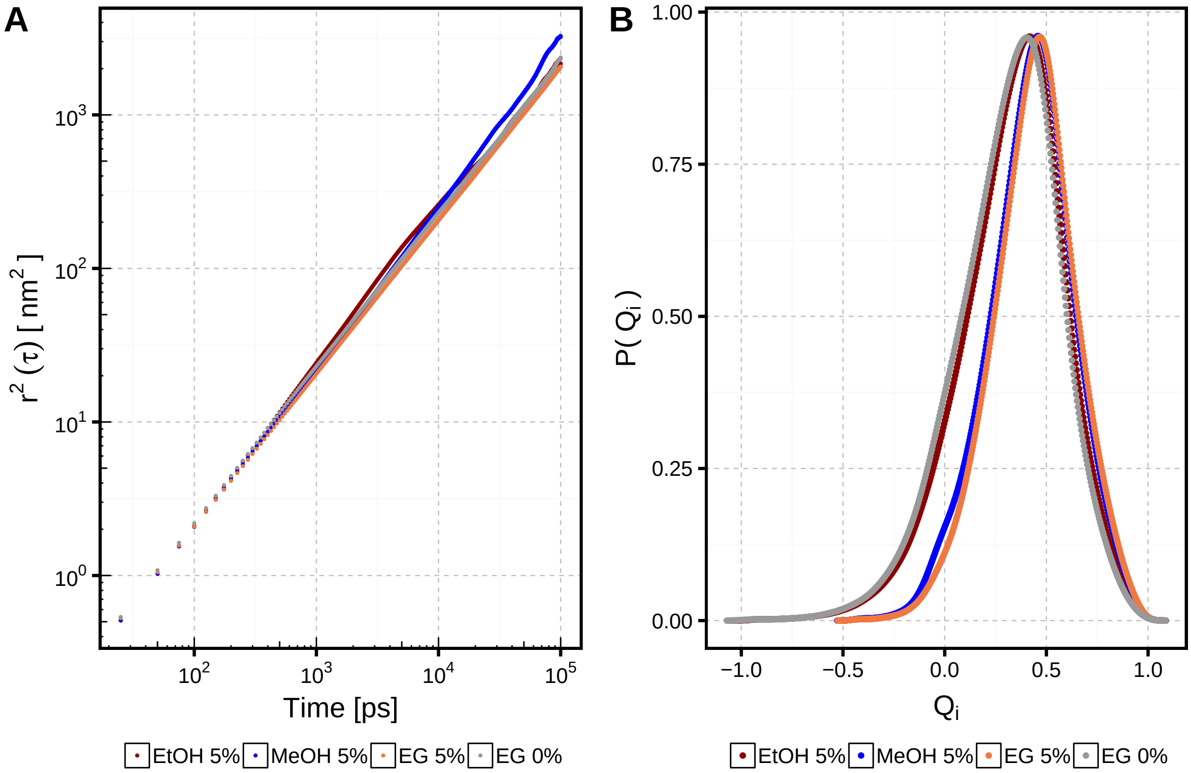


Figure S7: Characterization of the solvent for the simulations with 5% EtOH, 5% MeOH, and 5% EG in comparison to the simulation without any alcohol added to the solvent. (A) Double logarithmic representation of MSD of the water molecules as indirect measure of the viscosity. The constant increase of the MSD during the simulation describes normal diffusion. While the MSD of the simulation with EG is slightly decreased compared to the simulation with water only, the MSD for the simulation with MeOH is increased and the MSD of the simulation with EtOH behaves as with water only. (B) Tetrahedral order parameter Q_i to quantify the ability of water molecules to form hydrogen bonds to adjacent water molecules. The distribution of Q_i is right shifted for EG and MeOH while EtOH does not significantly influence the hydrogen bonding network of the water molecules.

To determine whether the interactions between EG and the surface residues are of hydrophobic or hydrophilic nature, we made surface density calculations based on Equations 7-11 with d set to 7 Å. The subsets of surface residues were grouped as follows:

active site residues:

41 42 81 84 119 120 121 150 175 177 183 184 185 186 188

remaining surface residues:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 17 18 24 27 28 29 30 31 32 33 34 35 36 37 38 44 45 46 47 48 49 50 51 52 53 59 60 61 62 63 64 65 66 67 68 69 70 71 72 75 76 77 79 80 83 85 86 87 88 89 90 91 92 94 95 96 97 105 114 115 116 117 118 122 123 124 126 127 128 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 151 152 153 154 156 157 158 159 160 162 163 164 165 166 167 168 169 170 171 172 173 174 176 178 179 180 181 182 187 189 190 191 192 193 194 195 196 197

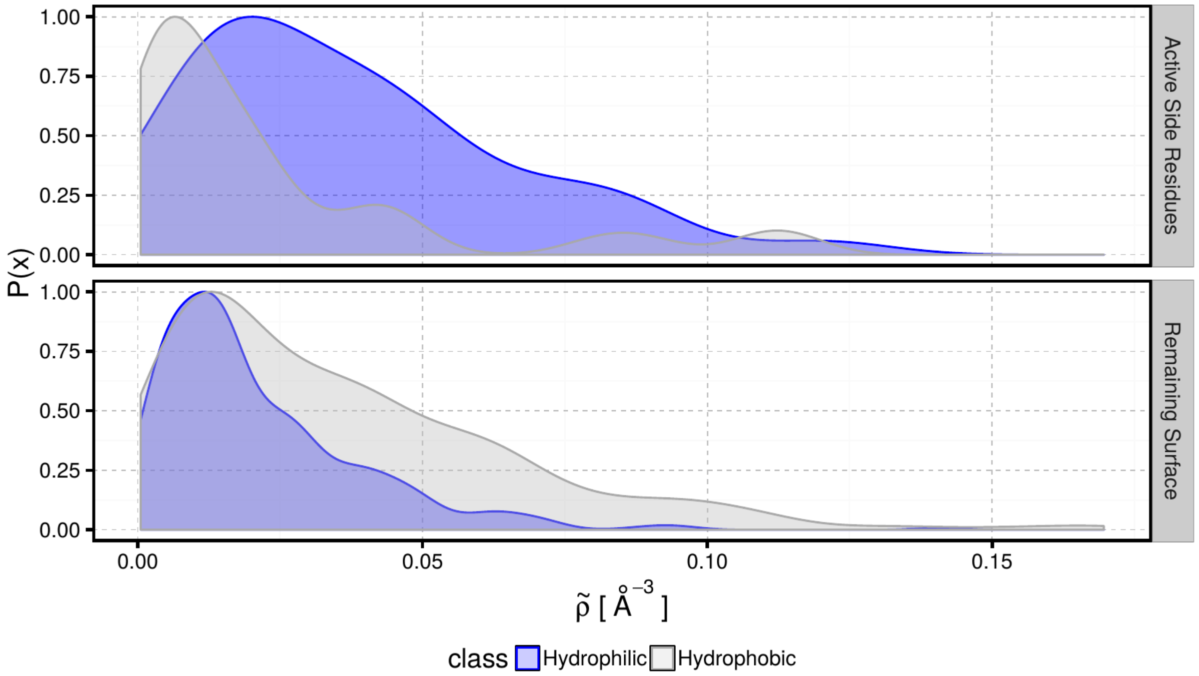


Figure S8: Distributions of EG densities at surface residues in the ative site in comparison to the EG densities at remaining surface residues. For both surface subsets the densities are seperately plotted for hydrophilic (blue) and hydrophobic (transparent) residues. While the EG densities at the remaining surface are quite similar for hydrophobic and hydrophilic residues, the densities of EG in the active site are significantly higher at hydrophilic residues. This indicates that the EG accumulation in the active site is mainly based on hydrogen-bonding interactions.

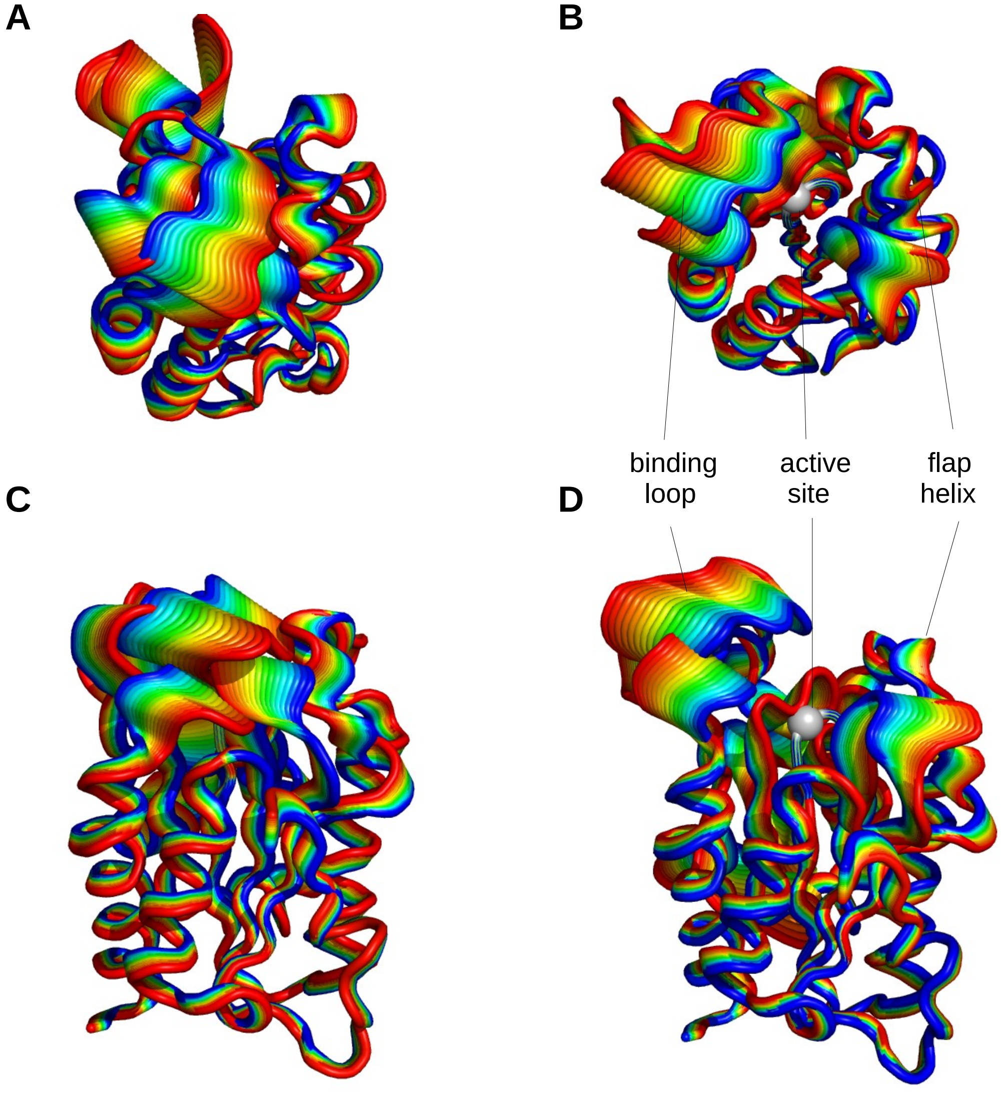


Figure S9: LRT model (cutoff 7Å) with repulsive A + C and attractive B + D force vectors with different forces analogous to the model with 13 Å in Figure 7. Here forces from 0 (blue) to 60 (red) in arbitrary units were applied.

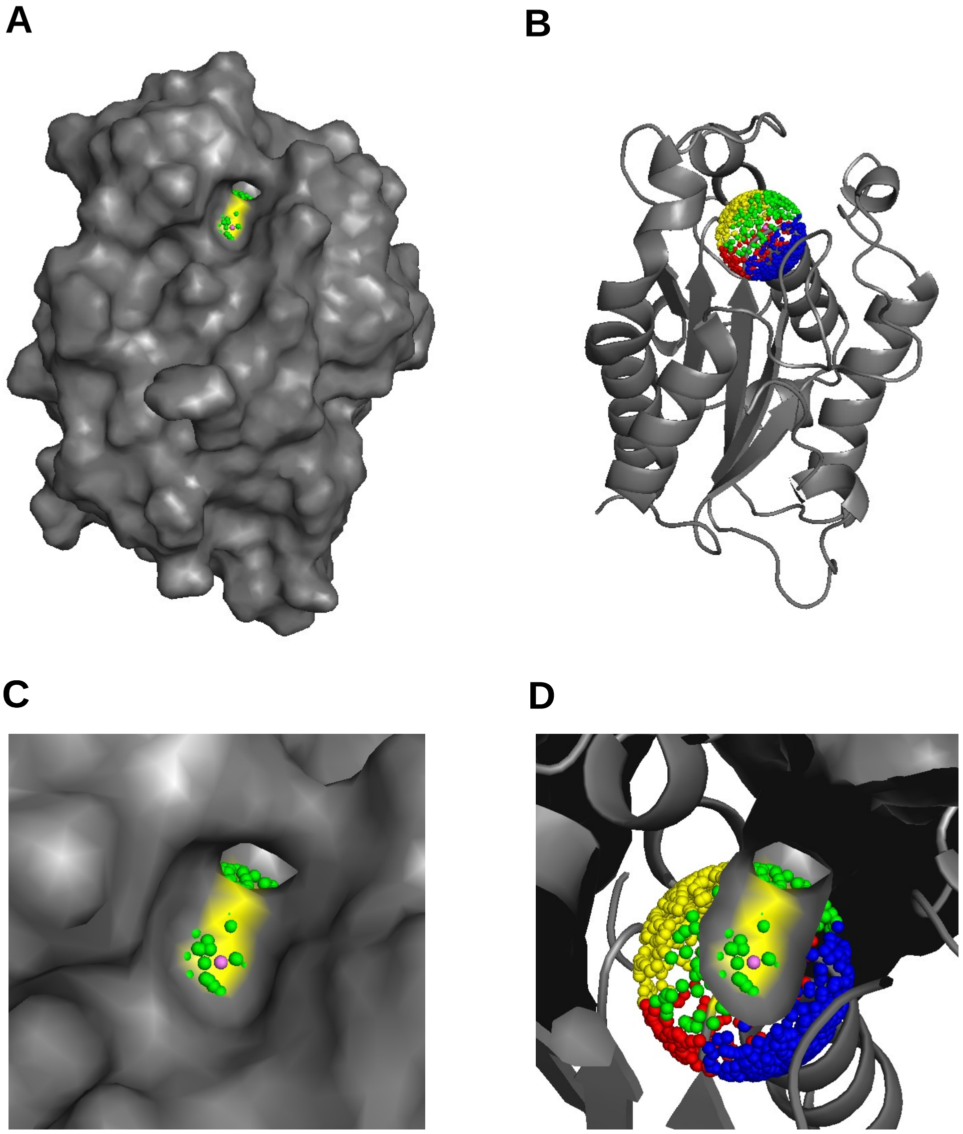
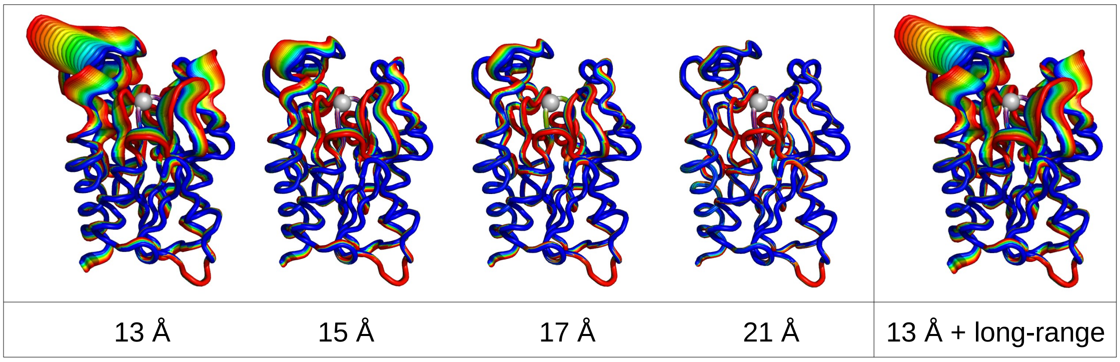


Figure S10: LRT null model (cutoff 7Å) with 1000 force directions for the external force vector analogous to the model with 13 Å in Figure 8.



13 Å 15 Å 17 Å 21 Å 13 Å + long-range

Figure S11: Comparison of the displacements of anisotropic network models with different cutoffs for connected residues to account for long-range interactions. The displacements only differ in their magnitude, which shows that the 13 Å modell does not disregard possible deviating long-range interactions. The 13 Å + long-range model was computed with decreasing interactions between residues with larger distances than 13 Å. Nevertheless, the displacement is the same as for the original 13 Å model.

References

(S1) Prompers, J. J.; Hilbers, C. W.; Groenewegen, A.; Schaik, R. C. V.; Pepermans, H. A. M. 1H, 13C, and 15N Resonance Assignments of *Fusarium solani pisi* Cutinase and Preliminary Features of the Structure in Solution. *Protein Sci.* **1997**, *6*, 2375–2384.

(S2) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577–2637.

(S3) R Development Core Team, R: A Language and Environment for Statistical Computing. 2008.

(S4) Grant, B. J.; Rodrigues, A. P. C.; ElSawy, K. M.; McCammon, J. A.; Caves, L. S. D. Bio3d: An R Package for the Comparative Analysis of Protein Structures. *Bioinformatics* **2006**, *22*, 2695–2696.

(S5) Cleveland, W. S. Robust Locally Weighted Regression and Smoothing Scatterplots. *J. Am. Stat. Assoc.* **1979**, *74*, 829–836.

5.2 Statistical Evaluation of HTS Assays for Enzymatic Hydrolysis of β-Keto Esters

The following manuscript:

- Buss, O.*,**Jager, S.***, Dold, S.-M., Zimmermann, S., Hamacher, K., Schmitz, K., & Rudat, J. (2016). Statistical Evaluation of HTS Assays for Enzymatic Hydrolysis of β-Keto Esters. PloS One, 11(1), e0146104.

presents a new method for selecting high-throughput assays for the development of drug precursors. The selection is based on a statistical evaluation of positive and negative controls of the assays to be analyzed. For this purpose, control reactions are carried out for the respective assay. These reactions are evaluated with different statistical distance metrics. The tests described above are performed before the actual experiment. Subsequently, each assay approach is assigned a performance. The assay with the highest performance will then be used for the screening. This straightforward procedure increases the success rate for screenings. We were able to demonstrate this experimentally using esterases as an example.

Contributions In this manuscript, I was responsible for creating the expression vectors for pNB-Est13 as well as FsC (esterase, cutinase) and for developing the expression protocol. Furthermore, I developed one of the presented assays and performed the statistical analysis of all involved assays. In accordance with the presented method I implemented a `R` library (`Assayvis`) and published it. I was also involved in the writing and conception of the manuscript. Furthermore, I was responsible for the content and the creation of Figure 2,3 and SI 1, SI 3. Kay Hamacher and Katja Schmitz helped to write the manuscript and improved it. Dold and Zimmerman helped to conduct the experiments. In this contribution, Buss and I share the first authorship.

PLOS | ONE

RESEARCH ARTICLE

Statistical Evaluation of HTS Assays for Enzymatic Hydrolysis of β-Keto Esters

O. Buß[3]☯*, S. Jager[1]☯*, S. -M. Dold[3], S. Zimmermann[2], K. Hamacher[1], K. Schmitz[4], J. Rudat[3]

1 Technische Universität Darmstadt, Computational Biology and Simulation, Darmstadt, Germany, **2** Karlsruhe Institute of Technology, Biomolecular Separation Engineering, Karlsruhe, Germany, **3** Karlsruhe Institute of Technology, Technical Biology, Karlsruhe, Germany, **4** Technische Universität Darmstadt, Biological Chemistry, Darmstadt, Germany

☯ These authors contributed equally to this work.
* oliver.buss@kit.edu (OB); jager@bio.tu-darmstadt.de (SJ)

Abstract

β-keto esters are used as precursors for the synthesis of *β*-amino acids, which are building blocks for some classes of pharmaceuticals. Here we describe the comparison of screening procedures for hydrolases to be used for the hydrolysis of *β*-keto esters, the first step in the preparation of *β*-amino acids. Two of the tested high throughput screening (HTS) assays depend on coupled enzymatic reactions which detect the alcohol released during ester hydrolysis by luminescence or absorption. The third assay detects the pH shift due to acid formation using an indicator dye. To choose the most efficient approach for screening, we assessed these assays with different statistical methods—namely, the classical Z'-factor, standardized mean difference (SSMD), the Kolmogorov-Smirnov-test, and *t*-statistics. This revealed that all three assays are suitable for HTS, the pH assay performing best. Based on our data we discuss the explanatory power of different statistical measures. Finally, we successfully employed the pH assay to identify a very fast hydrolase in an enzyme-substrate screening.

OPEN ACCESS

Citation: Buß O, Jager S, Dold S-M, Zimmermann S, Hamacher K, Schmitz K, et al. (2016) Statistical Evaluation of HTS Assays for Enzymatic Hydrolysis of β-Keto Esters. PLoS ONE 11(1): e0146104. doi:10.1371/journal.pone.0146104

Editor: Monika Oberer, University of Graz, AUSTRIA

Received: July 28, 2015

Accepted: December 14, 2015

Published: January 5, 2016

Copyright: © 2016 Buß et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Data are within the paper and its Supporting Information files. All additional files are available from https://zenodo.org/record/21536 (DOI 10.5281/zenodo.21536).

Funding: This work was supported by Federal Ministry of Education and Research (Germany) funding code 031A095B, project: Molecular Interaction Engineering MIE (http://www.forschung-mie.de/english/index.php). The authors acknowledge support by Deutsche Forschungsgemeinschaft and Open Access Publishing Fund of Karlsruhe Institute of Technology.

Introduction

β-amino acids are intermediates in the synthesis of a great variety of pharmaceutically important compounds [1, 2]. The objective of this study is to select an HTS assay to screen for one enzyme for a two-step reaction cascade for the synthesis of *β*-amino acids. These occur in a number of biologically active compounds such as paclitaxel, bleomycin and the lipopeptide YM-170320 [3, 4]. Paclitaxel is used as a drug in the treatment of certain types of cancers [5]. As the correct configuration of the stereocenters is essential for biological activity, synthesis strategies with high enantioselectivity are desirable [2].

A variety of synthesis strategies have been established for the production of chiral *β*-amino acids. However, the limitations of these strategies are unfavourable for industrial production [6, 7]. In addition to the chemical approaches, enzymatic strategies have been established to produce chiral *β*-amino acids. Since all enzymatic approaches established in a industrial scale

Competing Interests: The authors have declared that no competing interests exist.

are based on kinetic resolutions, their theoretical yield is limited to 50% at most [8]. Both the enzymatic and chemical synthesis strategies are still a subject of research for the production of β-amino acids [9]. In order to achieve higher yields than 50%, β-amino acids can also be obtained with high enantiomeric excess by enzymatic conversion of β-keto acids using transaminase [9]. However, the substrates of this synthesis strategy, the β-keto acids, are not stable and decarboxylate. To avoid this side reaction β-keto acids can be generated *in situ* by hydrolysis of a β-keto ester catalyzed by a hydrolase. In the synthesis of natural and non-natural substrates hydrolases are a beneficial and commonly used enzyme class [10] and a number of lipases and esterases are commercially available [11]. Many enzymes are well characterized, but often there is no perfect match between the requirements of an efficient catalysis reaction and the properties of the biocatalyst. Besides being able to acquire hydrolases commercially, there is the possibility to test hydrolases from different sources, such as biomolecular libraries containing purified enzymes, microorganisms from the environment or hydrolases variants from directed evolution and from in *silico* gene data basis. The limiting factor is the fast and reliable identification of the best fitting enzyme.

HTS-assays for hydrolases

While investment in both, experimental time and costs, can increase sample throughput in large screenings with standard analytical hardware [12, 13], this route can quickly become prohibitively expensive. High throughput assays permit simultaneous measurement of samples in 96- to 1536-well plates so that $10^5$to 10^7samples may be screended per day [14–16]. This way, large biomolecular libraries can be screened for optimal enzymes [16].

A wide variety of assays has been established for the detection of efficent hydrolases *in vivo* or *in vitro* [17, 18]. HTS assays for lipases and esterases have been extensively reviewed [19, 20]. HTS assays may directly detect the reaction products or convert these products for indirect detection. For example a simple indirect HTS-assay detects the pH-shift during a hydrolysis reaction by a pH-indicator molecule [21].

Enzyme activity can be detected by a number of parameters such as microbial growth or changes in spectral properties, temperature or electrochemical potential as well as by luminescence [17]. Chromatography methods are well suited for medium-throughput screenings [13]. The assessment of temperature change due to exothermic reactions has been reported, as a technically sophisticated measure requiring specialized infrared cameras, which are no standard laboratory equipment [22]. In direct HTS assays product formation is monitored by a change in a physical quantity associated with the product.

For this purpose, chromogenic and fluorogenic substrates are frequently applied in these assays [18]. Additional analytical reactions are also commonly used [18]. These mostly synthetic substrates allow for an easy readout by changes in absorbance, fluorescence or by chemiluminescence [12]. However, when a substrate analog with an artificial chromophore or chromogenic group is employed in the the optimization process instead of the substrate of interest this may lead to an enzyme optimized processing the analog but not the substrate of interest ("You get what you screen for" You and Arnold *et al.* 1996) [13, 19, 23, 24].

Disadvantages of assay based on non-natural fluorophores/chromophores

A variety of hydrolase screening assays is based on non-natural substrates with chromophoric groups. Well-known examples are umbelliferyl-, 4-nitrobenzofurazane- and 4-nitrophenyl-assays [25–27]. The disadvantage of this assays is caused by the difference between the properties of the substrate to the non-natural substrate. The most active enzymes in this kind of screening do not have to be the most active enzyme for the natural substrate. As an example, for screening hydrolases, 4-nitrophenyl esters are popular, because standard photometric plate readers can easily monitor the reaction by detection of the colored product at 348 to 405 nm [13]. However, 4-nitrophenyl esters *per se* are no substrate with industrial pertinence [28, 29]. Furthermore, the absorption of the liberated 4-nitrophenol at 405 nm depends on the pH-value, so that the pH value needs to be controlled. Alternatively, absorbance may be measured at the isosbestic point of 4-nitrophenol at 348 nm [30, 31]. Moreover, 4-nitrophenyl esters are more readily hydrolyzed than esters comprising aliphatic alcohol residues like methanol or ethanol. When 4-nitrophenyl esters are used to in hydrolase screening, the hits may therefore exhibit lower activity towards the actual substrate [19, 32]. In 2008, J. Córdova *et al.* [33] reported enzymatic activity of bovine serum albumin (BSA) showing high hydrolysis rates for 4-nitrophenyl esters at 80–160°C. Under these conditions, a spontaneous hydrolysis of 4-nitrophenyl esters is likely, so that the observed hydrolysis may not be necessarily due to an enzymatic activity of BSA. Another argument against the use of 4-nitrophenyl esters is the potential reaction with nucleophilic amino acid side chains as it was shown for the reaction of 4-nitrophenyl acetate with L-tyrosine esters by B.S. Hartley *et al.* in 1953 and the acetylation of insulin in the reaction with 4-nitrophenyl acetate [34]. Modification of the enzyme by acyl-group transfer may lead to artifacts during screening. Taken together, 4-nitrophenyl esters are are often not suitable as substrate analogues for screening. If possible, alternatives to the fluorophoric and chromophoric non-natural substrates should be used.

Indirect assays as an alternative approach

If it is impossible to directly detect the conversion of the native substrate, the alternative approach is to further convert the products for indirect detection. A simple approach are colorimetric pH-assays, which can be employed if one of the products is an acid that subsequently deprotonates or a base that lowers the proton concentration in the medium. This change in pH value can be monitored by an indicator dye [35, 36]. Enzymatically coupled systems that transform one of the products yielding a detectable compound are more complex. As more reaction steps and compounds required for quantification more parameters need to be optimized and the readout is more prone to errors. One of the first examples for a convenient indirect assay was the NADH-dependent, coupled enzyme assay for urease by Kaltwasser *et al.*(1966) [37].

For the hydrolysis of β-keto esters, we compared three different indirect assays for the activity of hydrolases. One assay relies on the change of the pH-value, the second is based on enzymatic oxidation of the released ethanol to acetic acid and the third, which we expected to be most sensitive, is based on the oxidation of ethanol to ethanal and hydrogen peroxide which is then converted by horseradish peroxidase (HRP) in a luminescence reaction. The first assay is the simplest one, because only a buffer- pH indicator system is needed and many examples are established for enzyme screening with pH-indicators [21, 35, 38, 39]. The second assay is also well-known and established for measurements of alcohol in food [40]. This assay was miniaturized to microtiter plates. The third assay is a modification of a chromogenic alcohol-oxidase (AOX)/peroxidase (HRP) ethanol assay for determination of ethanol in beverages, which normally based on 2,2'-azino-bis(3-ethylbenzthiazoline-6-sulphonic acid (ABTS) as chromogenic substrate [41]. This assay was modified to a luminometric assay by adding luminol instead of a chromogenic substrate. The luminometric measurements should be more sensitive due to photons are released by the detection reaction. For the first time the system was established in a flow system with separated bioreactors by Marschall and Gibson [42]. For the quantification of ethanol, we adapted the luminol-AOX-HRP system to microtiter plates by using design of experiments (DoE). The aim was to optimize the system for endpoint measurements of enzymatic hydrolysis reactions. Therefore, for the first time a luminometric AOX-HRP-system was tested for quantification of ethanol in 96-well plates. In this study all assays were tested for HTS compatibility.

Statistical evaluation methods

To compare the quality of high-throughput assays the Z'factor is frequently employed [14, 43]. In addition, the assessment of measurement procedures can be based on traditional statistic parameters like signal-to-noise-ratio(S/N), signal to background ratio (S/B), and coefficient of variation (CV) (Table 1). CV is the ratio of mean to variance and often used as quality control for assays [44]. The S/B-ratio is a criterion that indicates whether the level of the signal is sufficiently high above the background. The rule of thumb for a good HTS assay is S/B >3. However, fluctuations in both the signal and the background are not considered. In contrast, the S/N-ratio takes into account the standard deviation of the background. The higher the S/N ratio, the less do background fluctuation influence the desired signal. Both measures indicate whether a sample could be distinguished from the background, however they don't quantify to what extent the positive and negative controls can be distinguished.

Therefore methods have been explicitly developed to evaluate high-throughput screenings, like Z'-factor and the Strictly Standardized Mean Difference (SSMD) [45].

J.H. Zhang (1999) defined the Z'-factor based on the normal distribution and the 3-sigma rule of thumb [43, 46], which implies that 99.7% of all samples lie within less than three standard deviations distance from the mean. The Z'-factor has become a common metric for HTS quality control as it allows to decide whether an assay is suitable to just distinguish positive samples from negative ones (Z'>0,5) or whether it can distinguish well performing samples from poor ones (Z'>0,8) [14, 43, 47].

The strictly standardized mean difference (SSMD) takes the mean difference of negatives and positives in proportion to the standard deviation. SSMD gained recognition in screening e.g. for antiviral drugs [48]. In 2007, X.D. Zhang derived the SSMD under the condition of independence of both distributions through maximum likelihood estimation (MLE) and method of moment (MM). Finally, for increased robustness against outliers, the median of absolute deviation ($SSMD_R$) can be used.

In this work, we applied for the first time these different SSMDs for evaluation of a biocatalyst screening assay. SSMD have already been used in RNAi screening and cell-based systems

Table 1. Overview of statistical measures for evaluation of HTS assays. σ = standard deviation; μ = mean; m = median; pos = data set of positive controls; neg = data set of negative controls; n = sample size; F_k = cumulative density function for data set k; max_j = maximum distance between two distributions; SSMD = Strictly standardized mean difference; x_i = *ith* value of the (ordered) data set x.

Measures	Definition
Kolmogorov-Smirnov-test (KS-test)	$KS = \max_j(F_1(x_j) - F_2(x_j))$
t-statistic	$t = \frac{\mu_{pos} - \mu_{neg}}{\sqrt{\frac{\sigma^2_{pos}}{n_{pos}} - \frac{\sigma^2_{neg}}{n_{neg}}}}$
Z'-factor	$Z = 1 - 3\frac{\sigma_{pos} - \sigma_{neg}}{\mu_{pos} - \mu_{neg}}$
$SSMD_{MM}$ (method of moment)	$\beta_{MM} = \frac{\mu_{pos} - \mu_{neg}}{\sqrt{\sigma^2_{pos} - \sigma^2_{neg}}}$
$SSMD_{ML}$ (maximum likelihood)	$\beta_{MLE} = \frac{\mu_{pos} - \mu_{neg}}{\sqrt{\sigma^2_{pos}\frac{n_{pos}-1}{n_{pos}} + \sigma^2_{neg}\frac{n_{neg}-1}{n_{neg}}}}$
$SSMD_R$ (robust)	$MAD = 1.4826m(\lvert x_i - m \rvert)$
	$\beta_R = \frac{m_{pos} - m_{neg}}{\sqrt{MAD^2_{pos} - MAD^2_{neg}}}$

doi:10.1371/journal.pone.0146104.t001

or for quantification of enzyme inhibition [49, 50] but never for assays of enzyme catalysis. We compared these measures to the Z'factor.

We also chose non-parametric methods like the KS-test and *t*-statistic. Non-parametric test statistic differs from parametric statistic in that no explicit distribution is given, but the procedure is solely based on the data without the need for any model. The *t*-statistic describes the difference of two sample distributions measured in standard errors of both means. This is suitable for small sample sizes, especially when the underlying distribution is unknown or not derivable [51].

In contrast to all other methods, the Kolmogorow-Smirnow-Test (KS-test) [52] calculates the maximum distance between two cumulative density functions (CDF). The CDF can be computed for finite number of data points as a cumulative sum of the frequency of occurrence of the (sorted) data. The unique benefit of this kind of calculation is that no previous knowledge about distribution or models is needed. We include the KS test here to contrast it with the other measures that do not rely on any CDF.

Materials and Methods

All enzymes and chemicals used in the assays were purchased from Sigma-Aldrich (St. Louis, USA) and Carl Roth (Karlsruhe, Germany). Reagents were dissolved in 40 mM potassium phosphate buffer (pH 7.2) unless stated otherwise. For all buffers and solutions deionized water was used.

Expression and purification of para-nitrobenzyl-esterase 13 (pNB-Est 13)

Expression was performed in *E. coli* BL 21 codonplus using a pET-22b vector system with a pelB-leader sequence for export to the periplasm of *E.coli*. After transformation cells were cultivated at 180 rpm and 30°C in shaking flasks in 400 mL LB-medium. After the OD600 had reached 0.4, expression was induced with isopropyl β-D-1-thiogalactopyranoside (IPTG, Carl Roth, final concentration 0.2 mM). The expression proceeded over 15 h at 20°C. The cells were harvested by centrifugation at 4700 rpm for 20 min at 4°C (Heraeus Multifuge X3 FR). Protein purification from the periplasm was carried out by osmotic shock in ddH_2O, using the protocol by Petersen *et al.* [53]. The supernatant was frozen in liquid nitrogen and lyophilized overnight (Lyophilizer, Beta 1–8 Martin Christ). The product was analyzed by SDS-polyacrylamide-gel electrophoresis (SDS-PAGE, S2 Fig).

Oxidative luminescence assay

Using the standard conditions of the AOX-HRP-ABTS system (described in the protocol of Sigma-Aldrich [54]) the luminescence signal was very weak for ethanol (2.0 mM) [55, 56]. To gain a much longer and more intensive luminescence reaction a design of experiments with a statistic based optimization program (Modde 10.1 Software (Umetrics, Sweden)) was carried out (S4 Fig and S5 Fig). For measuring luminescence, white 96-well plates (Greiner, Austria) were used. The final reaction volume per well was 200 μL. Reaction mixtures were prepared on a Tecan Freedom Evo 200 liquid handling station (LHS). The LHS is equipped with one eight-port liquid handling arm (LiHa), a standard robotic plate handling arm (RoMa) equipped with a centric gripper, a 96-channel liquid handling arm (MultiChannelArmTM (MCA 96), Tecan) with an eccentric gripper, and an integrated spectrophotometer (Infinite M200 Pro, Tecan). Pipetting precision and accuracy of aqueous solutions was determined by pipetting on an analytical balance as described by Oelmeier *et al.*, 2010 [57]. A variation coefficient of less than 1.6% was determined for volumes between 20 and 1,000 μL. The concentrations of luminol, HRP, AOX and ethanol were varied in DoE to search for the maximum of luminescence intensity (slope). For quantification of ethanol in samples different dilutions of ethanol (0 to 2 mM) were added to the assay mixture. Finally a 1:8330 dilution of AOX (10–40 U/mg, Sigma-Aldrich) containing of 2 μg of HRP (150 U/mg, Sigma-Aldrich) were added to the assay mixture, containing 0.5 mM luminol (Sigma-Aldrich) dissolved in 5% (v/v) DMSO (Carl Roth) and 95% 40 mM sodium phosphate buffer, pH 8.0. For statistical evaluation, 2.0 mM ethanol was used as a positive control without hydrolase and ester substrate (see also subsection statistical analysis).

All ingredients, except HRP, were mixed at room temperature and then incubated under shaking at 150 rpm at 30°C for 15 min. The reaction was started by adding 15 μL of HRP solution (total activity 0.3 U) to the reaction mixture and mixing with a 96-tip automatic pipette of the evo workstation. Afterwards the 96-well plates were placed in the Infinite M200 Pro reader to measure luminescence. For each measure point luminescence intensity was integrated over 350 ms. Each well was repeatedly measured for up to 1 h. The working temperature was 30°C +/- 2°C. The luminescence raw data was processed either as mean or as numeric integral over the duration of the experiment.

Oxidative photometric assay

To determine the ethanol concentration the commercial ethanol kit from R-Biopharm AG (Germany) was used. According to the manufacturer, this assay is carried out in a total volume of 3 mL measured in cuvettes at 340 nm in an ordinary spectrophotometer. The manufacturer's instructions were adapted to the 96-well plate format using 100 μL per well maintaining the ratio of compounds. The assay was started by addition of a 1:10 dilution of aldehyde dehydrogenase solution. The plate was briefly centrifuged to eliminate bubbles. The ethanol concentration was determined by an ethanol calibration curve in the range from 0 to 3 mM (S3 Fig). According to the other assays a positive control for the evaluation test contained only 2.0 mM ethanol without ester and hydrolase (see also subsection statistical analysis).

pH indicator assay for endpoint measurements

The protocol from Moris-Varas *et al.* was adapted [38] for endpoint measurements. A weak 2.5 sodium phosphate buffer ($pK_{a2} = 7.2$) was adjusted to pH 7.0. Bromothymol blue (pK_a = 7.1) [58, 59] was dissolved under heating and stirring in this buffer to yield a 0.54 mM stock solution. The assay system contained 10% (v/v) indicator stock solution. According to the other assays a positive control for the evaluation contained only 2.0 mM HCl without ester and hydrolase (see also subsection statistical analysis). Absorbance was measured at the absorption maxima of bromothymol blue at 440 nm and 620 nm in transparent 96-well plates in a conventional plate reader (Epoch, Biotek).

pH indicator assay for enzymatic activity

The hydrolytic activity was also determined based on β-keto acid formation as an indicator of product formation using the pH indicator bromothymol blue. For the determination of activity of hydrolases substrate concentration was 2.0 mM. The calibrations were done by adding different concentrations of HCl (0 to 2 mM) to the buffer in the presence of each tested ester to determine a calibration line (S3 Fig) [39]. During the experiment the pH range was between pH 6.0 to pH 7.0. In this range, the 3-oxo-3-phenylpropanoic acid (β-keto acid) is almost completely dissociated. The calculated $pK_a$for 3-oxo-3-phenylpropanoic acid is 3.56 [60]. For comparison of hydrolases equal masses of protein from 10 mg/mL stock solutions were used. The enzyme concentration of the stock solution was tested by Bradford assay. For each enzyme the blank was determined in buffer with bromothymol blue and without substrate. The incubation temperature was 30°C and 2 mM (concentration *in situ*) ethyl benzoyl acetate was added to the reaction mixture containing hyrolase (see Table 2). The specific activity $[mol/(min \cdot mg)]$ was calculated by the slope at the beginning of the reaction at 620 nm. The specific activity was converted by the number of active sites into the turnover number [1/s].

Statistical analysis

To evaluate the quality of all proposed assays, we used the equations shown in Table 1. For the pH-assay and spectroscopic ethanol-assay μ_{pos} was the mean absorbance of the positive controls and μ_{neg} that of the negative controls. In case of the oxidative luminescence assay μ_{pos} was the mean intensity or integral of the positive controls and μ_{neg} of the negative controls. For the pH indicator assay the positive control contained 2 mM HCl in 5 mM sodium phosphate buffer (pH 7.2), and the negative control was made from plain buffer. For the two different ethanol based assays, the positive control consisted of assay buffer and 2 mM ethanol, while plain assay buffer was used for the negative controls. Between 44 and 48 positives and negatives were measured for each assay. The measurements took place in 96-well plates. SSMD, student *t*-statistic as well as the KS-statistic and other metrics were computed in `R` [69]For the *t*-statistic we used the t.test function with Welch correction and for the KS test the ks.test function in `R`. All other metrics were implemented in `R`. All plots were created using the `ggplot2` library [70]. For *t*-statistics the Welch correction has been designed to account for unequal variances of both groups (positive and negative) [71]. The criteria for evaluation of the assays based on the statistic measures are listed in Table 3.

Software contribution

`R` is an environment for interactive analysis of statistical data in bioinformatics offering many additional software packages. We combined all approaches and implemented the `Assay-Toolbox` package in `R` to make the applied methods accessible to a wide community. Software link: http://www.cbs.tu-darmstadt.de/htsassay.zip

Results

Upon hydrolysis of β-keto ethylesters, ethanol is released and a β-keto acid is formed. Afterwards the acids can decarboxylate to carbon dioxide and acetophenone. A small amount of

Table 2. Enzymes for screening. For comparison of enzyme activity equal protein concentrations (mg/mL) were used. The concentrations of all hydrolase solutions were determined by Bradford assay. Stock solutions of hydrolases were 10 mg/mL. All enzymes but pNB-Est13 were were purchased comerially. pNB-Est13 was hetrologous expressed in *E. coli*.

Abbreviation	Hydrolase type	Origin	Molecular weigth
PPL	Lipase	*Porcine pancreas*	50 kDa [61]
TLL	Lipase	*Thermomyces lanuginosus*	30 kDa [62]
RML	Lipase	*Rhizomucor miehei*	29 kDa [63]
CRL	Lipase	*Candida rugosa*	60 kDa [64]
ALBC	Amano lipase PS	*Burkholderia cepacia*	28 kDa [65]
ALPF	Amano lipase	*Pseudomonas fluorescens*	32 kDa [66]
ALM	Amano lipase M	*Mucor javanicus*	21 kDa [67]
pNB-Est13*	Esterase M	*Bacillus licheniformis*	55 kDa [68]
HRP	Peroxidase	*Armoracia rusticana*	
AOX	Alcohol oxidase	*Pichia pastoris*	

doi:10.1371/journal.pone.0146104.t002

Table 3. Criteria for performance evaluation of HTS-assays [43, 47, 51, 71, 86, 87].

Z'-factor	*t*-statistic	SSMD	Performance
$0.8 \leq Z' \leq 1.0$	Null-hypothesis rejected	SSMD ≥ 3.0	excellent assay
$0.5 \leq Z' \leq 0.8$			good assay
$0.5 \leq Z' \leq 0.0$			weak assay
$0.0 \leq Z'$	Null-hypothesis accepted	SSMD ≤ 3.0	"yes/no" type assay

doi:10.1371/journal.pone.0146104.t003

carbon dioxide forms carbonic acid in water, which lowers the pH-value as well [72] (Fig 1). We used the pH shift (Fig 1(B)) visualized by the indicator bromothymol blue as a measure of product formation. We choose this indicator due to its expected pH value of the enzyme coupled system consisting of lipase and transaminase. In addition we created a new luminometric assay (Fig 1(C)) for use in 96-well plates. This assays detects the formation of the product ethanol by an enzymatic reaction. Ethanol was oxidized by alcohol-oxidase to yield acetaldehyde and hydrogen peroxide, which is used to oxidize luminol to 3-aminophthalate by horseradish peroxidase (Fig 1(C)). This well-known chemiluminescent reaction is frequently used in immunoassays such as Western Blot or ELISA [73]. For the luminescence-based assay we expected a higher sensitivity as the emitted light can be accumulated (high signal) and there is no stray light from excitation (low background) [74]. Other alcohols can also be oxidized by alcohol oxidase [75] so that this assay would be applicable to the hydrolysis of different types of esters. For comparison we used the commercial ethanol assay from r-Biopharm, based on the oxidation of ethanol to acetic acid in a two-step oxidation by alcohol dehydrogenase and aldehyde dehydrogenase [40]. In this reaction cascade, two equivalents of NADH are generated, which can be detected by UV/Vis spectroscopy at 340 nm. The aim was to identify the best assay out of these three, i.e. the one with the highest dynamic range and most stable readout. The system also has to be sensitive enough to operate reliably with small quantities of substrate and enzymes to save costs in high-throughput screening.

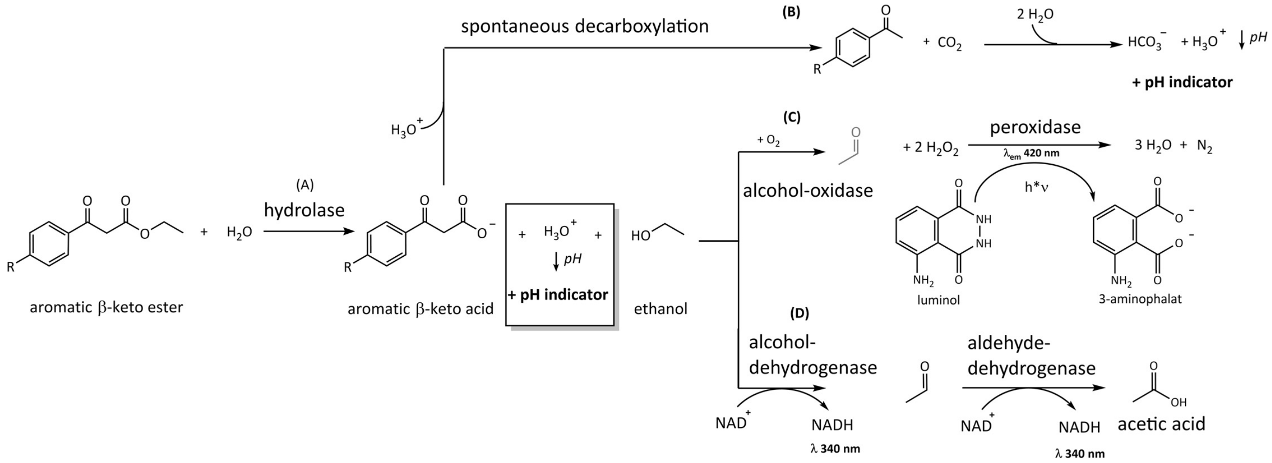


Fig 1. Overview of assays for hydrolysis of *β*-keto esters. (A) Hydrolysis reaction of *β*-keto ethylester catalyzed by hydrolase. (B) pH-assay: photometric detection of pH change due to acid formation and deprotonation with bromothymol blue. Ethanol based assays: (C) Oxidative luminescence assay using alcohol oxidase, horseradish peroxidase and ethanol (D) Photometric detection of ethanol by oxidation by dehydrogenases under conversion of $NAD^+$to NADH.

doi:10.1371/journal.pone.0146104.g001

The first step was to optimize all assays by testing different enzyme and compound concentrations without hydrolases, with HCl or ethanol as signal inducing molecules, mostly for the luminescence assay. For comparison of the three HTS assays, endpoint measurements were done. The two step oxidative luminescence assay was optimized by design of experiments (DoE) to maximize the duration of luminescence and the sensitivity for endpoint measurements. The isochronic induction of luminescence in all wells was very important for the reliable quantification of ethanol. The assay mixtures were automatically pipetted by the Tecan evo pipetting robot that can simultaneously pipette 96 wells. It turned out that a pre-incubation time of 15 min without peroxidase was necessary. Without pre-incubation, the resulting luminescence signal was too weak for the quantification of ethanol. Luminescence enhancers like 4-iodophenol [76], were added in order to maximize the luminescence output, however no sufficient signal amplification was observed so that these additives were omitted. Due to the relatively weak signal the luminescence assay was only suitable for endpoint determinations. For quantification, we calculated both the numeric integral and the mean of luminescence over the measured time. A clear luminescence signal was detectable for about 50 min.

Likewise, the commercial spectrometric ethanol assay was only suitable for endpoint determinations when carried out in 96-well plates. The assay was tested for kinetic measurements with the result that no correlation between substrate concentration and reaction rate of hydrolase was observed. The pre-assembled commercial assay might not be suitable for kinetics, because the concentration of NAD and alcohol-dehydrogenase can not separately be varied. In contrast to these two assays, the pH indicator assay has no additional enzymes. The pH indicator, bromothymol blue, has two absorption maxima in the UV-Vis spectrum at 440 and 620 nm and a pK_aat the desired pH value for the cascade reaction of a hydrolase in combination with a ω-transaminase. The extinction coefficient at both wavelengths depends on the pH-value so that both were used for evaluation measurements.

To evaluate these assays, we performed simple tests with positive and negative controls without the real substrate and hydrolases. One-half of each plate was filled with the positive control and the other half with the negative control. For the pH-assay, we used 2 mM HCl as positive control, which corresponds to the maximum product concentration, if all substrate was converted. For evaluation of the oxidative luminescence and spectrometric ethanol assay, we added 2.0 mM ethanol as positive control for the tested enzyme coupled assays. Reaction buffer without substrate and hydrolase was used as the negative control for all assays. In Fig 2 the distribution of negative and positive controls for each assay is shown. For evaluation we used the criteria for evaluation listed in Table 3.

The KS-statistic value was almost equal for all assays, because all distributions seem to have a maximal distance. KS-statistic shows that all positives and negatives have a clearly separated distribution, but it cannot answer the question whether the degree of separation is acceptable for HTS. Consequently KS-statistic test is not necessarily helpful to compare the performance of these types of assays. Z'-factor, *t*-statistic and SSMD allowed for a more detailed evaluation: Due to the low performance in all statistical tests the Lum(int) and pH-based assay at 440 nm were considered unsuitable and were rejected. The pH-based assay at 620 nm performed best, because it has the largest dynamic range (Fig 2(b)) and thus is a reliable test for the applicability of our statistical measures.

SSMD yielded similar results as the Z'-factor and student *t*-test, showing that this measure, that was developed for RNAi screens, can be employed for the evaluation of enzymatic reactions (Fig 3). Robust $SSMD_R$ differs from all metrics for the pH-based assay at 440 nm which indicates that outliers exist for this assay. The negative consequence of outliers can be avoided, if at least triplicates are measured. In contrast to all other measures $SSMD_R$ is suitable for selection of sensitive assays with more outliers. On the one hand, additional metrics like the

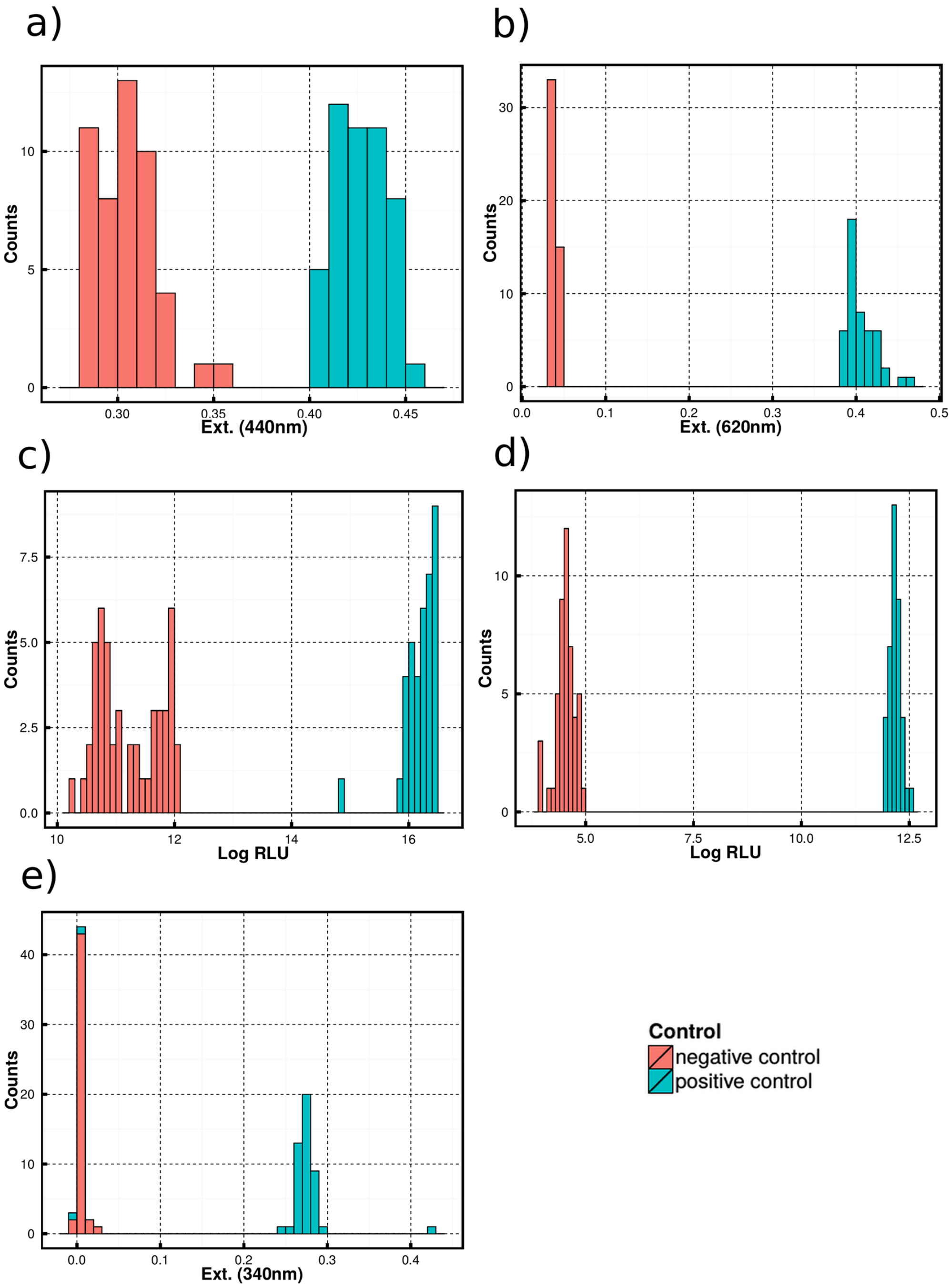
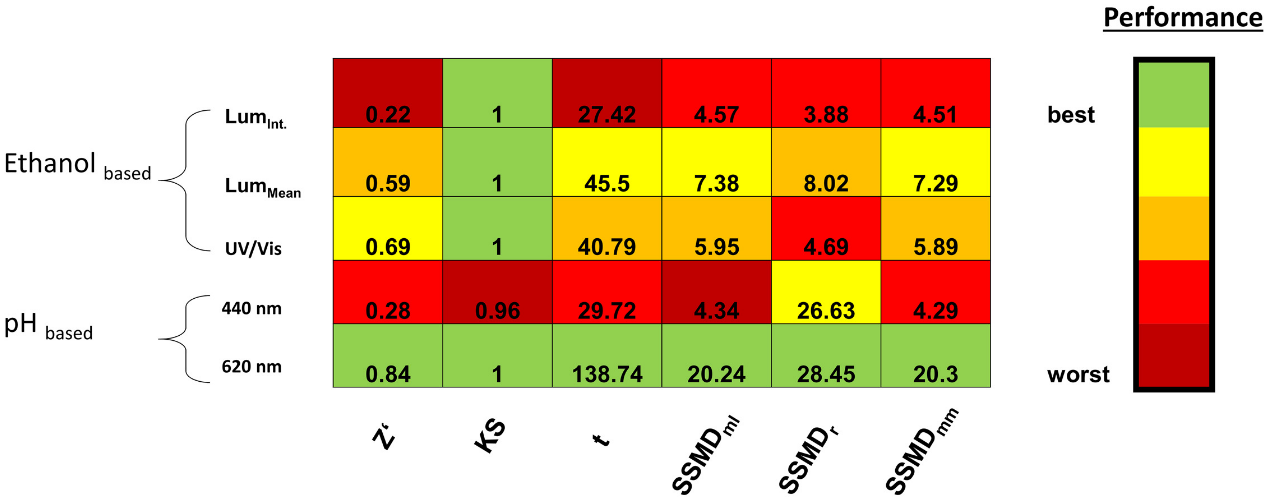


Fig 2. Histogram of positive and negative controls of different HTS assays. For each control 44–48 values were measured a) pH-indicator assay at 440 nm b) pH-indicator assay at 620 nm c) luminescence ethanol assay (mean luminescence intensity) d) luminescence ethanol assay (integrated luminescence intensity) e) photometric ethanol assay at 340 nm.

doi:10.1371/journal.pone.0146104.g002



		Z'	KS	t	$SSMD_{ml}$	$SSMD_r$	$SSMD_{mm}$
Ethanol based	$Lum_{Int.}$	0.22	1	27.42	4.57	3.88	4.51
	Lum_{Mean}	0.59	1	45.5	7.38	8.02	7.29
	UV/Vis	0.69	1	40.79	5.95	4.69	5.89
pH based	440 nm	0.28	0.96	29.72	4.34	26.63	4.29
	620 nm	0.84	1	138.74	20.24	28.45	20.3

Performance
best
worst

Fig 3. Matrix of different statistical parameters for evaluation of HTS assays. For each parameter the assays were ranked from the best (green) to the worst assay (red). The assays were grouped by the kind of detection. Ethanol quantification: Lum(int), Lum(mean), UV/Vis (ethanol dehydrogenase assay). pH-indicator assay at 440 and at 620 nm.

doi:10.1371/journal.pone.0146104.g003

mentioned $SSMD_R$ are particular useful in case of assays with outliers, because in some screening approaches they have a minor influence on the evaluation. However on the other hand statistics depend on explanatory power of the experimental data. As a conclusion of this, it should be carefully considered which metrics as well as processed raw data are suitable for a certain HTS evaluation.

As the pH indicator assay at 620 nm obtained the highest statistical scores it was chosen for a subsequent small scale screening of a set of different hydrolases with different derivates of ethyl benzoyl acetate (EBA; Fig 4). First of all, we determined for each ester a calibration curve with HCl concentrations between 0 and 2 mM at 620 nm (S3 Fig). The strong acid HCl (pK_a = -6) nearly dissociates to 100% in a wide range of the pH-scale. In contrast the weak β-keto acid (pK_a = 3.56) can only completely dissociate, when the pH is clearly above the pK_a. A weak sodium phosphate buffer (pH 7.0) was very important to of the pH change during the reaction. The range of pH-values during the experiment was between pH 6 and pH 7. This was the prerequisite for the calculation of the enzyme activity at the beginning of an activity test by the absorbance over the time. The concentrations were calculated by the different calibration curves of each substrate.

We tested each hydrolase in combination with each ester to determine which hydrolases are the most active ones and which substrates have the highest accessibility (Fig 4).

When adding the enzymes to the reaction buffer containing bromothymol blue we observed a color change for Amano lipase. However, upon substrate addition we detected no change in pH-value, so we assume that interactions between the enzyme and the pH-indicator, a polyaromatic molecule, may have interfered with the reaction or its colorimetric detection. Therefore, it was not possible to screen the activity of Amano lipase of *Pseudomonas fluorescens*(ALFP). Such interfering effects of indicator molecules and enzymes are well known, see for instance Banyai [12, 77]. An alternative option in pH-indicator screening might be the use of 4-nitrophenol (pK_a = 7.16) as an indicator molecule. It has only one aromatic ring and is therefore

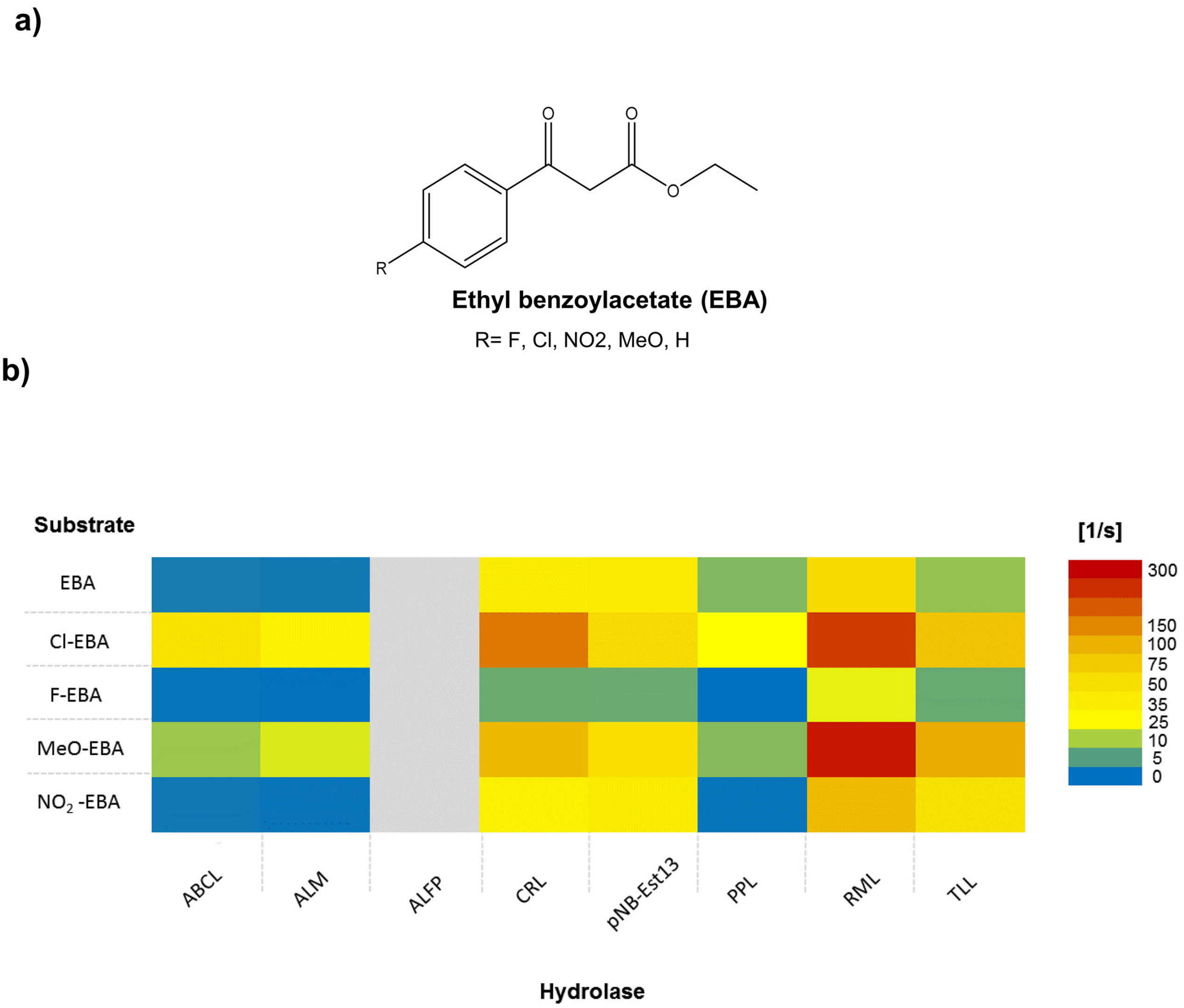


Fig 4. Screening of different β-keto esters against different hydrolases with pH-indicator assay. (for abbreviations see Table 3) The concentration of substrates was 2.0 mM in 2.5 mM sodium phosphate buffer (pH 7.0). The reactions were carried out at 30°C for 30 min. The activities (μmol/min) were normalized to μmol actives sites per s. (grey: no measurement possible) Para-nitrobenzyl-esterase 13 (pNB-Est13) was purified by us (described in methods). a) Structure of ethyl benzoylacetate with different substituents used as substrates in the screening. b) Activity matrix for 8 different enzymes (ABCL, ALM, ALFP, CRL, pNB-Est13, PPL, RML and TLL) against the respective substrates. The turnover number is in [1/s], for illustration turnover values were color coded from blue (low) to red (high).

doi:10.1371/journal.pone.0146104.g004

less hydrophobic [77, 78]. None of the other hydrolases showed any change in absorbance with only the pH indicator in the absence of substrate.

The experiment revealed that every substrate was converted by hydrolases. The lipase from *Rhizomucor miehei*(RML) showed the highest activity of the tested enzymes while the porcine pancreas lipase mix (PPL) showed the lowest hydrolysis activity towards aromatic β-keto esters. The only esterase in the screening (pNB-Est13), showed the fourth highest activity for all substrates. This shows that esterases may be considered as alternatives to lipases. This is of interest as the activity of lipases depends on hydrophobic surface interactions so that the substrate spectrum is limited. Esterases could help to broaden the substrate range for enzymatic β-keto acid and β-amino acid synthesis towards more hydrophilic compounds. Taken together, the assay results confirm, that the pH-assay at 620 nm can indeed be used for substrate-hydrolase screenings, albeit other indicator dyes may have to be tested to reduce non-specific interactions.

To explain the high activity of RML against all substrates, we consider the previous work of Rehm *et al.* [79]. Thus we compared the active site volumes and structural conformations like the opened as well as closed state of RML,*Candida rugosa* lipase (CRL) and TLL [79]. For example, in water, the lipase active site is covered by a mobile element, the lid, which opens upon substrate binding due to the hydrophobic interface [63]. In contrast to RML and TLL, CRL possesses a large and complex lid (residues 66–92), consisting of a short and a long α-helix. A comparison of the closed and open crystal structure of CRL revealed that the lid has to refold partially (S1b Fig). As expected for this reason, a slower opening and closing when compared to RML and TLL could be shown using Molecular Dynamics [79]. In contrast to CRL, a fast rigid body movement of the lid was suggested for RML and TLL. This opening event is suggested to be the rate limiting step during catalysis [64]. In addition several studies proposed that conformational rearrangement during the catalysis could have a much greater influence on the activity than binding energy inside the substrate pocket [80–82]. The discussion on the influence of the dynamic on the activity of the enzyme is still on going [83]. Although CRL possesses a even bigger lid than RML and TLL, the same approximated active site volume was calculated for all three lipases (RML: 59.5 nm^3; CRL 60.3 nm^3; TLL 57.3 nm^3, S1 Fig)

Beside this, the substrate properties of the ethyl benzoyl acetate substituents might have an influence on the activity caused by polarization and steric differences.

Conclusion

We compared three different assays based on a set of positive and negative controls by end point measurements. For this purpose we applied for the first time an alcohol oxidase-peroxidase-luminescence assay for ethanol quantification in 96-well format. Additionally, we evaluated a pH-indicator and a commercial ethanol assay for HTS. We applied several statistical measures for biocatalysis assay evaluation and found that classical Z'-factor, SSMD and *t*-statistic can indicate whether an assay is suitable for HTS. The pH-indicator assay based on color change of bromothymol blue at 620 nm upon acid formation performed best. However, strictly considered is the most accurate way to evaluate HTS assays to test the complete coupled assays (consisting of: enzyme(s) of interest, substrate(s) and assay compounds), then all necessary assay compounds, substrates and catalysts have to be tested in combination with each other. Otherwise possible cross-interactions of all assay substances can not be excluded when not the complete system is tested. But this approach is not suitable for a screening with the variation of more than one compound, caused by the exponentially growing number of experiments when two or more different compounds are tested. A further obstacle for evaluation of HTS assays by the whole testing system is that the hydrolases have to be previously inactivated in presence of the substrate for an end-point measurement. Moreover, the limiting factors for an evaluation are often enzymes as well as substrates availability (e.g. environmental samples). Therefore, a robust reduced evaluation approach might be suitable and weighed against potential benefits of a complete coupled system. In addition the non-coupled evaluation system could be also adapted for more complex screenings e.g. in case for enzymes from microbial lysates. To accomplish this a standard lysate from the expression host could be included into the evaluation setup. Beside this, using the pH assay in a screen of a panel of mostly commercially available lipases we identified the RML as the most efficient enzyme for the hydrolysis of the tested set of aromatic β-keto ethyl esters. We were able to explain those findings by structural models of the enzymes' binding pocket and lid as well as by comparison of our data with descriptive literature on lipases dynamics, including MD.

Furthermore, we demonstrated that the esterase pNB-Est-13 is also suitable for aromatic β-keto ester hydrolysis. This may help to broaden the substrate spectrum since esterases may accept more hydrophilic substrates than lipases.

Outlook

The pH-assay is a very useful method in the search for efficient hydrolases for a cascade reaction as it is robust, inexpensive and allows to record reaction kinetics. Our best-performing hydrolase, RML, could be used in combination with ω-transaminase, in cascade reactions for the synthesis of β-amino acids. All tested β-keto esters could be utilized for cascade reaction with hydrolase and ω-transaminase [84]. The newly proposed enzymatic oxidative luminescence assay, still requires further optimization. However, it bears the potential for a more sensitive assay, due to the cumulative measurement and the lower background compared to fluorescence and absorption measurements [85].

Beside this, we want to extend our modelling efforts to validate our hypothesis for the activity of the RML and to make predictions for the activity lipases and esterases like the pNB-Est-13, with a broad substrate tolerance.

Supporting Information

S1 Fig. Structural investigation of RML and CRL. a) Surface representation of CRL open (grey, 1CRL) and closed (blue, 1THG) state. Structural alignment of CRL open (grey, 1CRL) and closed (blue, 1THG). b) Surface representation of RML open (grey, 4TGL) and closed (blue, 3TGL) state. Structural alignment of RML open (grey, 4TGL) and closed (blue, 3TGL). c) Cavity volume (blue) of CRL in profile and front view. d) Cavity volume (blue) of RML in profile and front view.
(TIF)

S2 Fig. pNB-Est13 esterase compared by SDS-PAGE. Equal volumes with the same protein concentration were analyzed. The separation was carried in in a 12.5% SDS- polyacrylamide-gel at 200 V. Crude pNB-Est13 esterase was used after osmotic shock with TES-buffer and ddH$_2$O.
(TIF)

S3 Fig. Calibration of HTS-assays. The error bars show standard deviation. (a) Spectrometric ethanol assay (Pearson: 0.998, p-value = $1.64 \cdot 10^{-8}$) (b) pH-indicator assay at 620 nm (Pearson: -0.999, p-value = $2.47 \cdot 10^{-8}$). (c) Mean Luminescence ethanol assay (Pearson: 0.967, p-value = $8.15 \cdot 10^{3}$) a) and b) were measured in triplicate. The assays are describe in the methods section.
(TIF)

S4 Fig. First step optimization of the luminescence assay, by varying the concentration of luminol/ethanol (a) and luminol/hydrogen peroxide (b) against the average luminescence intensity over the time. The surface model plot was created by Modde 10.1.
(TIF)

S5 Fig. Contour plot to optimize the luminescence assay. The ethanol concentration (0 to 2.5 mM) was plotted against the HRP concentration (0 to 0.87 U/mL),AOX concentration (0 to 18.5 mU/mL)and luminol concentration (0 μM;37.5 μM;75 μM). Was scaled Contrasting the activity. A) luminescence slope per s: blue = high activity; red = low / no activity. (B) averaged luminescence: red = high activity blue = low / no activity. The plot was created by Modde 10.1.
(TIF)

Acknowledgments

The authors would like to thank for the German Federal Ministry of Education and Research (BMBF) for funding within the molecular interaction engineering (MIE) project (funding code 031A095B). In addition, we gratefully acknowledge Christoph Syldatk at Karlsruhe Institute of Technology (KIT), Technical Biology, for the provision of laboratory materials.

Author Contributions

Conceived and designed the experiments: OB. Performed the experiments: OB. Analyzed the data: OB SJ. Contributed reagents/materials/analysis tools: SJ SZ JR. Wrote the paper: OB SJ KS KH JR. Supervision of lab work: SMD.

References

1. Kudo F, Miyanaga A, Eguchi T. Biosynthesis of natural products containing [small beta]-amino acids. Natural Product Reports. 2014; 31(8):1056–1073. doi: 10.1039/C4NP00007B PMID: 24926851
2. Liljeblad A, Kanerva LT. Biocatalysis as a profound tool in the preparation of highly enantiopure β-amino acids. Tetrahedron. 2006 Jun; 62(25):5831–5854. Available from: http://www.sciencedirect.com/science/article/pii/S0040402006005412. doi: 10.1016/j.tet.2006.03.109
3. Sugawara T, Tanaka A, Tanaka K, Nagai K, Suzuki K, Suzuki T, et al. YM-170320, a Novel Lipopeptide Antibiotic Inducing Morphological Change of Colonies in a Mutant of Candida tropicalis pK233. The Journal of Antibiotics(Tokyo). 1998; 51(4):435–438. doi: 10.7164/antibiotics.51.435
4. Umezawa H, Maeda K, Takeuchi T OY. New antibiotics, bleomycin A and B. The Journal of Antibiotics (Tokyo). 1966; 19:200–209.
5. Rowinsky EK, Cazenave LA, Donehower RC. Taxol: A Novel Investigational Antimicrotubule Agent. Journal of the National Cancer Institute. 1990 Aug; 82(15):1247–1259. doi: 10.1093/jnci/82.15.1247 PMID: 1973737
6. Liu M, Sibi MP. Recent advances in the stereoselective synthesis of β-amino acids. Tetrahedron. 2002; 58(40):7991–8035. doi: 10.1016/S0040-4020(02)00991-2
7. Weiner B, Szymanski W, Janssen DB, Minnaard AJ, Feringa BL. Recent advances in the catalytic asymmetric synthesis of β-amino acids. Chem Soc Rev. 2010; 39:1656–1691. doi: 10.1039/b919599h PMID: 20419214
8. Tasnádi G, Forró E, Fülöp F. An efficient new enzymatic method for the preparation of β-aryl-β-amino acid enantiomers. Tetrahedron: Asymmetry. 2008; 19(17):2072–2077. Available from: http://www.sciencedirect.com/science/article/pii/S0957416608005697. doi: 10.1016/j.tetasy.2008.08.009
9. Rudat J, Brucher BR, Syldatk C. Transaminases for the synthesis of enantiopure β-amino acids. AMB Express. 2012 Jan; 2(1):11. doi: 10.1186/2191-0855-2-11 PMID: 22293122
10. Crout DHG, Christen M. Biotransformations in Organic Synthesis. In: Modern Synthetic Methods 1989 SE—1. vol. 5 of Modern Synthetic Methods. Springer Berlin Heidelberg; 1989. p. 1–114. Available from: http://dx.doi.org/10.1007/978-3-642-83758-6_1.
11. Daiha KdG, Angeli R, de Oliveira SD, Almeida RV. Are Lipases Still Important Biocatalysts? A Study of Scientific Publications and Patents for Technological Forecasting. PLoS ONE. 2015 06; 10(6): e0131624. doi: 10.1371/journal.pone.0131624
12. Kazlauskas RJ. Quantitative Assay of Hydrolases for Activity and Selectivity Using Color Changes. In: Enzyme Assays. Wiley-VCH Verlag GmbH & Co. KGaA; 2005. p. 15–39. Available from: http://dx.doi.org/10.1002/3527607846.ch1.

13. Reetz MT. High-throughput Screening Systems for Assaying the Enantioselectivity of Enzymes. In: Enzyme Assays. Wiley-VCH Verlag GmbH & Co. KGaA; 2005. p. 41–76. Available from: http://dx.doi.org/10.1002/3527607846.ch2.
14. Inglese J, Johnson RL, Simeonov A, Xia M, Zheng W, Austin CP, et al. High-throughput screening assays for the identification of chemical probes. Nature Chemical Biology. 2007 Aug; 3(8):466–479. doi: 10.1038/nchembio.2007.17 PMID: 17637779
15. Carroll SS, Inglese J, Mao SS, Olsen DB. Drug Screening: Assay Development Issues. In: Molecular Cancer Therapeutics. John Wiley & Sons, Inc.; 2004. p. 119–140. Available from: http://dx.doi.org/10.1002/047165616X.ch7.
16. Reymond JL, Fluxa VS, Maillard N. Enzyme assays. Chemical Communications. 2008;p. -46. Available from: http://dx.doi.org/10.1039/B813732C.
17. Sicard R, Reymond JL. Introduction. In: Enzyme Assays. Wiley-VCH Verlag GmbH & Co. KGaA; 2005. p. 1–14. Available from: http://dx.doi.org/10.1002/3527607846.ch.
18. Schmidt M, Bornscheuer UT. High-throughput assays for lipases and esterases. Biomolecular Engineering. 2005; 22(1–3):51–56. Directed Enzyme Evolution. Available from: http://www.sciencedirect.com/science/article/pii/S1389034405000109. doi: 10.1016/j.bioeng.2004.09.004 PMID: 15857783
19. Bornscheuer UT. High-Throughput-Screening Systems for Hydrolases. Engineering in life sciences. 2004; 4(6):539–542. doi: 10.1002/elsc.200402157
20. Beisson F, Tiss A, Rivière C, Verger R. Methods for lipase detection and assay: a critical review. European Journal of Lipid Science and Technology. 2000; 102(2):133–153. doi: 10.1002/(SICI)1438-9312(200002)102:2%3C133::AID-EJLT133%3E3.0.CO;2-X
21. Bornscheuer UT, Altenbuchner J, Meyer HH. Directed evolution of an esterase: screening of enzyme libraries based on ph-Indicators and a growth assay. Bioorganic & Medicinal Chemistry. 1999; 7(10):2169–2173. doi: 10.1016/S0968-0896(99)00147-9
22. Reetz M, Hermes M, Becker M. Infrared-thermographic screening of the activity and enantioselectivity of enzymes. Applied Microbiology and Biotechnology. 2001; 55(5):531–536. doi: 10.1007/s002530100597 PMID: 11414316
23. You L, Arnold FH. Directed evolution of subtilisin E in Bacillus subtilis to enhance total activity in aqueous dimethylformamide. Protein Engineering. 1996; 9(1):77–83. doi: 10.1093/protein/9.1.77 PMID: 9053906
24. Schmidt-Dannert C, Arnold FH. Directed evolution of industrial enzymes. Trends in Biotechnology. 1999; 17(4):135–136. doi: 10.1016/S0167-7799(98)01283-9 PMID: 10203769
25. Demirjian D, Shah P, Morís-Varas F. Screening for Novel Enzymes. In: Fessner WD, Archelas A, Demirjian DC, Furstoss R, Griengl H, Jaeger KE, et al., editors. Biocatalysis—From Discovery to Application. vol. 200 of Topics in Current Chemistry. Springer Berlin Heidelberg; 1999. p. 1–29. Available from: http://dx.doi.org/10.1007/3-540-68116-7_1.
26. Goddard JP, Reymond JL. Enzyme assays for high-throughput screening. Current Opinion in Biotechnology. 2004; 15(4):314–322. Available from: http://www.sciencedirect.com/science/article/pii/S0958166904000874. doi: 10.1016/j.copbio.2004.06.008 PMID: 15358001
27. Reetz MT, Zonta A, Schimossek K, Jaeger KE, Liebeton K. Creation of Enantioselective Biocatalysts for Organic Chemistry by In Vitro Evolution. Angewandte Chemie International Edition in English. 1997; 36(24):2830–2832. doi: 10.1002/anie.199728301
28. Menger FM, Ladika M. Origin of rate accelerations in an enzyme model: the p-nitrophenyl ester syndrome. Journal of the American Chemical Society. 1987 May; 109(10):3145–3146. doi: 10.1021/ja00244a047
29. Levine MN, Lavis LD, Raines RT. Trimethyl Lock: A Stable Chromogenic Substrate for Esterases. Molecules. 2008; 13(2):204–211. doi: 10.3390/molecules13020204 PMID: 18305412
30. Main AR, Miles KE, Braid PE. The determination of human-serum-cholinesterase activity with o-nitrophenyl butyrate. Biochemical Journal. 1961 Apr; 78(4):769–776. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1205470/. doi: 10.1042/bj0780769 PMID: 13765461
31. Hotta Y, Ezaki S, Atomi H, Imanaka T. Extremely Stable and Versatile Carboxylesterase from a Hyperthermophilic Archaeon. Applied and Environmental Microbiology. 2002 Aug; 68(8):3925–3931. doi: 10.1128/AEM.68.8.3925-3931.2002 PMID: 12147492
32. Moore J, Arnold F. Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents. Nature biotechnology. 1996 April; 14(4):458–467. doi: 10.1038/nbt0496-458 PMID: 9630920
33. Córdova J, Ryan JD, Boonyaratanakornkit BB, Clark DS. Esterase activity of bovine serum albumin up to 160°C: A new benchmark for biocatalysis. Enzyme and Microbial Technology. 2008 Feb; 42(3):278–283. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0141022907003353. doi: 10.1016/j.enzmictec.2007.10.007

34. Hartley BS, Kilby BA. The reaction of p-nitrophenyl esters with chymotrypsin and insulin. Biochemical Journal. 1954 Feb; 56(2):288–297. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1269615/. doi: 10.1042/bj0560288 PMID: 13140189

35. Persson M, Palcic MM. A high-throughput pH indicator assay for screening glycosyltransferase saturation mutagenesis libraries. Analytical biochemistry. 2008 Jul; 378(1):1–7. Available from: http://www.sciencedirect.com/science/article/pii/S0003269708001358. doi: 10.1016/j.ab.2008.03.006 PMID: 18405657

36. Pratt R, Faraci W, Govardhan C. A direct spectrophotometric assay for d-alanine carboxypeptidases and for the esterase activity of β-lactamases. Analytical biochemistry. 1985; 206:204–206. doi: 10.1016/0003-2697(85)90106-X

37. Kaltwasser H, Schlegel HG. NADH-dependent coupled enzyme assay for urease and other ammonia-producing systems. Analytical Biochemistry. 1966 Jul; 16(1):132–138. Available from: http://www.sciencedirect.com/science/article/pii/0003269766900881. doi: 10.1016/0003-2697(66)90088-1 PMID: 4290701

38. Moris-Varas F, Shah A, Aikens J, Nadkarni NP, Rozzell JD, Demirjian DC. Visualization of enzyme-catalyzed reactions using pH indicators: rapid screening of hydrolase libraries and estimation of the enantioselectivity. Bioorganic & Medicinal Chemistry. 1999 Oct; 7(10):2183–2188. doi: 10.1016/S0968-0896(99)00149-2

39. He N, Yi D, Fessner WD.Flexibility of Substrate Binding of Cytosine-5'-Monophosphate-N-Acetylneuraminate Synthetase (CMP-Sialate Synthetase) from Neisseria meningitidis: An Enabling Catalyst for the Synthesis of Neo-sialoconjugates. Advanced Synthesis & Catalysis. 2011; 353(13):2384–2398. doi: 10.1002/adsc.201100412

40. Beutler HO, Michal G. Neue Methode zur enzymatischen Bestimmung von Äthanol in Lebensmitteln. Fresenius' Zeitschrift für analytische Chemie. 1977; 284(2):113–117. doi: 10.1007/BF00447345

41. Azevedo AM, Prazeres DMF, Cabral JMS, Fonseca LP. Ethanol biosensors based on alcohol oxidase. Biosensors and Bioelectronics. 2005; 21(2):235–247. Available from: http://www.sciencedirect.com/science/article/pii/S0956566304004889. doi: 10.1016/j.bios.2004.09.030 PMID: 16023950

42. Marshall RW, Gibson TD. Determination of sub-nanomole amounts of hydrogen peroxide using an immobilized enzyme flow cell. Application to the determination of ethanol. Analytica Chimica Acta. 1992; 266(2):309–315. doi: 10.1016/0003-2670(92)85057-D

43. Zhang JH. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. Journal of Biomolecular Screening. 1999 Apr; 4(2):67–73. doi: 10.1177/108705719900400206 PMID: 10838414

44. Rodbard D. Statistical Quality Control and Routine Data Processing for Radioimmunoassays and Immunoradiometric Assays. Clinical Chemistry. 1974 Oct; 20(10):1255–1270. Available from: http://www.clinchem.org/content/20/10/1255.abstract. PMID: 4370388

45. Zhang XD. A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays. Genomics. 2007 Apr; 89(4):552–561. Available from: http://www.sciencedirect.com/science/article/pii/S0888754307000079. doi: 10.1016/j.ygeno.2006.12.014 PMID: 17276655

46. Vysochanskij DF, Petunin YI. Justification of the 3σ rule for unimodal distributions. Theory of Probability and Mathematical Statistics. 1980; 21(25–36).

47. Walla PJ. Assay Development, Readers and High-Throughput Screening. In: Modern Biophysical Chemistry. Wiley-VCH Verlag GmbH & Co. KGaA; 2014. p. 323–338. Available from: http://dx.doi.org/10.1002/9783527683505.ch13.

48. Cao J, Forrest JC, Zhang X. A screen of the NIH Clinical Collection small molecule library identifies potential anti-coronavirus drugs. Antiviral Research. 2015 Feb; 114(0):1–10. Available from: http://www.sciencedirect.com/science/article/pii/S0166354214003313. doi: 10.1016/j.antiviral.2014.11.010 PMID: 25451075

49. Andruska N, Mao C, Cherian M, Zhang C, Shapiro DJ. Evaluation of a luciferase-based reporter assay as a screen for inhibitors of estrogen-ER α-induced proliferation of breast cancer cells. Journal of biomolecular screening. 2012 Aug; 17(7):921–32. Available from: http://jbx.sagepub.com/cgi/content/long/17/7/921. doi: 10.1177/1087057112442960 PMID: 22498909

50. Zhou H, Xu M, Huang Q, Gates AT, Zhang XD, Castle JC, et al. Genome-scale RNAi screen for host factors required for HIV replication. Cell host & microbe. 2008 Nov; 4(5):495–504. Available from: http://www.sciencedirect.com/science/article/pii/S1931312808003302. doi: 10.1016/j.chom.2008.10.004

51. Student. The Probable Error of a Mean. Biometrika. 1908 Mar; 6(1):1–25 CR—Copyright 169; 1908 Biometrika Trust. Available from: http://www.jstor.org/stable/2331554. doi: 10.1093/biomet/6.1.1

52. Smirnov NV. On the Estimation of the Discrepancy between Empirical Curves of Distribution for two Independent Samples. Bulletin de l'Université d'État a Moscou Série internationale Section A, Mathématique et mécanique. 1939; 2:3–14.

53. Petersen S, Fojan P. The thermal stability of the Fusarium solani pisi cutinase as a function of pH. Journal of Biomedicine and Biotechnology. 2001 Jan; 1(2):62–69. doi: 10.1155/S1110724301000249 PMID: 12488611

54. Sigma-Aldrich; 2015. Available from: http://www.sigmaaldrich.com/technical-documents/protocols/biology/enzymatic-assay-of-alcohol-oxidase.html.

55. Janssen FW, Ruelius HW. Alcohol oxidase, a flavoprotein from several basidiomycetes species. Biochimica et Biophysica Acta (BBA)—Enzymology. 1968; 151(2):330–342. doi: 10.1016/0005-2744(68)90100-9

56. Keesey J. Biochemica information. Boehringer Mannheim Biochemicals, Indianapolis. 1987;p. 56–59.

57. Oelmeier SA, Dismer F, Hubbuch J. Application of an aqueous two-phase systems high-throughput screening method to evaluate mAb HCP separation. Biotechnology and Bioengineering. 2011; 108(1):69–81. doi: 10.1002/bit.22900 PMID: 20717969

58. Phosphorsäure. In: Römpp Online. Georg Thieme Verlag KG; 2003.

59. Sigma-Aldrich; 2015. Available from: http://www.sigmaaldrich.com/catalog/product/sial/114413?lang=de&region=DE.

60. ChemBioDraw 14. Perkin Elmer; 2014.

61. Birner-Gruenberger R, Scholze H, Faber K, Hermetter A. Identification of various lipolytic enzymes in crude porcine pancreatic lipase preparations using covalent fluorescent inhibitors. Biotechnology and Bioengineering. 2004; 85(2):147–154. Available from: http://dx.doi.org/10.1002/bit.10894. doi: 10.1002/bit.10894

62. Li N, Zong MH, Ma D. Thermomyces lanuginosus lipase-catalyzed regioselective acylation of nucleosides: Enzyme substrate recognition. Journal of Biotechnology. 2009; 140(3–4):250–253. Available from: http://www.sciencedirect.com/science/article/pii/S0168165609000467. doi: 10.1016/j.jbiotec.2009.02.003

63. Brady L, Brzozowski AM, Derewenda ZS, Dodson E, Dodson G, Tolley S, et al. A serine protease triad forms the catalytic centre of a triacylglycerol lipase. Nature. 1990 Feb; 343(6260):767–770. doi: 10.1038/343767a0 PMID: 2304552

64. Grochulski P, Li Y, Schrag JD, Bouthillier F, Smith P, Harrison D, et al. Insights into interfacial activation from an open structure of Candida rugosa lipase. Journal of Biological Chemistry. 1993; 268(17):12843–7. Available from: http://www.jbc.org/content/268/17/12843.abstract. PMID: 8509417

65. Dalal S, Singh PK, Raghava S, Rawat S, Gupta MN. Purification and properties of the alkaline lipase from Burkholderia cepacia A.T.C.C. 25609. Biotechnology and Applied Biochemistry. 2008; 51(1):23–31. doi: 10.1042/BA20070186 PMID: 18052929

66. Sugiura M, Oikawa T, Hirano K, Inukai T. Purification, crystallization and properties of triacylglycerol lipase from Pseudomonas fluorescens. Biochimica et Biophysica Acta (BBA)—Lipids and Lipid Metabolism. 1977; 488(3):353–358. doi: 10.1016/0005-2760(77)90194-1

67. Hideko I, Harumi O, Hiroh I, Setsuzo T. Studies on lipase from Mucor javanicus. Biochimica et Biophysica Acta (BBA)—Lipids and Lipid Metabolism. 1975; 388(3):413–422. doi: 10.1016/0005-2760(75)90100-9

68. Pütz A. Isolierung, Identifizierung und biochemische Charakterisierung Dialkylphthalat spaltender Esterasen. Heinrich-Heine-Universität Düsseldorf; 2006.

69. Team RDC. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2008. Available from: http://www.r-project.org.

70. Wickham H. ggplot2. Rice University, (Housten); 2010. Available from: http://ggplot2.org.

71. Legendre P, Borcard D. Statistical comparison of univariate tests of homogeneity of variances. Submitted for publication in Journal of Statistical Computation. 2008;(514:). Available from: http://biol09.biol.umontreal.ca/BIO2041e/MS_THV.pdf.

72. Housecroft C, Sharpe A. Inorgaic Chemistry. Pearson Education Limited; 2005.

73. Arakawa H, Maeda M, Tsuji A, Kambegawa A. Chemiluminescence enzyme immunoassay of dehydroepiandrosterone and its sulfate using peroxidase as label. Steroids. 1981 Oct; 38(4):453–464. doi: 10.1016/0039-128X(81)90079-9 PMID: 6458929

74. Dyke KV, Dyke CV, Woodfork K. Luminescence Biotechnology-Instruments and Applications. CRC PRESS; 2001.

75. Couderc R, Baratti J. Oxidation of Methanol by the Yeast, Pichia pastoris. Purification and Properties of the Alcohol Oxidase. Agricultural and Biological Chemistry. 1980 Oct; 44(10):2279–2289. doi: 10.1271/bbb1961.44.2279

76. Thorpe GH, Kricka LJ, Moseley SB, Whitehead TP. Phenols as enhancers of the chemiluminescent horseradish peroxidase-luminol-hydrogen peroxide reaction: application in luminescence-monitored enzyme immunoassays. Clinical Chemistry. 1985 Aug; 31(8):1335–1341. Available from: http://www.clinchem.org/content/31/8/1335.abstract. PMID: 3926345

77. Banyai E. Indicators. Oxford: Pergamon Press; 1972.

78. Janes LE, Löwendahl AC, Kazlauskas RJ. Quantitative Screening of Hydrolase Libraries Using pH Indicators: Identifying Active and Enantioselective Hydrolases. Chemistry—A European Journal. 1998 Nov; 4(11):2324–2331. doi: 10.1002/(SICI)1521-3765(19981102)4:11%3C2324::AID-CHEM2324%3E3.0.CO;2-I

79. Rehm S, Trolder P, Pleiss J. Solvent-induced lid opening in lipases: A molecular dynamics study. Protein Science. 2010;p. 592122–2130.

80. Stillman TJ, Baker PJ, Britton KL, Rice DW. Conformational Flexibility in Glutamate Dehydrogenase: Role of Water in Substrate Recognition and Catalysis. Journal of Molecular Biology. 1993; 234 (4):1131–1139. Available from: http://www.sciencedirect.com/science/article/pii/S0022283683716657. doi: 10.1006/jmbi.1993.1665 PMID: 8263917

81. Eisenmesser EZ, Millet O, Labeikovsky W, Korzhnev DM, Wolf-Watz M, Bosco DA, et al. Intrinsic dynamics of an enzyme underlies catalysis. Nature. 2005 Nov; 438(7064):117–121. Available from: http://dx.doi.org/10.1038/nature04105 http://www.nature.com/nature/journal/v438/n7064/suppinfo/nature04105_S1.html. doi: 10.1038/nature04105 PMID: 16267559

82. Eisenmesser EZ, Bosco DA, Akke M, Kern D. Enzyme Dynamics During Catalysis. Science. 2002; 295 (5559):1520–1523. Available from: http://www.sciencemag.org/content/295/5559/1520.abstract. doi: 10.1126/science.1066176 PMID: 11859194

83. Tuñón I, Laage D, Hynes JT. Are there dynamical effects in enzyme catalysis? Some thoughts concerning the enzymatic chemical step. Archives of Biochemistry and Biophysics. 2015; 582:42–55. Special issue in computational modeling on biological systems. Available from: http://www.sciencedirect.com/science/article/pii/S0003986115002684. doi: 10.1016/j.abb.2015.06.004

84. Crismaru CG, Wybenga GG, Szymanski W, Wijma HJ, Wu B, Bartsch S, et al. Biochemical properties and crystal structure of a β-phenylalanine aminotransferase from Variovorax paradoxus. Applied and Environmental Microbiology. 2013 Jan; 79(1):185–95. doi: 10.1128/AEM.02525-12 PMID: 23087034

85. Dyke KV, Woodfork K. Light Probes. In: Luminescence Biotechnology. CRC Press; 2001. p. 2–29. Available from: http://dx.doi.org/10.1201/9781420041804.ch1.

86. Iversen PW, Eastwood BJ, Sittampalam GS, Cox KL. A Comparison of Assay Performance Measures in Screening Assays: Signal Window, Z' Factor, and Assay Variability Ratio. Journal of Biomolecular Screening. 2006 Apr; 11(3):247–252. Available from: http://jbx.sagepub.com/content/11/3/247.abstract. doi: 10.1177/1087057105285610 PMID: 16490779

87. Zhang XD. Novel Analytic Criteria and Effective Plate Designs for Quality Control in Genome-Scale RNAi Screens. Journal of Biomolecular Screening. 2008 Jun; 13(5):363–377. doi: 10.1177/1087057108317062 PMID: 18567841

6 Graph-based Analysis of MD Simulations

6.1 *StreaM* - a Stream-based Algorithm for Counting Motifs in Dynamic Graphs

In the following publication:

- Schiller B.*, **Jager, S.**, Hamacher K., Strufe T. (2015) StreaM - A Stream-Based Algorithm for Counting Motifs in Dynamic Graphs. In: Dediu AH., Hernandez-Quiroz F., Martín-Vide C., Rosenblueth D. (eds) Algorithms for Computational Biology. AlCoB 2015. Lecture Notes in Computer Science, vol 9199. Springer

the StreaM algorithm for motif (connectivity pattern) counting was developed for use in MD trajectories of proteins. The algorithm specializes in counting motifs (4-vertex) in high-resolution, dynamic graphs. As an example we used simulations of the pNB-Est13 and converted them into CG dynamic unit sphere graphs. In these graphs, StreaM counts all different 4-vertex motifs for each time-step. This methodology was used to quantify fluctuations of the complete protein as well as only secondary structural elements. We also achieved a maximum speed-up of up to 2300-fold in direct comparison with related work.

Contributions For this publication, I created a 50 ns simulation of an esterase (pNB-Est13). This simulation was the basis for the run-time evaluation and for all subsequent evaluations. Furthermore, I implemented the interfaces between the StreaM algorithm and the MD trajectories. For this purpose I implemented a CG tool, which converts MD simulations into dynamic graphs. I used it to create the graphs for the evaluation. Further contributions are the statistical analysis of simulations and graphs as well as the creation of Figure 7 and help to write and motivate the paper. Benjamin Schiller performed run-time benchmarks and defined the algorithm. Kay Hamacher and Thorsten Strufe improved the manuscript. In this article I am the second author.

StreaM - a Stream-based Algorithm for Counting Motifs in Dynamic Graphs

Benjamin Schiller[1], Sven Jager[2], Kay Hamacher[2,3], and Thorsten Strufe[1]

[1] Privacy and Data Security, Dept. of Computer Science, TU Dresden, Germany
[2] Computational Biology and Simulation, Dept. of Biology, TU Darmstadt, Germany
[3] Dept. of Physics, Dept. of Computer Science, TU Darmstadt, Germany

Abstract. Determining the occurrence of motifs yields profound insight for many biological systems, like metabolic, protein-protein interaction, and protein structure networks. Meaningful spatial protein-structure motifs include enzyme active sites and ligand-binding sites which are essential for function, shape, and performance of an enzyme. Analyzing their dynamics over time leads to a better understanding of underlying properties and processes. In this work, we present *StreaM*, a stream-based algorithm for counting undirected 4-vertex motifs in dynamic graphs. We evaluate *StreaM* against the four predominant approaches from the current state of the art on generated and real-world datasets, a simulation of a highly dynamic enzyme. For this case, we show that *StreaM* is capable to capture essential molecular protein dynamics and thereby provides a powerful method for evaluating large molecular dynamics trajectories. Compared to related work, our approach achieves speedups of up to $2,300$ times on real-world datasets.

1 Introduction

Motifs, the basic building blocks of any complex network, have been widely studied in the past [31]. They are often used in the analysis of biological networks, most notably protein-protein interactions [27,2,9,36,10], DNA [40,18], cellular networks [21], and protein structure networks [22]. Because of their general applicability, they have also been studied and used in other fields, e.g., to understand patterns in real and generated languages [6], to analyze and improve Peer-to-Peer networks [15,26], and for the generation of Internet PoP maps [13,12]. Recently, temporal motifs that describe how interactions between components change over time have been investigated [23,17] as well as degree-based signatures [30].

The problem of counting motifs in a static graph has been widely studied. The first approaches like ProMotif [16], mfinder [20], MAVisto [39], and NeMoFinder [8] provided tools to count motifs of small sizes but performed rather poorly, especially on larger graphs. This changed with the development of Fanmod [43], a very efficient algorithm that all recent approaches have been compared to. New algorithms like Kavosh [19] improve the efficiency of enumerating all subgraphs. G-Tries [35] is based on the idea of creating dedicated representations of sub graphs and ACC [28] uses combinatorial techniques to speed-up the computation. Furthermore, parallelized approaches [44,37] have been developed as well as approximations [14,42].

While the analysis of static networks is important, dynamic or time-dependent networks have recently gained a lot of attention. Analyzing dynamics of biological processes and systems is important for synthetic as well as for computational biology [3]. For example, the analysis of enzyme dynamics helps to understand how it works, and thus, reveals opportunities for improving its functionality. Analyzing the dynamics of amino acids to identify spatial arrangements that correspond to active sites or other functionally relevant features is important for protein classification and structure prediction [22,7]. Due to spatial amino acid arrangements, some motifs occur only in stable structure elements like α-helices and β-sheets and some represent general interactions. Moreover counting such motifs in dynamic graphs or motifs in any kind of biological network seems to be a promising approach to gain insight into various biological systems [33,41,4]

A common way to analyze the protein dynamics is the solution of Newton's equation of motion, i.e., molecular dynamics (*MD*). This method is used to quantify motions, mechanics, and spatial motifs within a single protein-structure as well as different molecular interactions. The MD approach approximates the time dependent behavior of a protein in its natural environment and results in a trajectory of atoms. One efficient way to analyze this trajectory is the use of graph-theoretic measures. Moreover, using dynamic graph measures like motif frequencies opens new opportunities for MD analysis. To this end, the trajectory has to be transformed into an amino acid contact map defined by distance cutoffs. Afterwards one can apply methods from graph theory to analyze the networks of transient contacts and identify flexible or rigid regions as well as important motifs. So far, only rough metrics like root mean square deviation (*RMSD*) of heavy atoms are utilized to capture protein dynamics. While the maximum number of contacts of an amino acid is approximately 6 [32], functional motifs are commonly considered on smaller sizes. For simplicity, we focus on 4-vertex motifs in this work.

The dynamics of time-dependent graphs are commonly modeled as a stream of updates that describe each change to the graph as an atomic operation [34]. Stream-based algorithms use these updates to continuously update graph-theoretic properties of interest. While such an algorithm for counting triangles in dynamic, undirected graphs has been developed [11], there is no approach for counting undirected 4-vertex motifs in dynamic graphs. Using snapshot-based algorithms like Fanmod, Kavosh, G-Tries, or ACC is expensive, especially when performing an analysis with a high granularity. In this work, we close this gap by developing a stream-based algorithm for updating the motif count in dynamic graphs.

The remainder of this paper is structured as follows: In Section 2, we introduce our terminology and define the problem of counting motifs in a dynamic graph. We introduce *StreaM*, a stream-based algorithm for counting undirected 4-vertex motifs in dynamic graphs, in Section 3. We present an evaluation of our algorithm against existing approaches in Section 4 as well as an application scenario were we showed a complete new approach of using the analytical power of

StreaM to capture essential protein dynamics in a large MD trajectory. Finally, we summarize and conclude our work in Section 5.

2 Background and terminology

In this Section, we introduce our terminology for graphs, dynamic graphs, and motifs. Then, we define the problem of counting motifs in dynamic graphs and discuss the general proceedings of analyzing dynamic graphs.

Graphs An *undirected, unweighted graph* $G = (V, E)$ is described by a vertex set $V = \{v_1, v_2, \dots\}$ and an edge set $E \subseteq \{\{v, w\} : v, w \in V \wedge v \neq w\}$. We define the neighborhood of v as $n(v) := \{w : \{v, w\} \in E\}$ and its degree as $d(v) := |n(v)|$. Then, the maximum degree of a graph is defined by $d_{max} := \max\limits_{v \in V} d(v)$.

Dynamic graphs As a *dynamic graph*, we consider a graph whose vertex and edge sets change over time. Each change of such a dynamic graph is then represented by an update of V or E that adds or removes an element. Hence, there are four different updates to a dynamic graph:

1. adding a new vertex (a$dd(v), v \notin V$),
2. adding a new edge (a$dd(e), e \notin E$),
3. removing an existing vertex (r$m(v), v \in V$), and
4. removing an existing edge (r$m(e), e \in E$).

Then, a dynamic graph is represented by its initial state $G_0 = (V_0, E_0)$ and an ordered list or stream of updates $u_1, u_2, u_3, \dots$. Their consecutive application transforms the graph over time:

$$G_0 \xrightarrow{u_1} G_1 \xrightarrow{u_2} G_2 \xrightarrow{u_3} G_3 \dots$$

Each state G_i can be seen as a separate snapshot at the respective point in time. We refer to a consecutive list of updates as a batch $B_{i,j} := \{u_{i+1}, \dots, u_j\}$ whose application transforms the dynamic graph G_i into G_j, i.e.,

$$G_i \xrightarrow{B_{i,j}} G_j$$

Motifs In this work, we consider *undirected 4-vertex motifs* (cf. Figure 1). They represent the 6 classes of isomorph, connected subgraphs of size 4. We denote them as $\mathcal{M} = \{m_1, m_2, \dots, m_6\}$.

Problem definition Counting the motifs in a given graph G means to determine the number of occurrences $\mathcal{F}(m_i)$ of each motif $m_i \in \mathcal{M}$. Assume a dynamic graph described by its initial state G_0 and a list of updates $U = (u_1, u_2, \dots, u_{|U|})$. Further assume a subset $S = (s_0, s_1, \dots, s_t), 0 \leq s_0, s_i < s_{i+1}, s_t \leq |U|$ of its

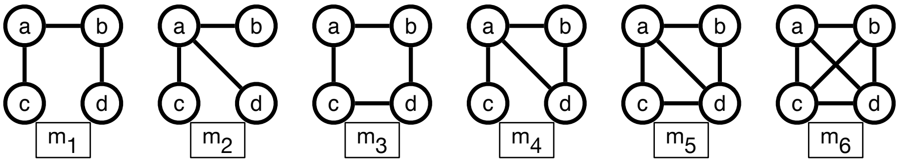


Fig. 1: The 6 undirected 4-vertex motifs $\mathcal{M} = \{m_1, m_2, m_3, m_4, m_5, m_6\}$

states which determines the granularity of the analysis. Then, the problem is to generate the motif count $\mathcal{F}_s$ for each state $s \in S$ of interest. Hence, the result of counting motifs in a dynamic graph is a list of motif frequencies $\mathcal{F}_{s_0}, \mathcal{F}_{s_1}, \dots, \mathcal{F}_{s_t}$ which describes how they change over time.

Analysis of dynamic graphs The properties of dynamic graphs can be analyzed at different granularities. At the highest granularity, properties like degree distribution, shortest-path lengths, or motif count are computed for each change, i.e., each possible state $G_0, G_1, G_2, \dots$ is analyzed. Lowering this granularity means to only consider a subset of these states, e.g., every 10th state $G_0, G_{10}, G_{20}, \dots$ or an arbitrary subset $G_0, G_{s_1}, G_{s_2}, \dots, s_i < s_{i+1}$.

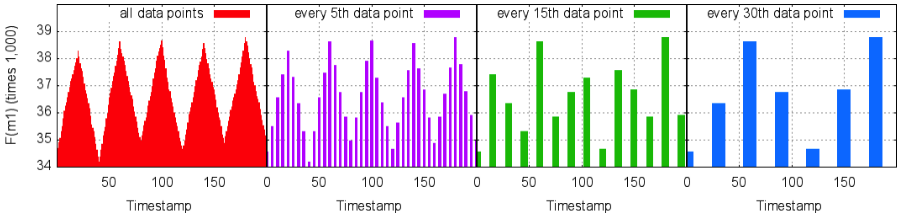


Fig. 2: $\mathcal{F}(m_1)$ in a random, dynamic graph analyzed with different granularities

As an example, take a random graph $G_0, \dots, G_{200}$ with 100 vertices and 500 edges where 20 random edge additions are always followed by 20 random edge removals. Figure 2 shows $\mathcal{F}(m_1)$, the number of occurrences of m_1 over time which fluctuates between 34,000 and 39,000. Performing an analysis at highest granularity shows the impact of each update, i.e., 201 data points. In case the granularity is lowered (only every 5^{th}, 15^{th}, or 30^{th} state is considered), local maxima are missed and the appearance of the development changes. Therefore, it is desirable to determine the properties of a dynamic graph at a high granularity.

Snapshot-based algorithms are executed separately for each snapshot $G_{s \in S}$ to obtain $\mathcal{F}_s$. Hence, the total runtime grows roughly linearly with the number of analyzed snaphots. In contrast, using *stream-based analysis*, the granularity does not influence the runtime. After computing $\mathcal{F}_0$ for the initial graph G_0 using any snapshot-based algorithm, each count $\mathcal{F}_{s_{i+1}}$ is computed by taking G_{s_i}, $\mathcal{F}_{s_i}$, and $B_{s_i,s_{i+1}}$ as input. Since each update is applied separately, an increase in granularity only requires to output the results with a higher frequency. Therefore, stream-based analysis should outperform snapshot-based approaches in case a high granularity is desired and the cost of applying the updates between two states is less than a complete re-evaluation.

3 Counting motifs in dynamic graphs

In this Secion, we describe basic insights regarding motifs in dynamic graphs. Then, we describe *StreaM*, a new stream-based algorithm for counting undirected 4-vertex motifs in dynamic graphs, and discuss its runtime complexity.

Basic insights Whenever an edge $e = \{a, b\}$ is added to a graph G_t, i.e., update $u_{t+1} = add(e)$, two things happen: existing motifs are changed and new motifs are formed. First, consider an existing motif m_i that consists of a, b, and 2 other vertices. The addition of e causes the motif to change into a different motif m_j which contains one more edge. We denote this operation as $(i \to j)$. Its execution decreases the occurrences of m_i and increases the occurrences of m_j, i.e.,

$$(i \to j) : \mathcal{F}_{t+1}(m_i) := \mathcal{F}_t(m_i) - 1,\ \mathcal{F}_{t+1}(m_j) := \mathcal{F}_t(m_j) + 1$$

Second, consider vertices c and d that do not form a connected component with a and b without e's existence. In case e connects the four vertices, a new motif m_k is formed. We denote this operation as $+(k)$. Its execution increases the occurrences of m_k, i.e.,

$$+(k) : \mathcal{F}_{t+1}(m_k) := \mathcal{F}_t(m_k) + 1$$

In case an existing edge is removed, i.e., $u_{t+1} = rm(e)$, the inverse happens: some motifs are changed and others are dissolved. We denote these operation as $(i \to j)^{-1}$ and $+(i)^{-1}$.

$$\begin{aligned} (i \to j)^{-1} &: \mathcal{F}_{t+1}(m_i) := \mathcal{F}_t(m_i) + 1,\ \mathcal{F}_{t+1}(m_j) := \mathcal{F}_t(m_j) - 1 \\ +(k)^{-1} &: \qquad\qquad\qquad\quad \mathcal{F}_{t+1}(m_k) := \mathcal{F}_t(m_k) - 1 \end{aligned}$$

Adding or removing a vertex with degree 0 has no effect on the motif count.

Each motif $m_i \in \mathcal{M}$ contains at least 3 and at most 6 edges. The addition and removal of edges leads to transitions between them (cf. Figure 3). For example, adding the missing edge to m_5 changes it to m_6 ($(5 \to 6)$) while removing any edge from m_6 changes it to m_5 ($(5 \to 6)^{-1}$). Adding edge $\{b, d\}$ to the disconnected set of nodes x creates a new motif m_1 ($+(1)$) which is dissolved by the removal of any of its 3 edges ($+(1)^{-1}$).

The main idea behind our new stream-based algorithm is to find and apply these operations to correctly update $\mathcal{F}$ for each edge addition and removal.

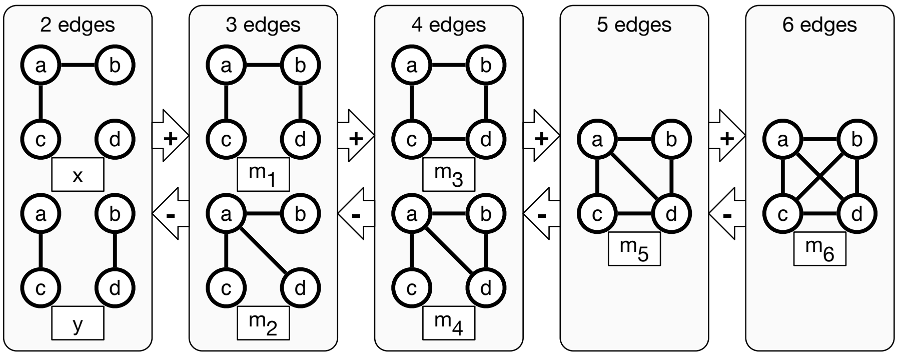


Fig. 3: Transitions between the motifs $m_i \in \mathcal{M}$ when adding and removing edges

StreaM Assume an update (addition or removal) of edge $e = \{a, b\}$. To correctly adapt $\mathcal{F}$, we need to consider all 2-vertex sets $\{c, d\} \in CD(a, b)$ such that a, b, c, and d form a motif if e exists. Either both vertices are connected to a or b directly or d is a neighbor of c which is connected to a or b. With

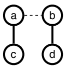
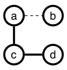
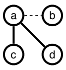
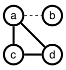
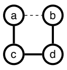
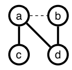
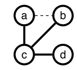
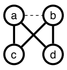
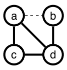
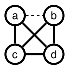
$$N(a, b) := (n(a) \cup n(b)) \backslash \{a, b\},$$

we can define $CD(a, b)$ as follows:

$$CD(a, b) = \{\{c, d\} : (c, d \in N(a, b), c \neq d) \vee (c \in N(a, b), d \in n(c) \backslash \{a, b\})\}$$

Besides $\{a, b\}$, 5 edges are possible between a, b, c, and d. We denote their existence as a quintuple $\mathcal{S}(a, b, c, d) = (ac, ad, bc, bd, cd)$, called their *signature*. At least two distinct edges must exist, the first connecting c and the second connecting d. Therefore, there are $2^5 - 2 \cdot 2^2 = 24$ possible signatures.

Table 1: Operation mapping $\mathcal{O}$ from signatures $\mathcal{S}(a, b, c, d)$ to operations

$\mathcal{S}$	10010 01100	10001 01001 00101 00011	11000 00110	11001 00111	10011 01101	11100 11010 10110 01110	10101 01011	11110	11101 11011 10111 01111	11111
$\mathcal{O}$	$+(1)$	$+(1)$	$+(2)$	$+(4)$	$(1 \rightarrow 3)$	$(1 \rightarrow 4)$	$(2 \rightarrow 4)$	$(3 \rightarrow 5)$	$(4 \rightarrow 5)$	$(5 \rightarrow 6)$
										

Each signature corresponds to a specific operation that must be executed to update $\mathcal{F}$. We define a function $\mathcal{O}$ that maps a signature $\mathcal{S}$ on the corresponding operation. The complete assignment of signatures to operations is given in

Table 1. In case the edge $\{a, b\}$ is removed instead of added, the inverse operation must be executed. As an example consider the signature (10010) which is isomorph to (01100). The addition of $\{a, b\}$ creates the motif m_1. Its removal dissolves the motif as a, b, c, and d are no longer weakly connected.

Based on $\mathcal{S}$ and $\mathcal{O}$, we can now describe the stream-based algorithm *StreaM* for updating the motif frequency in an undirected graph (cf. Algorithm 1). For an edge $\{a, b\}$ that is added or removed (described by *type*), we first determine the set $CD(a, b)$ of all pairs of vertices connected to a and b. For each pair $\{c, d\} \in CD(a, b)$, the required operation $o = \mathcal{O}(\mathcal{S}(a, b, c, d))$ is determined from the signature of a, b, c, and d. If $\{a, b\}$ is added, the operation o is executed. Otherwise, the inverse operation o^{-1} is executed.

```
Data: $G, \{a,b\}, type \in \{add, rm\}$
begin
    for $\{c,d\} \in CD(a,b)$ do
        $o = \mathcal{O}(\mathcal{S}(a,b,c,d))$ ;                /* operation */
        if $type = add$ then
            execute $o$ ;                                  /* add edge */
        else if $type = rm$ then
            execute $o^{-1}$ ;                             /* remove edge */
    end
end
```

Algorithm 1: *StreaM* for maintaining $\mathcal{F}$ in dynamic graphs

Complexity discussion *StreaM* iterates over the $|CD(a, b)| \leq 5 \cdot (d_{max})^2$ elements of $CD(a, b)$. For each element $\{c, d\}$, it computes the signature which can be done in $5 \cdot O(1)$ time, assuming hash-based datastructures are used for adjacency lists. In addition, $\mathcal{F}$ is incremented or decremented which has time complexity of $O(1)$ as well. Therefore, processing a single edge addition or removal with *StreaM* has time complexity of

$$O((d_{max})^2) \cdot (O(1) + O(1)) = O((d_{max})^2)$$

4 Evaluation

In this Section, we evaluate the runtime performance *StreaM*. First, we briefly discuss our evaluation setup. Then, we evaluate the runtime dependence of the algorithm to batch size as well as vertex degree. We compare the runtime of our algorithms to related work in scenarios where the analysis is performed at high granularity and on dynamic graphs obtained from MD simulations consisting of 20, 000 snapshots. Finally, we show that *StreaM* is a powerful and unique approach to capture essential molecular dynamics and gain insights on secondary structure focused amino acid interactions.

Evaluation setup All measurements are executed on an *HP ProLiant DL585 G7* server with 64 *AMD OpteronTM 6282SE* processors with 2.6GHz each running a Debian operating system. We implemented *StreaM* in the Java-based framework DNA (Dynamic Network Analyzer) [38] for the analysis of dynamic graphs. The framework including our implementation of *StreaM* is available on the project's GitHub page[4]. We compare our approach with four popular snapshot-based approaches for counting motifs: Fanmod [43], Kavosh [19], G-Tries [35], and ACC [28]. For all approaches, we use the original programs provided by their authors[5][6][7][8]. The cmd-line version of Fanmod as well as G-Tries and Kavosh are implemented in C++. We compiled them from the original sources using GCC version 4.7.2. Like our approach, ACC is implemented in Java and executed using a 64-bit JVM with version 1.7.0_25.

Complexity of *StreaM* Now, we validate the runtime complexity of *StreaM* discussed in Section 3. We generated undirected random graphs (R) as well as power-law graphs (PL) using the Barabási-Albert model with 500 vertices and 5,000, 10,000, and 15,000 edges. First, we generated 200 batches for each graph with a growing number of random edge exchanges. A random edge exchange is performed by selecting two random edges $e_1 = (a, b)$ and $e_2 = (c, d)$ and exchanging their end points, i.e., transforming the edges to $e'_1 = (a, d)$ and $e'_2 = (c, b)$. This implies 4 updates which are added to the respective batch: $rm(e_1)$, $rm(e_2)$, $add(e'_1)$, and $add(e'_2)$. The i^{th} batch contains i edge exchanges, denoted as $E^x(i)$. Second, we created 200 batches for each graph, each containing 250 random edge addition, denoted as $E^+(250)$. The application of each batch leads to an increase of the average and maximum degree by 1, hence 200 over all. For all graphs and batch types, we recorded the average per-batch runtime of 20 repetitions while performing an analysis using *StreaM*. In the first case, we expect the runtime to grow linearly with the number of updates $|B|$ because $E^x(i)$ does not change d_{max} significantly during the application of each batch B. Furthermore, we expect the runtime of *StreaM* to grow quadratically with the batches $E^+(250)$) since the maximum degree is increased by 1 with each batch.

Figure 4a shows the per-batch runtimes for the analysis of random and power-law graphs for $E^x(i)$. For all graphs, the runtime grows linearly with each batch. The per-batch runtimes of applying $E^+(250)$ to both graph types is shown in Figure 4b. For all datasets, the runtime appears to grow quadratically with average and maximum degree which are increased by approximately 1 with each batch. As expected, the runtime of *StreaM* increases linearly with the batch size (cf. Figures 4a. Since the application of $E^x(i)$ does not alter the degree distribution, average and maximum degree stay constant which results in this linear increase of the runtime. Furthermore, the runtime of *StreaM*, in dependence of

---

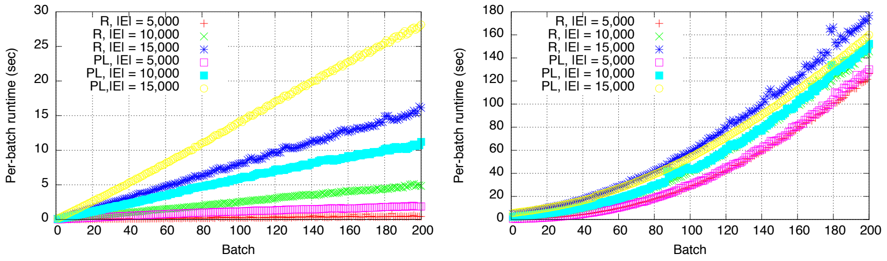
[4] `https://github.com/BenjaminSchiller/DNA`

[5] `http://theinf1.informatik.uni-jena.de/~wernicke/motifs/`

[6] `http://lbb.ut.ac.ir/Download/LBBsoft/Kavosh/`

[7] `http://www.dcc.fc.up.pt/gtries/`

[8] `http://www.ft.unicamp.br/docentes/meira/accmotifs/`



(a) Growing edge exchange ($E^x(i)$)　(b) Random edge addition ($E^+(250)$)

Fig. 4: Per-batch runtime for random (R) and power-law (PL) graphs of size $|V| = 500$ with different edge count $|E|$ depending on batch type

the maximum vertex degree d_{max}, is bounded from above by a quadratic function as its algorithmic complexity of $O((d_{max})^2)$ implies. These results validate the complexity discussion. The runtime of *StreaM* grows linearly with the batch size and depends quadratically on the maximum degree. Therefore, *StreaM* can be executed without performance penalties for arbitrary granularity.

Analysis with high granularity Next, we show that *StreaM* outperforms snapshot-based approaches for analyses with high granularities. As initial graphs, we consider 12 different datasets that have already been used in the evaluation of ACC [28] and other snapshot-based approaches. The datasets originate from a wide range of areas including biological, social, and traffic networks (cf. Table2). Their size ranges from 418 to 12,905 vertices with an average degree between 1.85 and 22.01. As dynamics, we created 1,000 batches each consisting of a single random edge exchange $E^x(1)$, i.e., $|B| = 4$. We measure the total time it takes each approach to determine the motif count of the resulting 1,001 states. In some cases, the execution of the snapshot-based approach did not finish after a whole week. We terminated these processes and extrapolated the runtime for the analysis of the 1,001 snapshots from the number of completed ones. The values for G-Tries on foldoc are excluded since it did not even process the first snapshot during this time. Especially for larger graphs, the repeated re-computation of the snapshot-based algorithms should perform much slower than the stream-based application of small batches. Therefore, we expect *StreaM* to outperform the snapshot-based approaches for all graphs.

Figure 5 depicts the resulting runtimes for each dynamic graph. For most datasets and approaches, *StreaM* performs between 10 and 1,000 times faster than the other approaches. The smallest speedup we observed was for the *wordsenglish* dataset, where *StreaM* is still 4.6 times faster than ACC, the best competitor. These results comply with our expectations. Since *StreaM* only computes the motif count of the complete graph once and then only updates the results for 4 updates between two states it performs much faster than all snapshot-based ap-

	$\lvert V\rvert$	$\lvert E\rvert$	d_{avg}		$\lvert V\rvert$	$\lvert E\rvert$	d_{avg}
ecoli	418	519	2.48	**odlis**	2,900	16,377	11.29
yeast	688	1,078	3.13	**epa**	4,271	8,909	4.17
roget	1,010	3,648	7.22	**pairsfsg**	5,018	55,227	22.01
airport	1,574	17,215	21.87	**california**	6,175	15,969	5.17
csphd	1,882	1,740	1.85	**wordsenglish**	7,381	44,207	11.98
facebook	1,899	13,838	14.57	**foldoc**	12,905	83,101	12.88

Table 2: Properties of datasets used for evaluation with high granularity

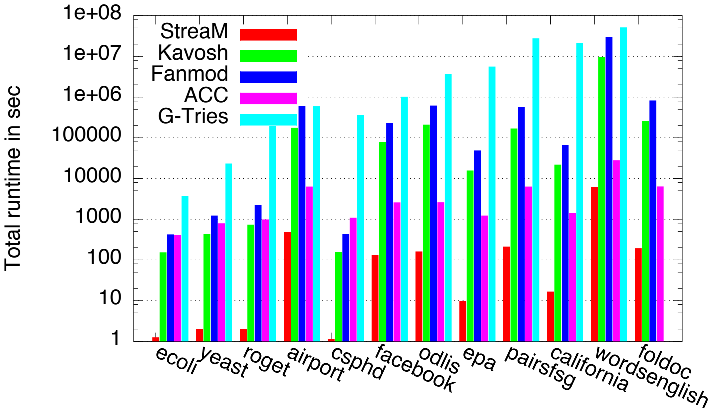


Fig. 5: Total runtime of analysis with high granularity (1,000 times $E^x(1)$)

proaches. Therefore, it becomes clear that *StreaM* outperforms snapshot-based algorithms in case an analysis with high granularity is desired.

MD-simulations case Motifs are essential for the structural classification of proteins which can be observed from amino acid interactions during MD simulations. In proteins, some motifs, like helices, occur only within structurally stable elements while others occur more frequently. Counting such structural motifs during an MD simulation of a dynamic enzyme is an interesting approach and allows the evaluation of essential molecular dynamics. Therefore, we analyze an MD simulation using *StreaM* as well as snapshot-based algorithms to investigate their performance on such a realistic dynamic graph. In addition, we investigate the general applicability of *StreaM* to gain new insights of capturing molecular dynamics from amino acid interaction motifs.

We created a graph time series from a molecular dynamics simulation of an enzyme, the para Nitro Butyrate Esterase-13 (*pNB-Est13*), a large carboxilic esterase. It is used as an additive of cleansing agents and holds a big potential towards plastic degradation [29]. *PNB-Est13* monomer consists of 491 residues with molecular weight of 35 to 40 kDa [29]. As a protein structure we used an homology model of *pNB-Est13*. We used *1C7J chain A*, *1QE3 chain B*, and *1C7I chain A* as templates for the homology model. MD simulations were performed with the *Yasara* software suite [24] using the *AMBER03* force-field [1] with constant temperature (313K), pressure (1 bar), and pH (7.4). At the beginning of the simulation, the box is filled with *Tip3* water and *NaCl* counter ions (0.9%). Afterwards, the protein was energy-minimized utilizing the *AMBER03* force field until convergence was reached (0.01 kJ/mol per atom during 200 steps) [25]. We equilibrated the solvent for 500 ps. Then we simulated for 50 ns and took snapshots every 2.5 ps. From these 20,000 snapshots, we generated a dynamic graph with 491 vertices, each modeling an amino acid C_α. An undirected edge is created between two vertices in case their Euclidean distance is shorter than $d = 7$Å.In addition, we generated dynamic graphs using $d \in [8, 12]$Å to create denser graphs. All 6 dynamic graphs consist of 491 vertices connected by 1,904 to 7,398 edges on average, depending on the distance threshold d (cf. d (cf. Table 3)). The average batch size ranges between 141.16 and 487.7 implying that 6.6% to 8.58% of all connections are changed between two snapshots. To compare the performance of *StreaM* to existing approaches, we analyzed these dynamic graphs consisting of an initial graph and 19,999 batches using our stream-based implementation and measured the total runtime of the execution. For existing approaches, we generated the 20,000 separate snapshots that represent the dynamic graph and analyzed each one separately. We added the execution times of all steps to obtain a single runtime.

The averages of 20 repetitions for all approaches and distance thresholds d are shown in Figure 6. For the standard distance threshold of 7Å, our stream-based approach takes 173 sec while Kavosh, the fastest competitor, takes 2,365 sec. The analysis using Fanmod and ACC takes around 10,000 sec while G-Tries runs for 400,000 sec. Hence, the speedup of *StreaM* lies between 13.7 and 2,300

	7Å	**8Å**	**9Å**	**10Å**	**11Å**	**12Å**
$\|V\|$	491	491	491	491	491	491
$\|E\|$	1,904	2,413	3,248	4,370	5,877	7,398
d_{avg}	7.76	9.83	13.23	17.8	23.94	30.13
$\|B\|$	141.16	176.54	278.92	366.02	422.32	487.70
$\frac{\|B\|}{\|E\|}$	7.42%	7.32%	8.58%	8.38%	7.18%	6.60%

Table 3: Properties of the dynamic graphs generated during MD simulations

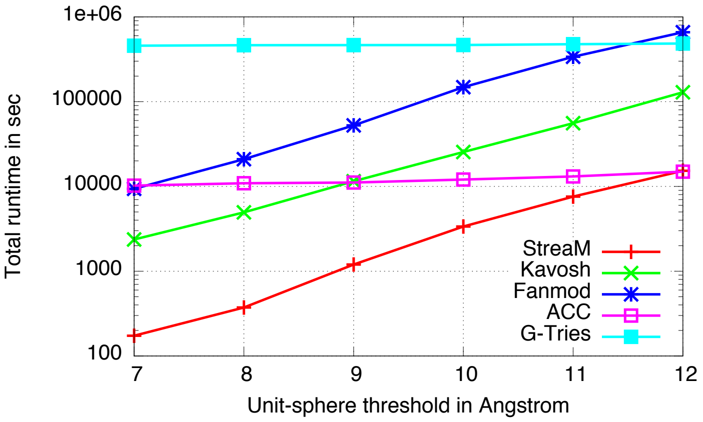
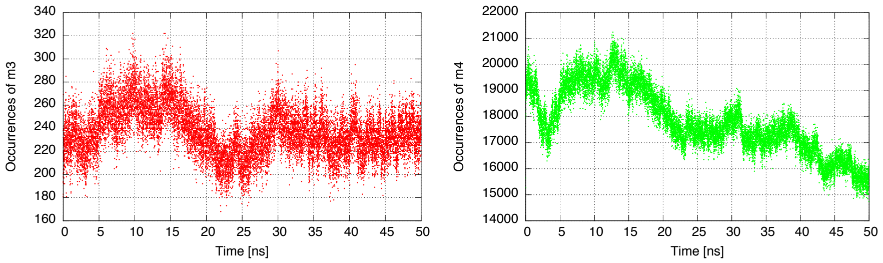


Fig. 6: Total runtime of MD graph analysis depending on distance threshold

times when compared to snapshot-based approaches for the standard case of $d = 7\text{Å}$ [5]. When increasing the distance threshold d, the runtimes of *StreaM*, Kavosh, and Fanmod increase in a similar way. Interestingly, the runtimes of ACC and G-Tries only increase slightly, indicating that they mainly depend on the number of vertices in a graph and not the number of edges.

As expected and indicated by our complexity discussion and evaluation before, the runtime of *StreaM* increases as the batch size and maximum vertex degree grows. Since the runtime of ACC does not increase as drastically with the distance threshold, the runtimes of both approaches are very close for the highest investigated threshold of 12Å. Notably, a distance threshold of 12Å is not realistic for amino acid contact prediction.

This performance evaluation shows that *StreaM* outperforms snapshot-based algorithms when analyzing realistic dynamic graphs, in our case from MD simulations. Except for unrealistically dense graphs, obtained with a distance threshold of 12Å, *StreaM* performs considerably faster than all other approaches. Hence, it allows a much faster analysis of dynamic biological networks such as the protein graphs obtained from MD simulations. For structural motifs during protein dynamics, a high granularity is very important to count transient interactions. Especially long term MD trajectories require fast and efficient algorithms to analyze transient amino acid interactions. To avoid unstable or unrealistic long term simulations, in case of protein unfolding or using incorrect force field parameter sets, *StreaM* indeed is powerful enough to monitor MD stability online in parallel to the execution of the MD simulation.



(a) $\mathcal{F}(m_3)$

(b) $\mathcal{F}(m_4)$

Fig. 7: Motif occurrences over time in the dynamic graph for $d = 7\text{Å}$

Interpretation of analysis results For the quantification of *pNB-Est13*'s dynamic behavior we now investigate two meaningful motifs: The structure of m_4 is typical for stabilizing effects between structure elements or loops. It is capable to describe 3 amino acids which are covalently connected within the backbone and interact with a flexible loop due to electrostatic or hydrophobic interactions. In contrast, m_3, a circle/loop containing 4 edges, can only be found in robust structure elements like α-helices and β-sheet. In case of our MD simulation, we observe that the occurrences of m_4 decreases over time (cf. Figure 7b). This means that the protein structure enlarges during simulation. In relation to the *RMSD*, we observe a Pearson correlation of -0.67 (p-value $< 2.2 \cdot 10^{-16}$, 95% conf. interval, -0.673 to -0.657). Similar for m_4 we observe a Pearson correlation of *RMSD* to m_3 with a value of -0.190 (p-value $< 2.2 \cdot 10^{-16}$, 95% conf. interval: -0.204 to -0.177). Clearly, the number of m_3 motifs remain nearly constant over time as shown in Figure 7a. This behavior agrees with the general assumption that m_3 can only be found in stable structure elements, such as α-helices, which is necessary for a constant *RMSD*. In case of MD graphs, a high granularity is indispensable to capture all transient amino acid interactions. In contrast, snapshot-based approaches do not allow for an analysis at such high granularity and therefore cannot generate similar insights. From these results, we can conclude that *StreaM* is capable of capturing essential molecular dynamics at high granularity - in particular important structural features based on secondary structure focused amino acid interactions. To this end, besides its outstanding performance, we showed that *StreaM* is a powerful new algorithm for the analysis of large MD trajectories.

5 Summary, Conclusion, & Future Work

As dynamic graphs have gained much attention in the recent past, not many approaches exist to efficiently analyze the time-dependent properties of such networks. In this work, we developed *StreaM*, a stream-based algorithm for counting undirected 4-vertex motifs in dynamic graphs. We evaluated the algorithm on generated datasets as well as realistic graphs obtained from MD simulations of *pNB-Est13*. We showed that using motifs for protein dynamic analysis helps to distinguish between structure elements and general interactions and might be a valuable, additional analysis procedure to assess local stability of MD trajectories whereas *RMSD* measures global stability. Our approach outperforms state-of-the-art by up to $2,300$ times on real-world datasets. Thereby, it enables the fine-grained analysis required to understand highly dynamic graphs over time.

Dynamic aspects are typically done on small motifs because the maximal number of contacts of an amino acid is approximately six. In the future, we will extend our work by generalizing the algorithm for arbitrary motif sizes and developing rule sets for other motif types. Moreover we will dynamically annotate individual amino acids and the respective motifs participate in, during a simulation. This will open new possibilities for bimolecular engineering and in particular enzyme engineering.

References

1. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. (2003)

2. Albert, I., Albert, R.: Conserved network motifs allow protein–protein interaction prediction. Bioinformatics (2004)
3. Alder, B.J., Wainwright, T.E.: Studies in Molecular Dynamics. Journal of Chemical Physics (1959)
4. Alon, N, D.P.H.I.H.F., Sahinalp, S.C.: Biomolecular network motif counting and discovery by color coding. Bioinformatics (2008)
5. Atilgan, A.R., et al.: Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophysics Journal (2001)
6. Biemann, C., et al.: Quantifying semantics using complex network analysis. In: COLING (2012)
7. Chakraborty, S., Biswas, S.: Approximation algorithms for 3-d common substructure identification in drug and protein molecules. (1999)
8. Chen, J., et al.: Nemofinder: Dissecting genome-wide protein-protein interactions with meso-scale network motifs. In: ACM SIGKDD (2006)
9. Chen, J., et al.: Labeling network motifs in protein interactomes for protein function prediction. In: IEEE ICDE (2007)
10. Colak, R., et al.: Dense graphlet statistics of protein interaction and random networks. In: Pacific Symposium on Biocomputing (2009)
11. Ediger, D., et al.: Massive streaming data analytics: A case study with clustering coefficients. In: IEEE IPDPSW (2010)
12. Feldman, D., Shavitt, Y.: Automatic large scale generation of internet pop level maps. In: IEEE GLOBECOM (2008)
13. Feldman, D., et al.: A structural approach for pop geo-location. Computer Networks (2012)
14. Gonen, M., Shavitt, Y.: Approximating the number of network motifs. Internet Mathematics (2009)
15. Hales, D., Arteconi, S.: Motifs in evolving cooperative networks look like protein structure networks. Networks and Heterogeneous Media (2008)
16. Hutchinson, E.G., Thornton, J.M.: Promotif—a program to identify and analyze structural motifs in proteins. Protein Science (1996)
17. Jurgens, D., Lu, T.: Temporal motifs reveal the dynamics of editor interactions in wikipedia. In: ICWSM (2012)
18. Kalir, S., et al.: Ordering genes in a flagella pathway by analysis of expression kinetics from living bacteria. Science (2001)
19. Kashani, Z.R.M., et al.: Kavosh: a new algorithm for finding network motifs. BMC Bioinformatics (2009)
20. Kashtan, N., et al.: Mfinder tool guide. Technical Report (2002)
21. Kim, J., et al.: Coupled feedback loops form dynamic motifs of cellular networks. Biophysical journal (2008)
22. Kleywegt, D.J.: D. J. Kleywegt. Recognition of spatial motifs in protein structures. Journal of Molecular Biology (1999)
23. Kovanen, L., et al.: Temporal motifs in time-dependent networks. Journal of Statistical Mechanics: Theory and Experiment (2011)
24. Krieger, E., et al.: Increasing the precision of comparative models with YASARA NOVA–a self-parameterizing force field. Proteins (2002)
25. Krieger, E., et al.: Fast empirical pKa prediction by Ewald summation. Journal of molecular graphics & modelling (2006)
26. Krumov, L., et al.: Leveraging network motifs for the adaptation of structured peer-to-peer-networks. In: IEEE GLOBECOM (2010)
27. Maslov, S., Sneppen, K.: Specificity and stability in topology of protein networks. Science (2002)

28. Meira, L.A.A., et al.: acc-motif detection tool. arXiv:1203.3415 (2012)
29. Michels, A., et al.: Verwendung von esterasen zur spaltung von kunststoffen (2011)
30. Milenkoviæ, T., Pržulj, N.: Uncovering biological network function via graphlet degree signatures. Cancer Informatics (2008)
31. Milo, R., et al.: Network motifs: simple building blocks of complex networks. Science (2002)
32. Miyazawa, S., Jernigan, R.L.: Residue–Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading. Journal of Molecular Biology
33. Panni, S., Rombo, S.E.: Searching for repetitions in biological networks: methods, resources and tools. Briefings in Bioinformatics (2015)
34. Rauch, M., et al.: Computing on data streams. In: DIMACS Workshop External Memory and Visualization (1999)
35. Ribeiro, P., Silva, F.: G-tries: an efficient data structure for discovering network motifs. In: ACM Symposium on Applied Computing (2010)
36. Royer, L., et al.: Unraveling protein networks with power graph analysis. PLoS computational biology (2008)
37. Schatz, M., et al.: Parallel network motif finding. Tech. rep. (2008)
38. Schiller, B., Strufe, T.: Dynamic network analyzer building a framework for the graph-theoretic analysis of dynamic networks. In: SummerSim (2013)
39. Schreiber, F., Schwöbbermeyer, H.: Mavisto: a tool for the exploration of network motifs. Bioinformatics (2005)
40. Shen-Orr, S.S., et al.: Network motifs in the transcriptional regulation network of escherichia coli. Nature genetics (2002)
41. Tran, N H, C.K.P., Zhang, L.: Counting motifs in the human interactome. Nature Communications
42. Wernicke, S.: Efficient detection of network motifs. IEEE ACM TCBB (2006)
43. Wernicke, S., Rasche, F.: Fanmod: a tool for fast network motif detection. Bioinformatics (2006)
44. Zhao, Z., et al.: Subgraph enumeration in large social contact networks using parallel color coding and streaming. In: ICPP (2010)

6.2 Motif Based Analysis of MD Simulations

The following section will give an elaborate demonstration of a motif-based analysis approach using four different examples. First, we analyze simulations of water in a cubic box at different densities and temperatures.

We observe a high correlation between the frequencies of 3-vertex motifs and the thermodynamic entropy – thus assessing conformational entropy to be a driving force.

Second, we apply our method to MD simulations of confined water within minerals. We extract structural and dynamic properties of water near the surface of a pore wall based on the frequencies of 3-vertex motifs.

Third, we use a small α-helical peptide as a *toy model*. We monitor the unfolding at 320K and compare 7-vertex motif counts with conventional metrics like the RMSD.
In the last example, previously identified motifs are taken and used to describe the formation of a molecular complex. This simulation consists of a complex of *Interleukin-8* (IL-8) and the synthetic inhibitor peptide *IL8RPLoops*.

While distance based metrics do not adequately describe the complex formations, 7-vertex motif counts made it possible to monitor and describe this process in detail. This part was done in collaboration with Benjamin Schiller, Thorsten Strufe, Michael Vogel and Kay Hamacher. The following parts are from a manuscript draft [1]. All illustrations and computations were made with the help of the following `R` librarys: `ggsci`, `StreaMD`, `ggplot2` and `cowplot` (*153–156*).

MD Simulation Protocols

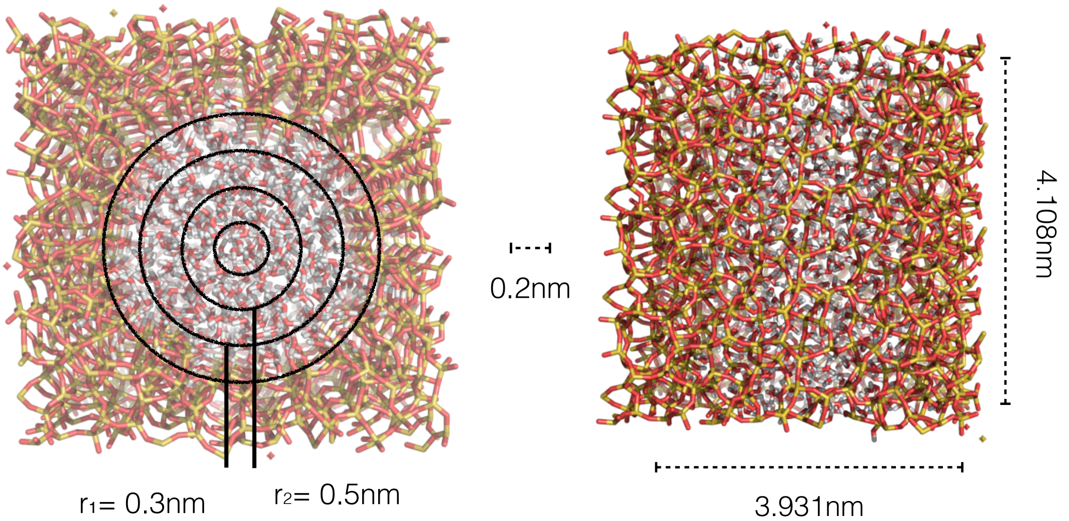
Simulation of SPC/E Water in a quadratic Box We simulate extended simple point charge water (SPC/E) which models the oxygen and the two hydrogen atoms of a water molecule as separate atoms. We use a bond length of 0.1 nm between oxygen and hydrogen atoms and an angle of 109.49 between the oxygen-hydrogen atoms. This results in a default distance of 0.17 nm between the two hydrogen atoms of a water molecule as illustrated in Figure 6.1. We perform 27 simulations with temperatures between 273.16 K and 647.29 K as well as densities ranging from 0.9998 kg/m^3 to 0.00000485 kg/m^3. A list of

---

[1] Motif Based Analysis of MD Simulations

all 27 configurations is given in Table 10.3. In all cases, we simulate 216 water molecules at a constant volume and temperature. After a short energy minimization, we perform an NVT equilibration for 2 ns using a weak Berendsen temperature coupling (*157*) to reach the target pressure. Afterward, we use temperature coupling with a *Nosé-Hoover* thermostat (*121*). Bond lengths were constrained using the LINCS (*118*) algorithm. The Lennard-Jones non-bonded interactions were evaluated using a cutoff distance of 1.4 nm. We set the van der Waals interaction cutoff to 0.95 nm and the integration step-size to 1 fs and obtain a total of 10 000 frames.

MD Simulation of SPC/E Water in mineral Confinement The MD simulations of the small pore are done using the `Gromacs` simulation software package (*158*). We use the same simulation-setups at four different temperatures: 200K, 250K, 270K, 350K. The temperature was set using the *Nosé-Hoover* thermostat (*119*, *121*). The radius of the silica pore amounts to 1.1 nm (*159*). The water density was set to $\rho = 1kg/m^3$ at all studied temperatures. Further simulation details can be found in previous work (*160*), where the structure and dynamics of the confined water were analyzed using conventional methods for data analysis. To ensure an equilibrated system, we discard the initial 3000 frames and model dynamic graphs from the following 1000 frames. We create 32 dynamic graphs with increasing distance r to the pore wall $r \in [0.3, 1.7]$ nm



(a) Schematic space for dynamic graphs (b) Side view of simulated water

Figure 6.1: Model of water in mineral confinement with pore dimensions

(cf. Figure 6.1) as well as four different temperatures. Every graph has a thickness of 0.2 nm. Basic statistics of all generated dynamic graphs are given in Table 10.4. The walls of the silica pores are rough so that Silicon atoms at the pore surface are located at these distances from the pore axis. Only water atoms within distance r and with a greater distance to the silica wall than to the center of the box is considered. To model the dynamic graphs we use a distance threshold d of 0.19 nm for the water molecules. Further details can be found in Section 6.2.

Simulation of a Molecular Complex The Il8RPLoops peptide is a rationally designed Il-8 capture agent. This peptide is formed of two helices linked with 6-amino-hexanoic acid which is used as a linker (*161*). The Il8RPLoops peptide has a high affinity for Il-8 and inhibits consequently its binding to CXCR1.

We create the structure of the IL8RPLoops, consisting of 18 amino acids, using modeller (*162*). We perform all simulations in `Gromacs` (*158*) using the `Gromos65atb` force field with parameters for the synthetic 6-amino-hexanoic acid linker (*163*). At the beginning of the simulation, the box is filled with TIP3P water, and sodium counter ions are placed until the cell is neutralized. We equilibrate the solvent according to a short steepest descent energy minimization and thus fix the protein-movement for 2 ns in the center. Afterwards, the system was equilibrated for 2 ns in the NVT-ensemble at a temperature of 300 K and for 5 ns in the NpT-ensemble at a temperature of 300 K and a pressure of 1 bar. During the equilibration, temperature was controlled using the velocity-rescale thermostat(*164*) ($\tau_\mathrm{T} = 0.1\,\mathrm{ps}$) and pressure was controlled using the Parrinello-Rahman (*120*) pressure coupling. ($\tau_\mathrm{P} = 0.5\,\mathrm{ps}$). Isothermal compressibility was set to $4.5 \times 10^{-5}\,\mathrm{bar}^{-1}$. Production runs were performed for 100 ns. The temperature was controlled using the Nosé-Hoover thermostat (*119*, *121*) ($\tau_\mathrm{T} = 1\,\mathrm{ps}$) and pressure was controlled using the Parrinello-Rahman barostat (*120*) ($\tau_\mathrm{P} = 1\,\mathrm{ps}$) during the production runs. Bond lengths were constrained using the LINCS (*118*) algorithm. An integration step size of 1.5 fs and a van der Waals interaction cutoff value of 0.95 nm0.95 nm. We recorded a total of 66 667 frames during the simulation.

Simulation of 1HU5 in SPC/E Water As the protein structure we used the solution NMR of ovispirin-1 (*61*). For water molecules, we used the SPC/E model, while the protein interacts through the `CHARMM27` (*104*) force field.

Additionally, we added 0.9 % NaCl solution to the simulation. The system was first energy minimized by conjugate gradient and equilibrated for 2 ns in the NVT-ensemble at a temperature of 300 K and for 5 ns in the NpT-ensemble at a temperature of 300 K and a pressure of 1 bar (for the second simulation 320 K). During the equilibration, temperature was controlled using the velocity-rescale thermostat(*164*) ($\tau_\mathrm{T} = 0.1\,\mathrm{ps}$) and pressure was controlled using the Parrinello-Rahman (*120*) ($\tau_\mathrm{P} = 0.5\,\mathrm{ps}$) and the isothermal compressibility was set to $4.5 \times 10^{-5}\,\mathrm{bar}^{-1}$. Production runs were perfomed for 100 ns. The temperature was controlled using the Nosé-Hoover thermostat (*119*, *121*) ($\tau_\mathrm{T} = 1\,\mathrm{ps}$) and pressure was controlled using the Parrinello-Rahman barostat (*120*) ($\tau_\mathrm{P} = 1\,\mathrm{ps}$) during the production runs. Bond lengths were constrained using the LINCS (*118*) algorithm. The Lennard-Jones nonbonded interactions were evaluated using a cutoff distance of 1.4 nm. The electrostatic interactions were evaluated using the particle mesh Ewald method with a real space cutoff 1.4 nm and a grid-spacing 0.12 nm. The equations of motion were integrated using a 2 fs time step.

Transforming MD Trajectories into Dynamic Graphs

In this section, we describe the process of transforming the trajectories obtained from the MD simulations into dynamic graphs.

Transformation for SPC/ E Water in a Box

We transform all 27 MD trajectories from simulations with different temperatures and water densities into dynamic graphs. All 648 hydrogen and oxygen atoms that form the 216 water molecules are represented as vertices. We create edges between atoms whose distance is below the threshold d of 0.19 nm. As a result, hydrogen atoms should always be connected to their corresponding oxygen atom. The existence of edges between hydrogen atoms of the same water molecule depend on their current movement and can be dissolved in case their distance increases significantly. The distances between all three atoms within a individual water molecule are below the threshold of $d = 0.19$ nm most of the time. Therefore, most water molecules form a motif of type m_2, i.e., are connected by three edges. In case the angle between the hydrogen atoms increases significantly, the distance between them increases such that the atoms form a motif of type m_1 instead. In addition, m_1 and m_2 also occur during the

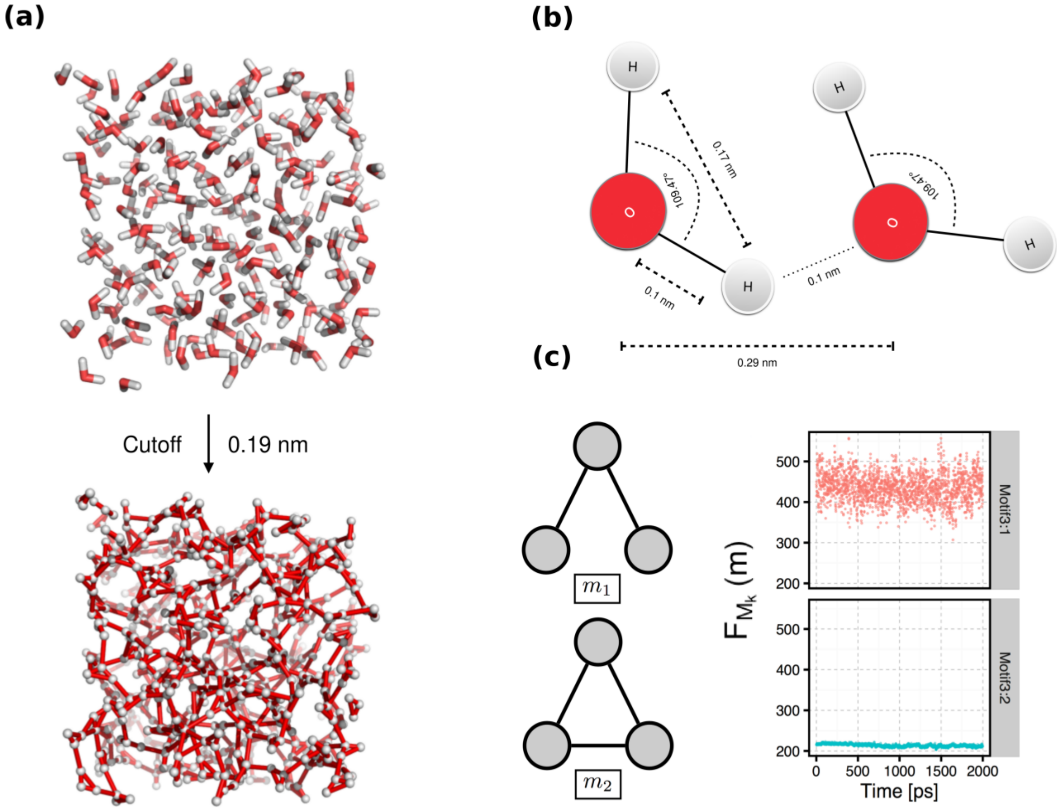


Figure 6.2: Spatial measures and visualizations of SPC/E water in a box

interaction of two or three water molecules. Therefore, we can quantify the number of interactions between water molecules as the the sum of both motif counts without the number of water molecules. Hence, we define the *number of interactions* as

$$I = F_{\mathcal{M}_3}(m_1) + F_{\mathcal{M}_3}(m_2) - \frac{|V|}{3} \tag{6.1}$$

and accordingly for each snapshot, we compute the average number of interactions I_{avg}, defined as

$$I_{avg} = I \cdot (\frac{|V|}{3})^{-1} = \frac{F_{\mathcal{M}_3}(m_1) + F_{\mathcal{M}_3}(m_2) - \frac{|V|}{3}}{\frac{|V|}{3}}. \tag{6.2}$$

The average of these values for all snapshots of a dynamic graph is defined as $\overline{I_{avg}}$.

6.2.1 Results

In this section, we analyze MD trajectories using the motif counts of their resulting dynamic graphs. We analyze the thermodynamics of SPC/E water and compare it with experimental entropy values in Section 6.2.1. In Section 6.2.2, we compare the dynamics of water near a pore surface to the water (bulk) in its center.

Thermodynamics of Water Molecules

In this section, we investigate the capability of graph-based analysis of MD trajectories to reflect the properties of hydrogen bonding networks and their complex rearrangements at different temperatures and densities – a necessary first approach to obtain coarse-grained models of configuration based water models. The importance stems from the fact that biological systems like cells, proteins, or other macro molecules are immersed in water. Its thermodynamic surface properties provide valuable insights into its role in various biological processes, such as protein folding and ligand binding (*165*). For this reason, there has been increasing interest in the evaluation of the entropy of water (and other solvents). Several methods have been proposed to determine it from the results of MD simulations (*166*, *167*). Examples are approaches based on perturbation theory, Kirkwood-Zwanzig thermodynamic integration, and Widom particle insertions (*168*). However, large computational demands render these methods impractical for large systems.

We use the counts of 3-vertex motifs (cf. Figure 6.3) to derive the number of interactions of water molecules and compare the results with experimental entropy values. Therefore, we use MD trajectories from 27 simulations of 216 water molecules in a cubic box with different temperatures and densities. Temperature and volume remain constant during each simulation. The temperatures range along the vapor-liquid saturation line from the triple point at 273.16 K to the critical point at 647.29 K and the densities between 0.00000485 kg/m^3 and 0.9998 kg/m^3. In experiments, entropy values between 61.21 $J/mol\,K$ and 227.89 $J/mol\,K$ have been observed for these configurations[2]. A complete list for the values of all 27 simulations is given in 10.3. For most frames, the count of m_2 is higher than 216 with an average around 225. This indicates close interactions between water molecules (cf. Figure 6.2). In

---

[2] Experimental data is taken from *Fundamentals of classical thermodynamics* (*169*)

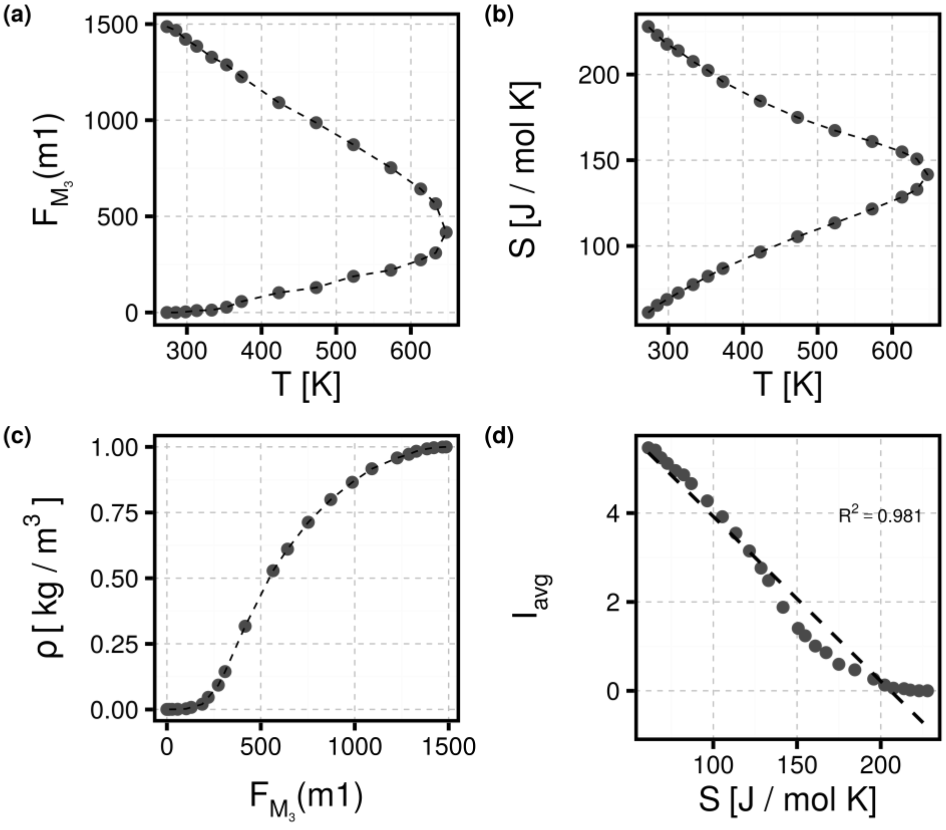


Figure 6.3: Comparison of thermodynamic entropy and motif-based interactions **(a)** Temperature against number of interactions. **(b)** Density against number of interactions. **(c)** ρ against raw motif count. **(d)** Experimental entropy against number of interactions.

some cases, the counts drops below 216, which can be explained by hydrogen atoms moving further away from each other such that their distance is above the threshold of 0.19 nm.

In 6.3, we present the results of our analysis and show the relations between the number of interactions and the physical properties temperature, density, and entropy. Furthermore, we contrast the number of interactions measured using our graph-based analysis to the entropy values measured during experiments. All simulations result in a two-phase thermodynamic model (cf. Figure 6.3). We observe that the number of interactions exhibits a high linear correlation with experimental entropy values over the whole temperature range for the liquid phase (p-value $= 4.284 \cdot 10^{-16}$, Pearson r=0.981). In the case of vapor, experimental values differ a bit from the number of interactions. This can be explained by the process of the distance-dependent graph transformation. At

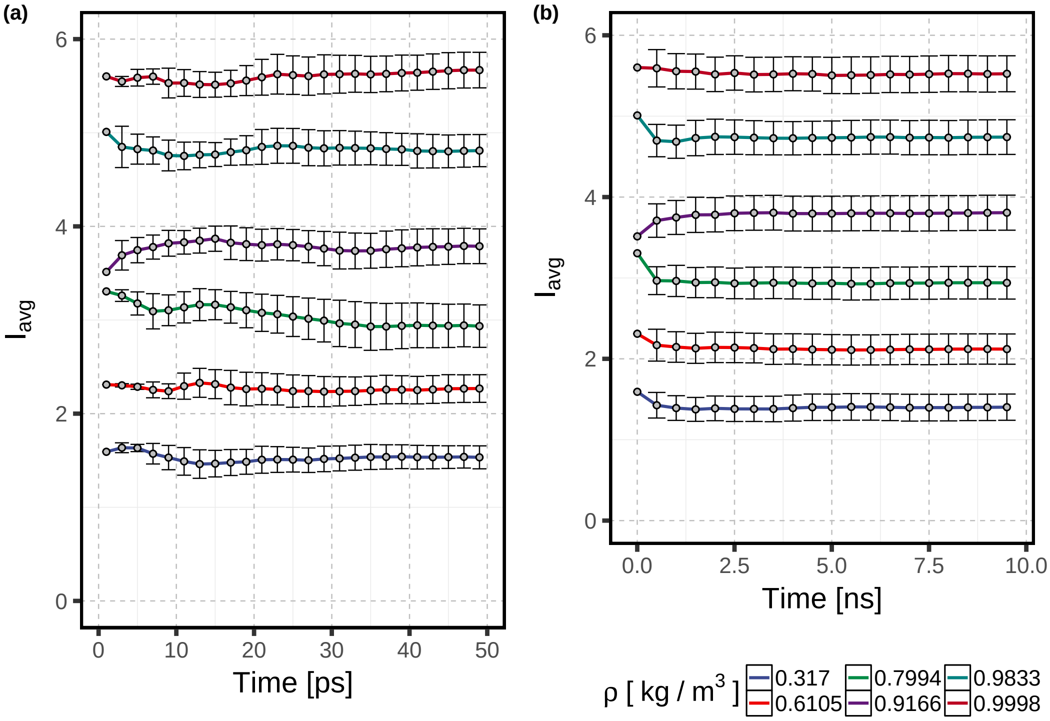


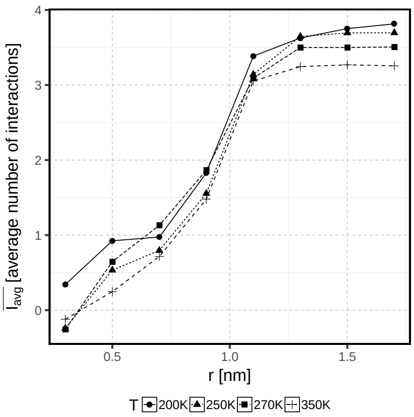
Figure 6.4: Convergence of the number of interactions for different ρ of water simulations. Left higher resoultion (ps scale), right ns scale.

low densities, the distances between molecules are high, leading to a graph that is not fully connected and resulting in a decrease of interactions and thereby motif occurrences. These observations are reflected by the densities against the number of interactions (cf. Figure 6.3). A particularly attractive feature of the graph-based analysis is its fast convergence for the number of interactions over time in case the system is well equilibrated. In Figure 6.4, we show the mean values and standard deviations of the number of interactions for time intervals of 100 ps for systems at six different temperatures. We observe that the number of interactions of liquid water converges after 10 ps to 100 ps.
We observed a high correlation between the number of interactions, measured using graph-based analysis, and the experimental entropy values of all 27 simulations of SPC/E water. With the rapid conversion of this property, we have shown that our graph-based approach for the analysis of MD trajectories originating from simulations of SPC/E water is applicable for studying the thermodynamics of water, including order and dynamics.

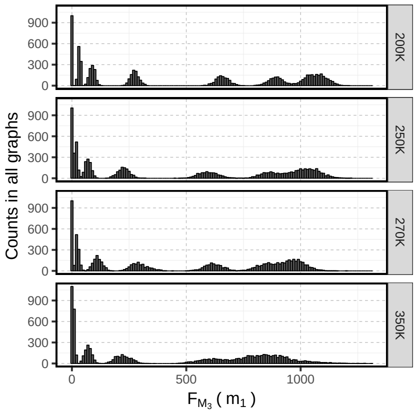
6.2.2 Dynamics of Water in Confinement of Minerals

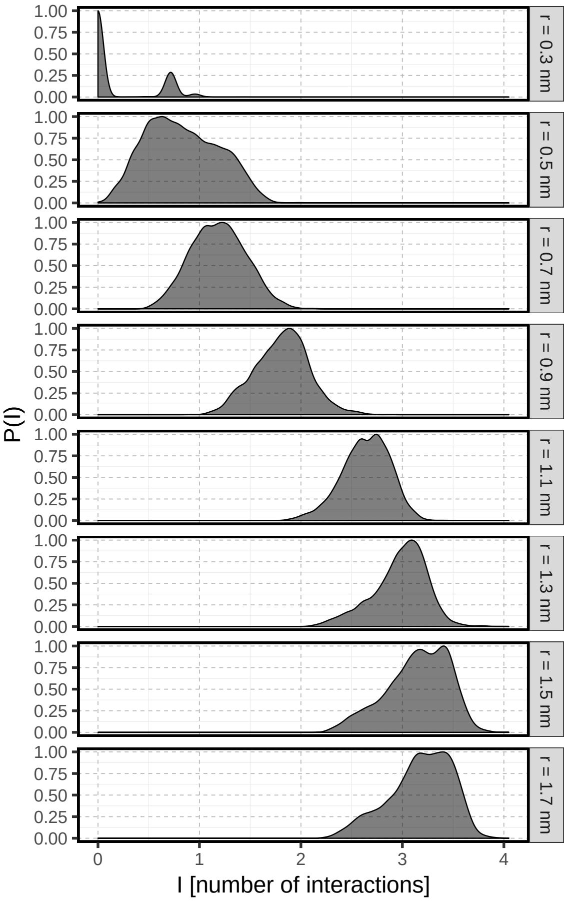
In this section, we analyze the dynamics of water in mineral confinement. We use trajectories from simulations of SPC/E water in a small silica pore at four temperatures: 200 K, 250 K, 270 K, and 350 K. For each frame, we model the atoms of water molecules as vertices that are located at distances $d \in [r', r]$ of the pore wall. Vertices are interconnected with distance thresholds $d = 0.19$ nm. We use nine different distance intervals from $[0.0, 0.3]$ nm, $[0.3, 0.5]$ nm, …, $[1.5, 1.7]$ nm and obtain a total of 32 dynamic graphs. The walls of the silica pore are rough such that the silicon atoms at its surface are located at distances between 0.3 nm and 1.7 nm from its axis. With an increasing distance to the pore wall, the number of vertices increases (cf. Table 10.4). This implies that the density increases in the pore center. Figure 6.5a shows $\overline{I_{avg}}$, the mean average number of interactions of all snapshots for all 32 dynamic graphs. $\overline{I_{avg}}$ increases with the distance to the pore wall. $\overline{I_{avg}}$ is a value which characterizes the motif connectivity of an average water molecule. A high value indicates a higher order and thus a smaller entropy thermodynamical $\mathcal{S}$ of all water molecules. The results obtained for the structure of water near the pore wall reveals imposed disorder in terms of $\overline{I_{avg}}$ values. Figure 6.5b shows the counts of $F_{\mathcal{M}_3}(m_1)$ observed for the dynamic graphs modeled from the MD trajectories at different temperatures. Figure 6.5b shows that $\overline{I_{avg}}$ decreases with increasing temperatures. This implies that the structural disorder – expressed later as the entropy of motifs – grows with increasing temperatures. These results are in good agreement with previous studies of the pore by Harach et. al (*160*). Figure 6.5c shows the raw motif counts as a histogram for all dynamic graphs with increasing distances from the pore surface at a specific temperature as a histogram. These results show that an increase in temperature leads to a flatter distribution and a disappearance of distinctive peaks. Hence, higher temperatures result in higher fluctuations of the absolute motif counts.

Now, we examine how the water dynamics are affected by the pore wall. To this end, we consider the structural order of the hydrogen bonding network inside the silica pore, expressed by the average number of interactions. We introduce the *motif-based entropy* $H_{\mathcal{M}_k}$ of a dynamic graph as a measure of its disorder. It quantifies the fluctuation of a time series and, thus, the dynamics of structural changes. Let P_m denote the relative counts of motif $m \in \mathcal{M}_k$



(a) $\overline{I_{avg}}$ depending on temperature





(b) Counts of $F_{\mathcal{M}_3}(m_1)$ depending on pore wall distance r

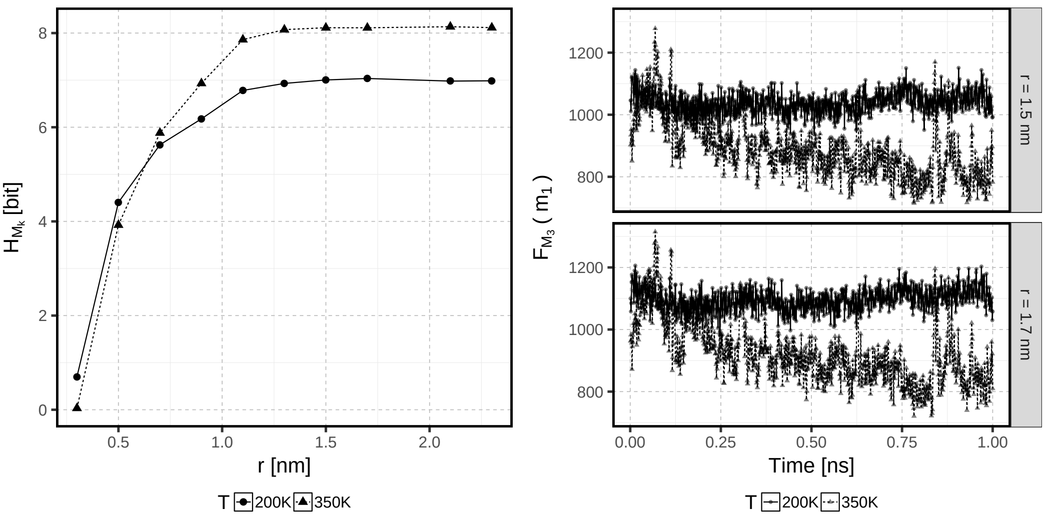
(c) Counts of I_{avg} depending on pore wall distance r

Figure 6.5: Motif-based properties depending on pore wall distance r and temperature

among all motifs occurring for all snapshots of the dynamic graph. Then, we define $H_{\mathcal{M}_k}$ as follows:

$$H_{\mathcal{M}_k} = -\sum_{m \in \mathcal{M}_k} P_m \cdot \log_2 P_m \tag{6.3}$$

In Figure 6.6a, the motif-based entropies for temperatures 200 K and 350 K is shown depending on the pore wall distance r. It is computed for 1,000 snapshots only, a time interval that corresponds to 1 ns. The counts of motif m_1 over time is shown in Figure 6.6b. At the pore surface, expressed by small entropy values, we observe water molecules with slow dynamics. The slowdown of dynamics near the silica wall can be explained by its structural relaxation which is hindered by an atomically rough and mainly static energy landscape



(a) Motif-based entropy $H_{\mathcal{M}_3}$ (b) Counts of m_1 over time

Figure 6.6: Motif-based results for simulations of water in mineral confinement

imposed by the mostly fixed wall atoms (*160*). All the results above are in agreement with previous MD studies on confined water (*160*).

Structure of an Helical peptide in SPC/E Water

In Section 6.1, 4-vertex motif counts are used to quantify structural elements (mainly α-helices) and their dynamics. However, as an example, dynamic graphs from a large esterase (pNB-Est13) were used, which has a multitude of structural elements. In order to investigate that motives are suitable for determining structural dynamics, a toy system consisting of exactly one structural element was required (e.g.,α-helices). For this reason, a small peptide (the solution NMR of ovispirin-1 (*61*)) has been selected which possesses a natural α-helical configuration (cf. Section 6.2).

Observations This peptide was simulated using MD with a temperature of 300 K (blue) and with 320 K (red). The peptide begins to unfold during the simulation with 320 K and starts refolding in a sheet-like structure at the end fo the simulation. This highly dynamic system exhibits three different secondary structural configurations, namely the helical structure, the bent-helix and the sheet-like structure. During the simulation with 300 K, the peptide remains in the α-helical configuration.

Network Transformation The peptide consists of 18 amino acids, each of which is combined to form a vertex. Therefore we are using each C-α as a vertex and create undirected edges between two vertices's in case their spatial cut-off (d) is shorter than $d \in [0.6, 0.8]$.

The resulting graphs are shown in Figure 6.7 (cf. frames at: 5, 25, 50, 75 ns). For this example we examine whether 7-vertex motifs are suitable for describing dynamic processes and structure transitions and compare this approach with classical methods like RMSD (Equation (4.16)). We have chosen 7-vertex motifs because they can display more complex topologies due to the additional vertices's. Furthermore, there exist far more (Number of motifs: 853) 7-vertex than 4-vertex motifs (number of motifs: 6). A visualization of all considered 7-vertex motifs are illustrated in Figure 10.1.

Evaluation For this "toy-model," we expect the 7-vertex motif counts, where motif counts with a high number of edges increase, or remain constant for stable structures (structure elements). Conversely, we expect the reverse for motifs with a few edges. In addition, we also expect to find strongly connected motifs which, due to their unique topology, are only found in helices.

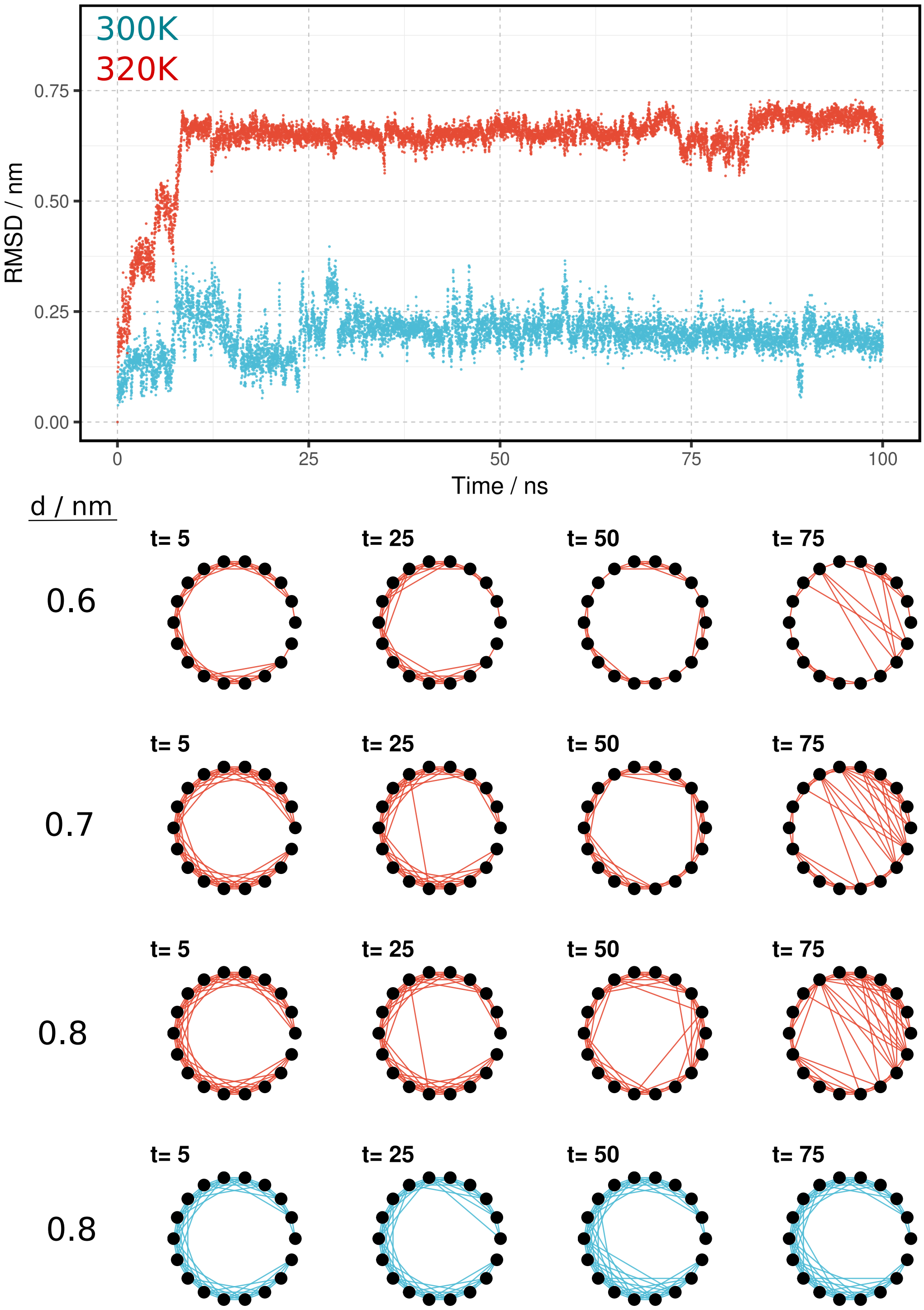


Figure 6.7: RMSD and graph-based representation of an helical molecule at two different temperatures. Blue stands for the low temperature simulation and red for the high temperature simulation.

If we focus on Figure 6.7, we can observe that the helix graph has a unique pattern. In this pattern, only defined motifs appear or do not appear. The lower graphs are sorted by increasing cut-off d, and the last one is the one from the simulation with 320 K and obtained with a cut-off $d = 0.8nm$.

One observes that the network becomes denser and denser with increasing d, but the graph of the helix remains constant over the whole simulation period. However, the RMSD values show a completely different impression of these simulations. The simulation at 300 K is more stable than the simulation at 320 K. This means that the helix is retained from the beginning of the simulation over the entire duration of the simulation at 300 K.

The simulation with the higher temperature shows that this helix changes its conformation into at least one state with an average RMSD of 0.69 nm. However, the course of the RMSD remains at this value showing only small fluctuations. Only at 72 ns, it changes minimally for the duration of 8 ns (cf. Figure 6.7). The fluctuations of the two straight lines indicate that the simulation at 300 K(blue) as the RMSD fluctuates slightly more.

At first glance at the graphs (e.g., Figure 6.7, 5 ns to 75 ns), however, a completely new picture emerges. One can observe the simulation process of both peptides by just looking at the graphs. The graph at 320 K displays a transformation into a completely different topology, while the other simulation (blue, 300 K) does not change much from the initial pattern.

In the following part, all 7-vertex motifs were counted. Two dynamic graphs were created for this purpose (300 K blue,320 K red; Figure 6.8).

However, out of total 853 motifs only 45 motifs with an average count above ten could be found. Since the simulation with 300 K remains in helical conformation for the duration of the whole simulation, this means that the remaining m_{808} motif classes do not occur in a helical topology. Figure 6.8 depicts motifs which display a similar course over time.

Almost all motifs displayed in Figure 6.8 (counts) show a stable course during the MD simulation at 300 K. This result reinforces the hypothesis of Section 6.1, that the course of special motif counts in stable structures remains stable over time and these are only present in a helical secondary structure. The only exception here is the motif with class m_{346}, this motif is also very strongly connected. Interestingly, the mean count for motif m_{345} shows a similar pattern as the rest of the motifs. If we focus on the topology of those motifs, it is noticeable that motif m_{346} differs only in one edge between vertex

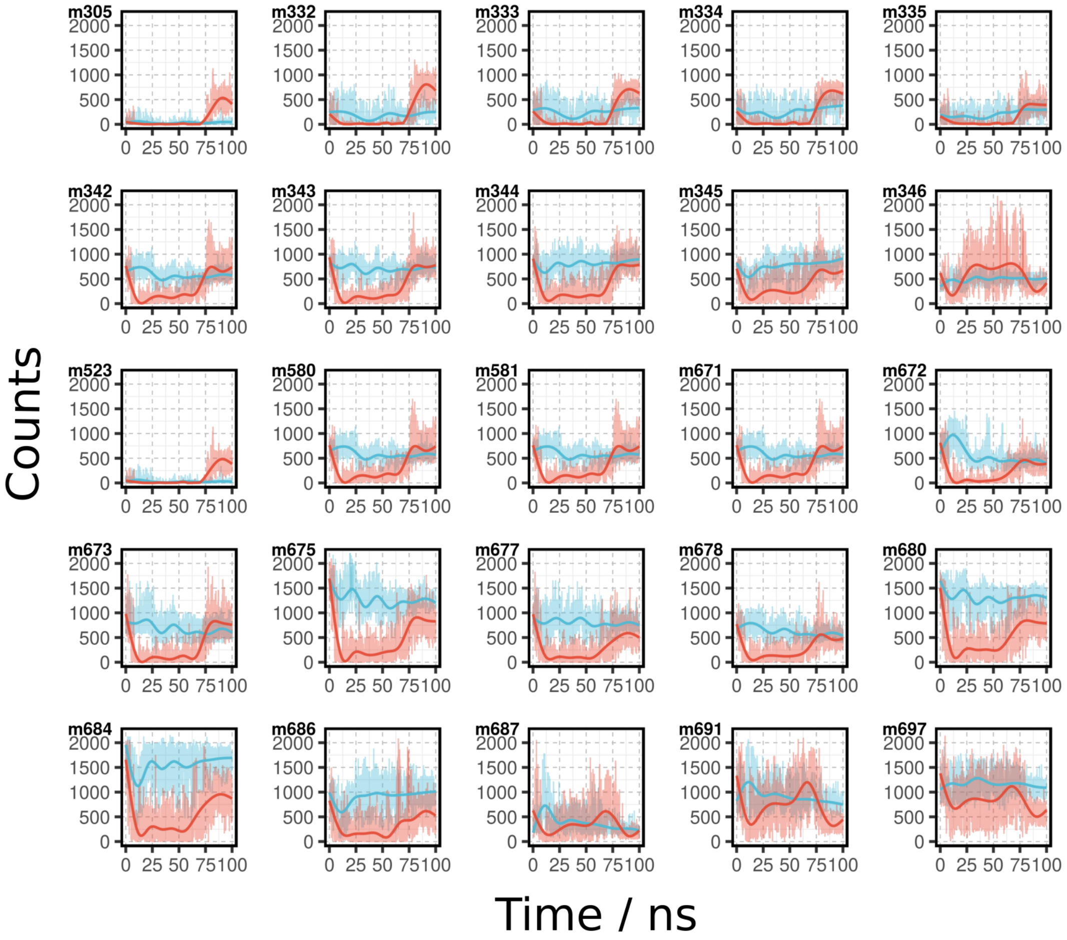


Figure 6.8: 7-vertex motif counts as a function of time for two 100 ns simulation of *1hu5* (300 K, blue; 320 K red)

one and six (cf. Figure 10.1). Inside an α-helix, these amino acids would not interact so often (dependent on d) because of the helical twist.

Motif m_{346} is a perfect example of a motif, in which due to its topology, is also found in other densely packed structural parts (e.g., helix-bends or junctions). Nevertheless, exactly when the helix starts to unfold, the molecule also starts to bend. At this location, the motifs are more frequent because the amino acid contacts increase. Moreover, the observed fluctuations, in particular, give an additional indication of the increasing configuration dynamics at this point.

Figure 6.9 shows the motif counts of the remaining motifs. Most of them are strongly connected (cf. Figure 10.1). Motif class m_{824}, m_{820}, m_{703}, m_{582}, m_{849}, m_{853} and m_{845} shows a similar course here. It looks like there's a huge "peak" in the middle, which fluctuates very strong (320 K; red). The count of

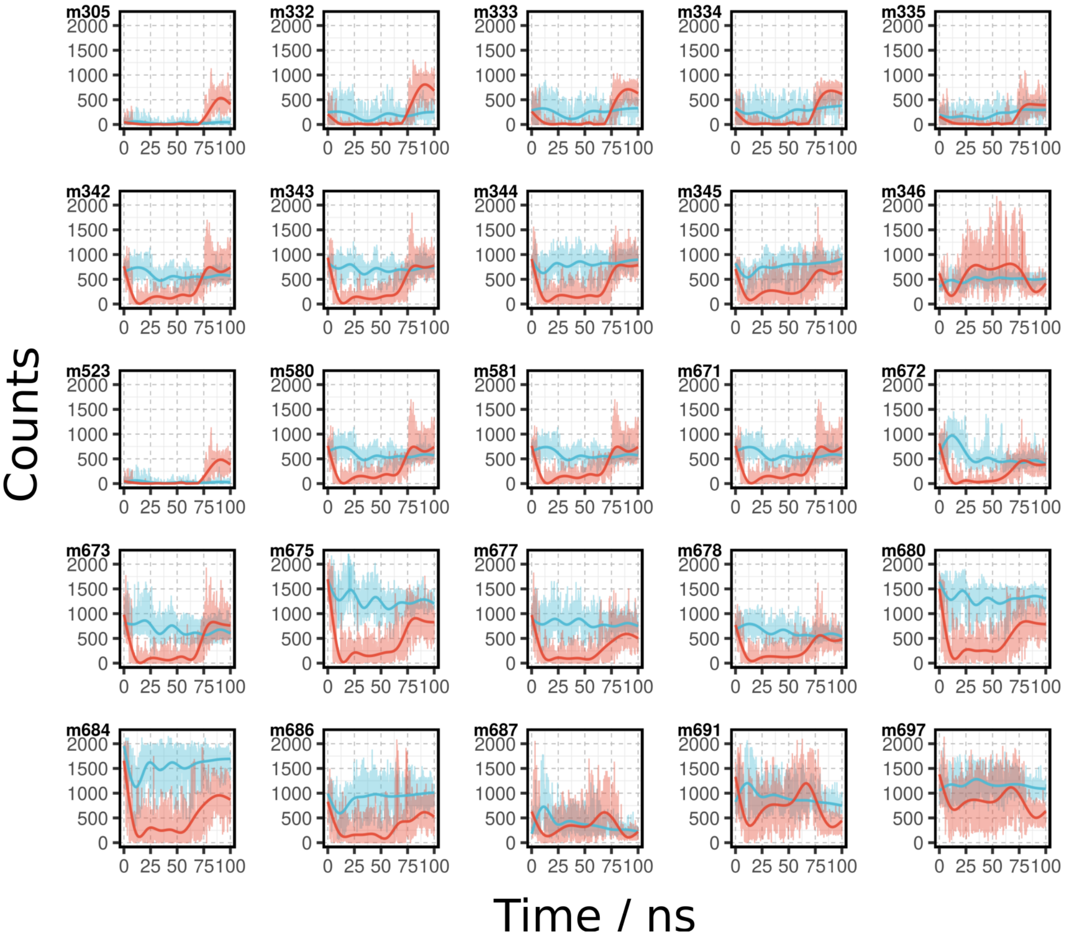


Figure 6.9: 7-vertex motif counts as a function of time for two 100ns simulation of *1hu5* (300 K, blue; 320 K red)

the simulation at 300 K remains stable over time (300 K; red). The course is also very similar to that of motif class m_{346}, which indicates that these motifs can also show the course of helix bending well.

In contrast to RMSD, motifs allow for a more detailed view of the structural process that an α-helix goes through at the two different temperatures. This points to the fact that k-motif counts can record precisely this process of folding (thermal) in the form of their increased counts as well as the underlying dynamics in the form of their fluctuations. A motif-based semantic to determine the dynamic of structural elements is much more expressive than distance-based metrics.

6.2.3 Structural Dynamics of a Protein Complex

In this section, we analyze MD simulations of a complex consisting of the Loops peptide and (*Complex*). We compare the counts of a 3-vertex motif with the counts of three 7-vertex motifs. We contrast these results with commonly used distance-based measures to investigate the expressiveness of our motif-based analysis approach. Visualizations of the simulates components at different points in time are shown in Figure 6.10. We present distance-based measures of the MD trajectory over time in Figure 6.11. Figure 6.11a shows the development of the euclidean distance between the centers of mass of both proteins over time. Starting around 2.8 nm, this distance decreases to less than 1.5 nm at the end of the simulation. This can be explained by strong interactions between both molecules. The RMSD, computed for the C_α atoms of Il-8 and the Loops peptide, is shown in Figure 6.11b. This basic measure is often used to monitor structural changes of MD trajectories. It increases rapidly during the second half of the simulation. This observation could potentially lead to the false interpretation that both structures start to unfold or drift away because they were enlarging their conformation (molecular unfolding). However, both proteins merge and form a stable complex.

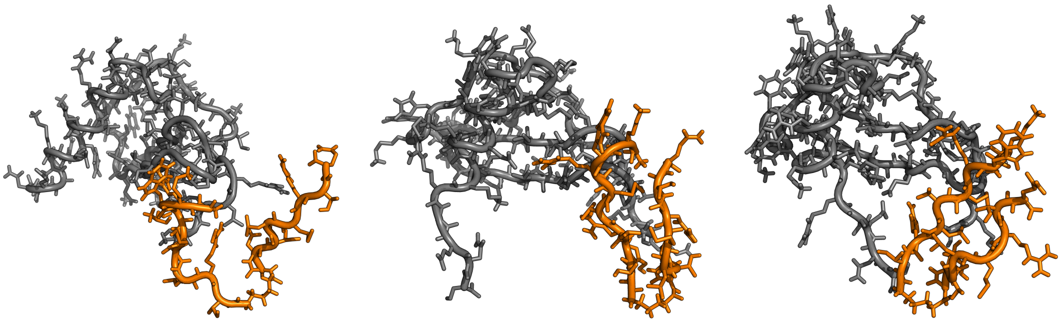
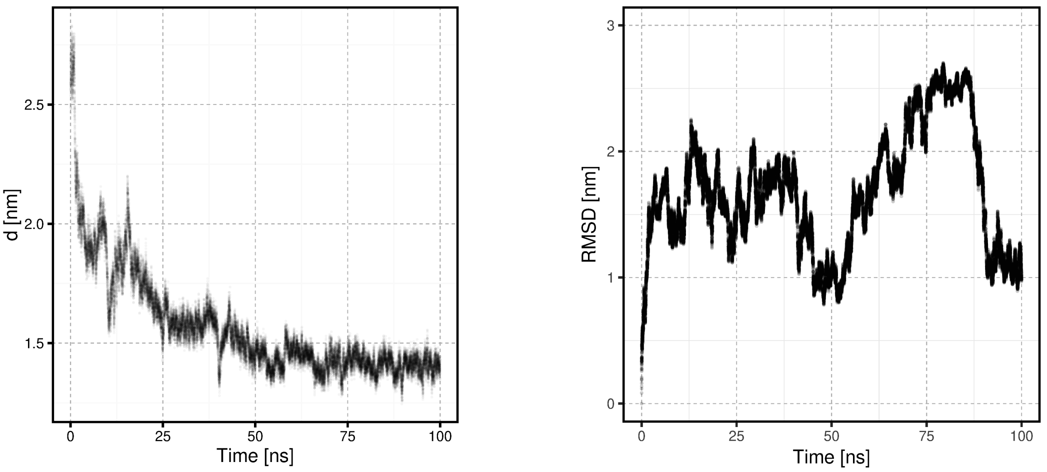

(a) Simulation time: 1 ns (b) Simulation time: 10 ns (c) Simulation time: 20 ns

Figure 6.10: Visualization of the simulated complex over time

We perform a motif-based analysis of the dynamic graph modeled from the MD trajectory using the unit-sphere model with a distance threshold of $d = 0.7$ nm. This threshold is appropriate to measure conformational dynamics in coarse grained models of proteins (*63*). We investigate the occurrences of the 3-vertex motif $m_1 \in \mathcal{M}_3$ (cf. Figure 10.5) with the counts of the three selected 7-vertex motifs $m_{21}, m_{101}, m_{416} \in \mathcal{M}_7$ (cf. Figure 6.12). The counts of all four motifs over time, normalized by their maximum observed values,

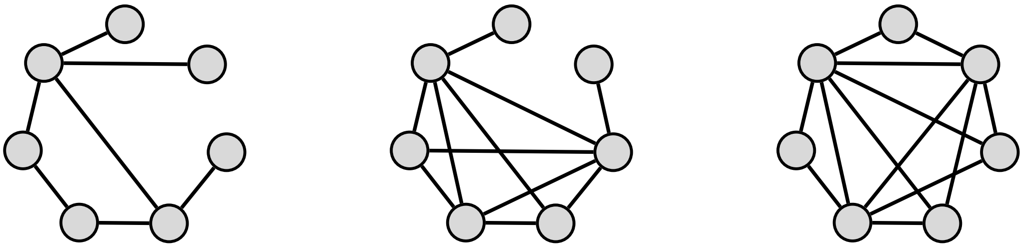


(a) Distance between centers of mass of protein and its binding partner.

(b) Root-Mean-Square Deviation for the full complex.

Figure 6.11: Distance-depended properties of the protein complex over time.

are shown in Figure 6.13. Each point represents the average counts of 600 snapshots while error bars show the respective standard deviation.



(a) m_{21} (7 edges)

(b) m_{101} (11 edges)

(c) m_{416} (13 edges)

Figure 6.12: Selected 7-vertex motifs ($\mathcal{M}_7$)

The relative counts of $m_1 \in \mathcal{M}_3$ does not change significantly over time. This implies that 3-vertex motifs are not well-suited to characterize biomolecular structures even though they express the interactions of molecular solvents well (cf. Sections 6.2.1 and 6.2.2). The counts of the three 7-vertex motifs clearly differ from each other and change over time. The motif m_{21} (cf. Figure 6.12a), connected by 7 edges, nearly disappears after the initial 40 ns. The more densely connected motif m_{101} (cf. Figure 6.12b) does not occur during the initial time frame while its counts increases afterward. The Counts of m_{416}, connected by 13 edges (cf. Figure 6.12c), increases over the whole simulated time period. The disappearance of m_{21} and the appearance of m_{101}

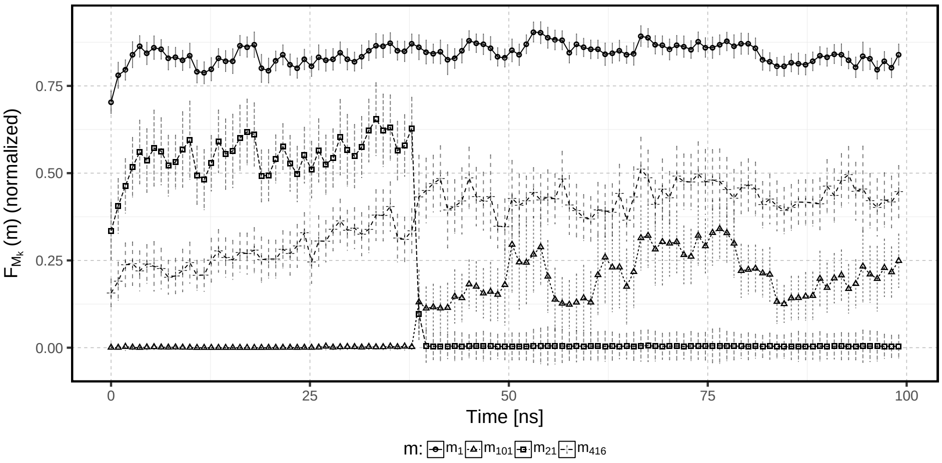


Figure 6.13: Counts of 3-vertex motif $m_1 \in \mathcal{M}_3$ and 7-vertex motifs $m_{21}, m_{101}, m_{416} \in \mathcal{M}_7$ over time for Complex

around 40 ns indicate a significant merging event of the two components of the complex. These results indicate that the increase of occurrences of densely connected motifs goes hand in hand with structural density and complexity. While the small 3-vertex motifs appear too simple to characterize structural changes in protein networks, the larger 7-vertex motifs are well-suited to analyze structural events. Counting motifs with only a few edges provides means to measure unfolding and structural enlargements. The occurrences of densely connected motifs provide insights into small molecular structures and folding processes.

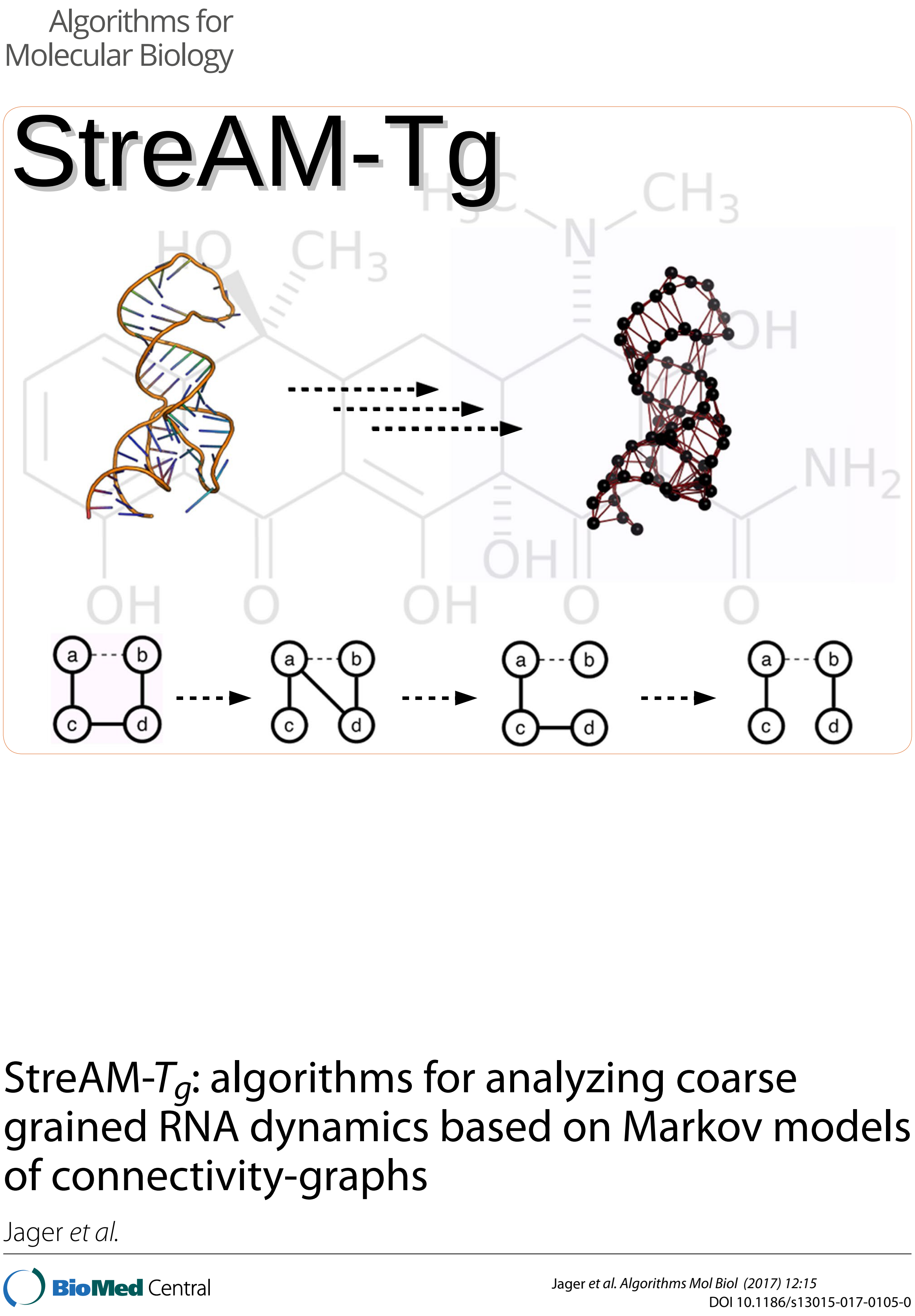
6.3 StreAM-Tg: algorithms for analyzing coarse grained RNA dynamics based on Markov models of connectivity-graphs.

In this work I have developed a new algorithm which calculates MSMs based on any dynamic graph and a given selection of 4-vertex motifs. However, these models could only be set up by expanding the motif space with the whole space of the adjacent matrices. Based on these models, conclusions can be drawn about the entropy of the stationary states of the system. The algorithm was developed especially for the application on RNA simulations. After evaluation of the simulations, design proposals for the riboswitch were derived.

Contributions For this publication, I have integrated my implementations, the `Tg` algorithm and the Algorithm `StreAM` (Benjamin Schiller) into a Julia Library. I was also responsible for the creation of all images and their contents. Furthermore, I took over the writing of the evaluation and the discussion. The `Tg` algorithm was formulated by me, Benjamin Schiller formulated the `StreAM` algorithm and helped me with the evaluation. I was responsible for the simulation of both trajectories and their evaluation and interpretation. Furthermore, I have written and motivated large parts of the paper. In this article, I am the first author.

- **Jager, S.***, Schiller B., Strufe T., Hamacher K. (2016) StreAM-Tg: Algorithms for Analyzing Coarse Grained RNA Dynamics Based on Markov Models of Connectivity-Graphs. In: Frith M., Storm Pedersen C. (eds) Algorithms in Bioinformatics. WABI 2016. Lecture Notes in Computer Science, vol 9838. Springer

After being published by LNCS (Lecture Notes in Computer Science), I was invited to expand the paper and methodology for another publication in *Algorithms for Molecular Biology*. Hence, the paper was extended with eight new MD simulations and a formulation of the algorithm for up to 10-vertex motifs. In this extension, I have optimized the algorithm by a factor of eight with respect to run-time. Furthermore, I performed an analysis of the algorithm with respect to robustness and accuracy and discussed it afterwards. The new MD simulations were carried out by Phillipp Babel and Malte Blumenroth. For every new simulation, I did the evaluation and discussion at this point. Kay Hamacher and Thorsten Strufe helped to write the manuscript and improved it. In this article I am the first author.



Algorithms for
Molecular Biology

StreAM-T_g: algorithms for analyzing coarse grained RNA dynamics based on Markov models of connectivity-graphs

Jager *et al.*

BioMed Central

Jager *et al. Algorithms Mol Biol* (2017) 12:15
DOI 10.1186/s13015-017-0105-0

Jager *et al. Algorithms Mol Biol (2017) 12:15*
DOI 10.1186/s13015-017-0105-0

Algorithms for
Molecular Biology

RESEARCH Open Access

StreAM-T_g: algorithms for analyzing coarse grained RNA dynamics based on Markov models of connectivity-graphs

Sven Jager[1*], Benjamin Schiller[2], Philipp Babel[1], Malte Blumenroth[1], Thorsten Strufe[2] and Kay Hamacher[3]

Abstract

Background: In this work, we present a new coarse grained representation of RNA dynamics. It is based on adjacency matrices and their interactions patterns obtained from molecular dynamics simulations. RNA molecules are well-suited for this representation due to their composition which is mainly modular and assessable by the secondary structure alone. These interactions can be represented as adjacency matrices of *k* nucleotides. Based on those, we define transitions between states as changes in the adjacency matrices which form Markovian dynamics. The intense computational demand for deriving the transition probability matrices prompted us to develop *StreAM*-T_g, a stream-based algorithm for generating such Markov models of *k*-vertex adjacency matrices representing the RNA.

Results: We benchmark *StreAM*-T_g (a) for random and RNA unit sphere dynamic graphs (b) for the robustness of our method against different parameters. Moreover, we address a riboswitch design problem by applying *StreAM*-T_g on six long term molecular dynamics simulation of a synthetic tetracycline dependent riboswitch (500 ns) in combination with five different antibiotics.

Conclusions: The proposed algorithm performs well on large simulated as well as real world dynamic graphs. Additionally, *StreAM*-T_g provides insights into nucleotide based RNA dynamics in comparison to conventional metrics like the root-mean square fluctuation. In the light of experimental data our results show important design opportunities for the riboswitch.

Keywords: RNA, Markovian dynamics, Dynamic graphs, Molecular dynamics, Coarse graining, Synthetic biology

Background

The computational design of switchable and catalytic ribonucleic acids (RNA) becomes a major challenge for synthetic biology [1]. So far, available models and simulation tools to design and analyze functionally complex RNA based devices are very limited [2]. Although several tools are available to assess secondary as well as tertiary RNA structure [3], current capabilities to simulate dynamics are still underdeveloped [4] and rely heavily on atomistic molecular dynamics (MD) techniques [5]. RNA structure is largely modular and composed of repetitive motifs [4] that form structural elements such as hairpins and stems based on hydrogen-bonding patterns [6]. Such structural modules play an important role for nano design [1, 7].

In order to understand RNA dynamics [8, 14] we develop a new method to quantify all possible structural transitions, based on a coarse grained, transferable representation of different module sizes. The computation of Markov State Models (MSM) have recently become practical to reproduce long-time conformational dynamics of biomolecules using data from MD simulations [15].

To this end, we convert MD trajectories into dynamic graphs and derive the Markovian dynamics in the space of adjacency matrices. Aggregated matrices for each nucleotide represent RNA coarse grained dynamics. However, a full investigation of all transitions is computationally expensive.

---

*Correspondence: jager@bio.tu-darmstadt.de
[1] Department of Biology, TU Darmstadt, Schnittspahnstr. 2, 64283 Darmstadt, Germany
Full list of author information is available at the end of the article

© The Author(s) 2017. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

To address this challenge we extend *StreaM*—a stream-based algorithm for counting 4-vertex motifs in dynamic graphs with an outstanding performance for analyzing (bio)molecular trajectories [16]. The extension *StreAM* computes one transition matrix for a single set of vertices or a full set for combinatorial many matrices. To gain insight into global folding and stability of an RNA molecule, we propose *StreAM-*T_g: It combines all adjacency-based Markov models for a nucleotide into one global weighted stochastic transition matrix $T_g(a)$. However, deriving Markovian dynamics from MD simulations of RNA is an emerging method to describe folding pathways [13] or to elucidate the kinetics of stacking interactions [11]. Especially MSM of atomistic aptamer simulations like the theophylline [12] and thrombin aptamer could help to understand structure-function relationships as well as the folding process [18]. Nonetheless, all the methods mentioned above rely on Root Mean Square Deviation (RMSD) computations in combination with clustering in order to identify relevant transition states. For *StreAM-*T_g, the transition states are given by small adjacency matrices representing structural motifs.

The remainder of this paper is structured as follows: In "Our approach for coarse grained analysis", we introduce the concept of *StreAM-*T_g as well as our biological test setup. We describe details of the algorithm in "Algorithm". We present runtime evaluations as well as application scenario of our algorithm in "Evaluation" for a synthetic tetracycline (TC) dependent riboswitch (TC-Aptamer). Furthermore, we investigate the influence upon ligand binding of four different TC derivates and compare them with a conventional method. Finally, we summarize our work in "Summary, conclusion, and future work".

Our approach for coarse grained analysis

Structural representation of RNA

Predicting the function of complex RNA molecules depends critically on understanding both, their structure as well as their conformational dynamics [17, 19]. To achieve the latter we propose a new coarse grained RNA representation. For our approach, we start with an MD simulation to obtain a trajectory of the RNA. We reduce these simulated trajectories to nucleotides represented by their ($C3'$) atoms. From there, we represent RNA structure as an undirected graph [20] using each $C3'$ as a vertex and distance dependent interactions as edges [3]. It is well known that nucleotide-based molecular interactions take place between more than one partner [21]. For this reason interactions exist for several edges observable in the adjacency matrix (obtained via a Euclidean distance cutoff) of $C3'$ coordinates at a given time-step. The resulting edges represent, e.g., strong local interactions such as Watson-Crick pairing, Hoogsteen, or $\pi-\pi$-stacking.

Our algorithm estimates adjacency matrix transition rates of a given set of vertices (nucleotides) and builds a Markov model. Moreover, by deriving all Markov models of all possible combinations of vertices, we can reduce them afterwards into a global weighted transition matrix for each vertex representing the ensemble that the nucleotide modeled as a vertex is immersed in.

Dynamic graphs, their analysis, and Markovian dynamics

A *graph* $G = (V, E)$ is an ordered pair of *vertices* $V = \{v_1, v_2, \ldots v_{|V|}\}$ and *edges* E. We refer to a single vertex of V as a. Here, we only consider *undirected graphs without self-loops*, i.e., $E \subseteq \{\{v, w\} : v, w \in V, v \neq w\}$. We define a self-loop as an edge that connects a vertex to itself. For a subset V' of the vertex set V, we refer to $G[V'] = (V', E')$, $E' := \{\{v, w\} \in E : v, w \in V'\}$ as the V'*-induced subgraph* of G. We refer to the powerset of V as $\mathbb{P}(V)$. The *adjacency matrix* $A(G) = A_{i,j}$ (Eq. 1) of a graph G is a $|V| \times |V|$ matrix, defined as follows:

$$A_{i,j} := \begin{cases} 0 & : i < j \wedge \{v_i, v_j\} \notin E \\ 1 & : i < j \wedge \{v_i, v_j\} \in E \\ \Diamond & : \text{otherwise} \end{cases} \tag{1}$$

Here, the symbol $\Diamond$ denotes for an undefined matrix entry. We denote the set of all adjacency matrices of size k as $\mathcal{A}_k$, with $|\mathcal{A}_k| = 2^{\frac{k\cdot(k-1)}{2}}$. In our current implementation k can takes values in $\{2, 3, 4, 5, 6, 7, 8, 9, 10\}$. With $concat(A)$, we denote the row-by-row *concatenation* of all defined values of an adjacency matrix A. We define the *adjacency id* of a matrix A as the numerical value of the binary interpretation of its concatenation, i.e., $id(A) = concat(A)_2 \in \mathbb{N}$. We refer to $id(V') := id(A(G[V']))$ as the adjacency id of the V'-induced subgraph of G. For example, the concatenation of the adjacency matrix of graph $G_1[V']$ (shown in Fig. 1) is $concat(A(G_1[V'])) = 011011$ and its adjacency id is $id(V') = 011011_2 = 27_{10}$.

As a *dynamic graph* $G_t = (V, E_t)$, we consider a graph whose edge set changes over time. For each point in time $t \in [1, \tau]$, we consider G_t as the *snapshot* or *state* of the dynamic graph at that time. The *transition of a dynamic graph* G_{t-1} to the next state G_t is described by a pair of edge sets which contain the edges added to and removed from G_{t-1}, i.e., (E_t^+, E_t^-). We refer to these changes as a *batch*, defined as follows: $E_t^+ := E_t \backslash E_{t-1}$ and $E_t^- := E_{t-1} \backslash E_t$. The *batch size* is referred as $\delta_t = |E_t^+| + |E_t^-|$ and the average batch size is refered as δ_{avg} and is defined as $\frac{\sum_t \delta_t}{\tau}$.

The *analysis* of dynamic graphs is commonly performed using *stream-* or *batch-based* algorithms. Both output the desired result for each snapshot G_t. Stream-based algorithms take a single update to the graph as

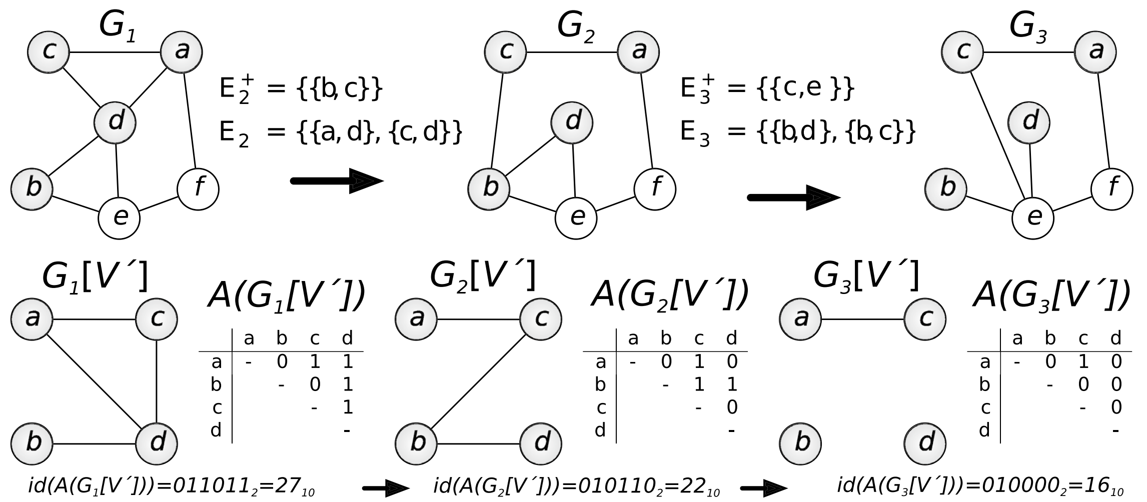


Fig. 1 Dynamic graph example. Example of a dynamic graph and induced subgraphs for $V' = \{a, b, c, d\}$. The *first row* shows the dynamic graph G_t and the second the induced subgraph V' with its respective adjacency matrix. At the *bottom* is a short example of how to compute the adjacency id for the displayed subgraphs

input, i.e., the addition or removal of an edge e. Batch-based algorithms take a pair (E^+_{t+1}, E^-_{t+1}) as input. They can always be implemented by executing a stream-based algorithm for each edge addition $e \in E^+_{t+1}$ and removal $e \in E^-_{t+1}$. We refer to $id_t(V')$ as the adjacency id of the V'-induced subgraph of each snapshot of G_t. The result of analyzing the adjacency id of V' for a dynamic graph G_t is a list $(id_t(V') : t \in [1, \tau])$. We consider each pair $(id_t(V'), id_{t+1}(V'))$ as an *adjacency transition of* V' and denote the *set of all transitions* as $\mathcal{T}(V')$. Then, we define the *local transition matrix* $T(V')$ of V' as a $|\mathcal{A}_k| \times |\mathcal{A}_k|$ matrix, which contains the number of transitions between any two adjacency ids over time, i.e., $T_{i,j}(V') := |(i+1, j+1) \in \mathcal{T}(V')|$ for an adjacency size k. From $T(V')$, we can derive a *Markov model* to describe these transitions.

By combining all possible $T(V')$ where $V' \in \mathbb{P}(V) : |V'| = k$ and $a \in V'$, we derive a transition tensor $C_a(V)$. Thus $C_a(V)$ has the dimensions of $|\mathcal{A}_k| \times |\mathcal{A}_k| \times (k-1)!\binom{|V|}{k-1}$.

We define the weighting matrix $W(V')$ with the dimensions of $|\mathcal{A}_k| \times (k-1)!\binom{|V|}{k-1}$. $W(V')$ contains the weighting for every subset $V' \in C_a(V)$. It is defined as $W(V') := \frac{S(V')}{\sum_{V' \in C_a(V)} S(V')}$. Here, $S(V')$ is a matrix containing the sum of every transition between adjacency $id(V')$ and every other $id(V')$ of the same matrix $T(V')$ for all $V' \in C_a(V)$. Hence $S(V')$ has the dimensions $|\mathcal{A}_k| \times (k-1)!\binom{|V|}{k-1}$. Thus $W(V')$ is considered as the local distribution weighted by its global distribution of transitions matrices of V'. Finaly, we define a global transition matrix, a vertex a is immeresd in, as $T_g(a) = \sum_{V' \in C_a(V)} W(V') \times T(V')$ with the dimensions $|\mathcal{A}_k| \times |\mathcal{A}_k|$.

For a local or global transition matrix the respective dominant eigenvector[1] is called π and represents the stationary distribution attained for infinite (or very long) times. The corresponding conformational entropy of the ensemble of motifs is $H := -\sum_i \pi_i \cdot \log \pi_i$. The change in conformational entropy upon, e.g., binding a ligand is then given as $\Delta H = H_{wt} - H_{complex}$.

MD simulation setup

We use a structure of a synthetic tetracycline binding riboswitch (PDB: 3EGZ, chain B, resolution: 2.2 Å, Fig. 2) [23] and perform six simulations: the TC-Aptamer with five different tetracycline types in complex and one without tetracycline. As tetracycline binding alters the structural entropy of the molecule [24] our proposed method should be able to detect changes in (local) dynamics due the presence of tetracycline. All simulations were performed using the GROMACS software package (version 2016). For water molecules, we used the

---

[1] Guaranteed to exist due to the Perron-Frobenius theorem with an eigenvalue of $\lambda = 1$.

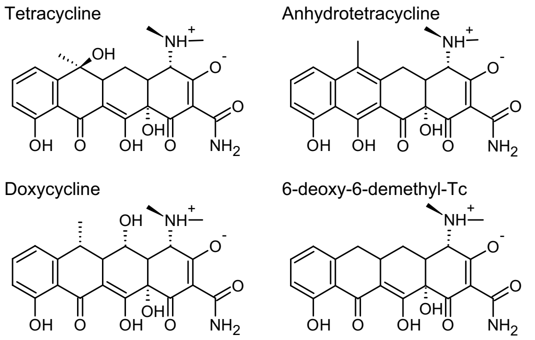


Fig. 2 TC-derivates. TC-derivates illustrated as chemical structures. Here we show the structure of Tetracycline (*left top*), Anhydrotetracycline (*right top*), Doxycycline (*left bottom*) and 6-deoxy-6-demethyl-Tetracycline (*right bottom*). The illustrated derivates share the characteristic 4-ring-structure and functional groups

TIP3P model, the RNA interact through the CHARMM force field, while the tetracycline analogs interact through a modified CHARMM force field from Aleksandrov and Simonson [25, 26]. The systems were first energy minimized and equilibrated for 1 ns in the NVT-ensemble at a temperature of 300 K and for 5 ns in the NpT-ensemble at a temperature of 300 K and a pressure of 1 bar. During the equilibration, temperature was controlled using the velocity-rescale thermostat [27] ($\tau_T = 0.1$ ps) and pressure was controlled using the Berendsen barostat [28] ($\tau_P = 0.5$ ps). Isothermal compressibility was set to $4.5 \times 10^{-5}\,\text{bar}^{-1}$, which is the corresponding value for water. Production runs were performed for 500 ns. The temperature was controlled using the Nosé-Hoover thermostat [29, 30] ($\tau_T = 1$ ps) and pressure was controlled using the Parrinello-Rahman barostat [31] ($\tau_P = 1$ ps) during the production runs. Bond lengths were constrained using the LINCS [32] algorithm. The Lennard-Jones nonbonded interactions were evaluated using a cutoff distance of 1.2 nm. The electrostatic interactions were evaluated using the particle mesh Ewald method with a real space cutoff 1.2 nm and a grid-spacing 0.12 nm. Long-range corrections to energy and pressure due to the truncation of Lennard-Jones potential were accounted for. The equations of motion were integrated using a 2 fs time step.

Tetracycline derivates

For the comparison of TC derivates we use tetracycline (tc), doxycycline (dc), anhydrotetracycline (atc) and 6-deoxy-6-demythyltetracycline (ddtc) in our MD simulation. These four analogs share the characteristic 4-ring-structure and functional groups of all tetracyclines. Still, the possibility and the mode of interaction with the RNA is an open question. The first ring of tetracycline carries a dimethylamino group, while the third ring carries a hydroxy and a methyl group facing towards the same direction away from the 4-ring-system. The detailed chemical structures are shown in Fig. 3. In comparison to these two rings the fourth, aromatic ring has an especially small steric volume on this side of the molecule. From tc over dc and atc to ddtc this steric volume is further reduced by shifting the aforementioned hydroxy and methyl group away from the fourth ring or eliminating some of them entirely. Note, that our graph-based approach is capable to easily distinguish between different modes of interaction upon changes in the, e.g., the side-chains of the rings. The molecular data of tc, dc, atc and ddtc was created using the Avogadro software [33]. Structures were manually constructed and moved into the extended conformation described to be 3 kcal/mol more stable than its twisted alternative by Alexandrov et al. [24]. The molecules were then fitted to the position of 7-chlorotetracycline (7-cl-tc) bound in the TC-Aptamer structure used for simulation. Note, that the geometry of 7-cl-tc was already present in the crystal structure of the TC-Aptamer. All considered antibiotics show different properties upon ligand binding. They range from high activity (tc, 7-cl-tc) to weak activity (dc, ddtc, atc) based on in vivo experiments [34].

Workflow

RNA trajectory and contact probability

An RNA trajectory X is represented as a list of T frames $X = (\vec{x}_{t_0}, \vec{x}_{t_1}, \ldots)$. Each frame $\vec{x}_t \in \mathbb{R}^{3n}$ contains the three-dimensional coordinates of the simulated system of the n atoms at the respective point in time t. We define a binary contact matrix $B(t)$ with dimensions $|V| \times |V|$. Its entries scan range between $\{0, 1\}$. A single contact $B_{i,j}(t)$ between one pair of atom coordinates $\vec{r}_i(t)$ and $\vec{r}_j(t)$ is generated if their Euclidean distance [L2-norm, $L2(\ldots)$] is shorter than d. Thus $B(t)$ entries are defined as follows:

$$B_{i,j}(t) := \begin{cases} 0 & : d < L2(\vec{r}_i(t) - \vec{r}_j(t)) \\ 1 & : d > L2(\vec{r}_i(t) - \vec{r}_j(t)) \end{cases} \qquad (2)$$

The contact probability of one pair of atom coordinates $\vec{r}_i$ and $\vec{r}_j$ is defined as:

$$P(X, \vec{r}_i, \vec{r}_j) = \frac{\sum_{t=1}^{T} B_{ij}(t)}{T}. \qquad (3)$$

Graph transformation

All considered MD simulations have a total length of 500 ns using an integration stepsize of 2 fs. We created snapshots every 250 ps resulting in 100,000 frames. We generated dynamic graphs $G_t = (V, E_t)$ containing $|V| = 65$ vertices (Table 1), each modelling a nucleic $3C'$ (Fig. 2). This resolution is sufficient to represent both small secondary structure elements as well as large

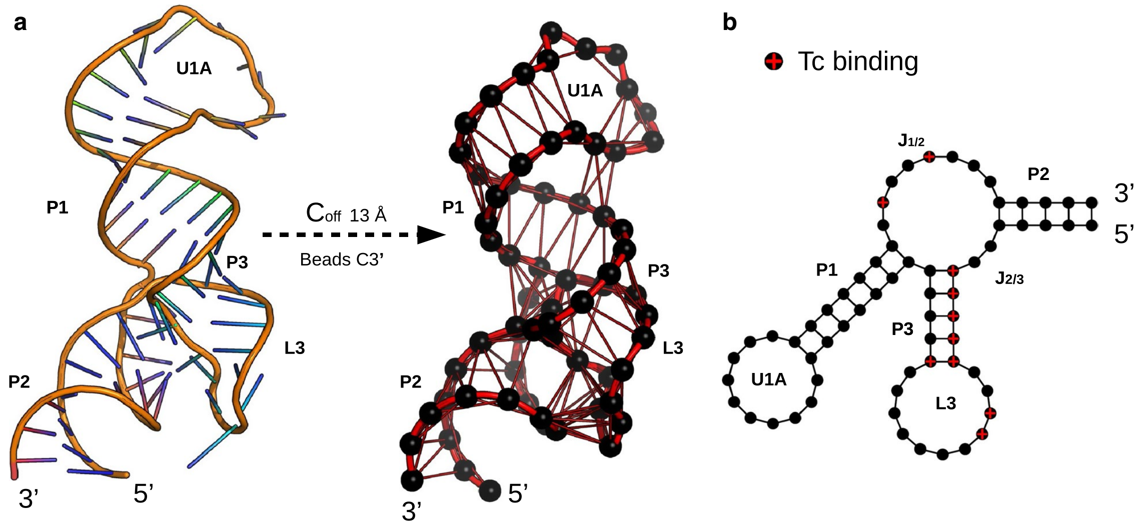


Fig. 3 Structural representation of TC-Aptamer. **a** Crystal structure of TC-Aptamer with a cut-off of 13 Å and using *C*3′ atom for coarse graining reveals edges for dominant WC base-pairings. Important structural parts are annotated according to [23]. **b** Secondary structure representation of TC-Aptamer. Nucleotides are displayed as vertices and connections are based on hydrogen-bonding patterns. Nucleotides participating in TC-binding are colored in *red*. Graphics were created using `Pymol` and `R` [39, 47]

Table 1 Details of the dynamic graphs obtained from MD simulation trajectories

	10 Å	**11 Å**	**12 Å**	**13 Å**	**14 Å**	**15 Å**	**Rand$_{g1}$**	**Rand$_{g2}$**	**Rand$_{g3}$**
$\|V\|$	65	65	65	65	65	65	500	500	500
$\|E\|$	94	129	189	241	298	353	500	1000	1200
δ_{avg}	6.1	15.6	19.4	18	19.6	23.8	80	100	120

$|V|$ is the number of vertices, $|E|$ the number of edges and δ_t is the average batch size of a simulation. We convert simulations to unit sphere dynamic graphs with $d \in [10, 15]$ Å

quaternary RNA complexes [35, 36]. We create undirected edges between two vertices in case their Euclidean cut-off (d) is shorter than $\{d \in N | 10 \leq d \leq 15\}$ Å (cmp. Table 1).

Markov state models (MSM) of local adjacency and global transition matrix

StreAM counts adjacency transitions (e.g. as a set $\mathcal{T}(V')$) of an induced subgraph for a given adjacency size. Now the transition matrix $T(V')$ can be derived from $\mathcal{T}(V')$ but not all possible states are necessarily visited in a given, finite simulation, although a "missing state" potentially might occur in longer simulations. In order to allow for this, we introduce a minimal pseudo-count [37] of $P_k = \frac{1}{|\mathcal{A}_k|}$. All models that fullfill $\{V' \in \mathbb{P}(V) : |V'| = k, a \in V'\}$ have the same matrix dimension and thus can be envisioned to be combined in a tensor $C_a(V)$. Now, $C_{a\ i,j,l}(V)$ is one entry of the tensor of transitions between adjacency *id* i and j in the l th transition matrix $T(V')$ with $|l| = \binom{|V|}{k-1} \times k - 1$. Thus $C_a(V)$ contains all $T(V')$ a specific vertex is immersed in and due to this it contains all possible information of local markovian dynamics. To derive $T_g(a)$ every entry $C_{a\ i,j,l}(V)$ is normalized by the count of all transitions of i in all matrices $S(V)_{j,l} = \sum_i C_{a\ i,j,l}(V)$. For a given set of l transition matrices $T(V')$ we can combine them into a global model with respect to their probability:

$$T_{g\ i,j}(a) = \sum_l \frac{S(V)_{jl}}{\sum_l S(V)_{jl}} \cdot C_{a\ i,j,l}(V). \tag{4}$$

Stationary distribution and entropy

As $T_g(a)$ (Eq. 4) is a row stochastic matrix we can compute its dominant eigenvector from a spectral decomposition. It represents a basic quantity of interest: the stationary probability $\vec{\pi} := (\pi_1, \ldots, \pi_i, \ldots)$ of micro-states i [37]. To this end we used the `markovchain` library in `R` [38, 39]. For measuring the changes in conformational entropy $H := -\sum_{i=1}^{|\mathcal{A}_k|} \pi_i \cdot \log \pi_i$ upon binding a ligand, we define $\Delta H = H_{wt} - H_{complex}$, form a stationary distribution.

Conventional analysis: root mean square fluctuation (RMSF)

The flexibility of an atom can be quantitatively assessed by its *Root-mean-square fluctuation* (RMSF). This measure is the time average L2-norm $L2(\ldots)$ of one particular atom's position $\vec{r}_i(t)$ to its time-averaged position $\bar{\vec{r}_i}$. The RMSF of an nucleotide i (represented by its respective $C3'$ atom) is defined as:

$$RMSF(X, r_i) := \sqrt{\frac{1}{T} \cdot \sum_{t=1}^{T} L2(\vec{r}_i(t), \bar{\vec{r}_i})^2} \qquad (5)$$

Algorithm

Overview

In this section, we introduce the required algorithms to compute $T_g(a)$. First, we describe *StreAM*, a stream-based algorithm for computing the adjacency $id(V')$ for a given V'. Afterwards we describe, the batch-based computation using $StreAM_B$ to derive $id_t(V')$. By computing the adjacency id of a dynamic graph $G_t[V']$ we derive a list $(id_t(V') : t \in [1, \tau])$ where each pair $[id_t(V'), id_{t+1}(V')]$ represents an adjacency transition. The respective transitions are than stored in $\mathcal{T}(V')$. Now, a single $T(V')$ can be derived by counting the transitions in $\mathcal{T}(V')$. At last we introduce *StreAM-T_g*, an algorithm for the computation of a global transition matrix $T_g(a)$ for a given vertex a from a dynamic graph $G_t[V]$. To this end, *StreAM-T_g* computes the tensor $C_a(V)$ which includes every single matrix $T(V')$ where $V' \in \mathbb{P}(V)$ and $|V'| = k$ with vertex $a \in V'$. Finally, *StreAM-T_g* computes $T_g(a)$ from $C_a(V)$.

StreAM and StreAM$_B$.

We compute the adjacency id $id(V')$ for vertices $V' \subseteq V$ in the dynamic graph G_t using the stream-based algorithm *StreAM*, as described in Algorithm 1. Here, $id(V') \in [0, |\mathcal{A}_{|V'|}|)$ is the unique identifier of the adjacency matrix of the subgraph $G[V']$. Each change to G_t consists of the edge $\{a, b\}$ and a type to mark it as addition or removal (abbreviated to *add*,*rem*). In addition to edge and type, *StreAM* takes as input the ordered list of vertices V' and their current adjacency id.

An edge $\{a, b\}$ is only processed by *StreAM* in case both a and b are contained in V'. Otherwise, its addition or removal has clearly no impact on $id(V')$.

Assume $pos(V', a), pos(V', b) \in [1, k]$ to be the positions of vertices a and b in V'. Then, $i = min(pos(V', a), pos(V', b))$ and $j = max(pos(V', a), pos(V', b))$ are the row and column of adjacency matrix $A(G[V'])$ that represent the edge $\{a, b\}$. In the bit representation of its adjacency id $id(V')$, this edge is represented by the bit $(i-1) \cdot k + j - i \cdot (i+1)/2$. When interpreting this bit representation as a number, an addition or removal of the respective edge corresponds to the addition or subtraction of $2^{k\cdot(k-1)/2-((i-1)\cdot k+j-i\cdot(i+1)/2)}$. This operation is performed to update $id(V')$ for each edge removal or addition. In the following, we refer to this position as $e(a, b, V') := \frac{|V'|\cdot(|V'|-1)}{2} - [(i-1) \cdot |V'| + j - \frac{i\cdot(i+1)}{2}]$.

```
Data: $V'$, $id$, $\{a,b\}$, $type \in \{add, rem\}$
begin
    if $a \in V' \wedge b \in V'$ ;                /* process only relevant edges */
    then
        if $type == add$ then
            $A := A + 2^{e(a,b,V')}$ ;             /* set corresponding bit to 1 */
        else
            $A := A - 2^{e(a,b,V')}$ ;             /* set corresponding bit to 0 */
        end
    end
    return $id$ ;
end
```

Algorithm 1: *StreAM*: stream-based computation of the adjacency id

Furthermore, in Algorithm 2 we show *StreAM*$_B$ for the batch-based computation of the adjacency id for vertices V'

```
Data: $V'$, $id_{t-1}$, $E_t^+$, $E_t^-$
begin
    $id_t(V') := id_{t-1}(V')$ ;                              /* init id with previous one */
    for all $\{a,b\} \in E_t^+$ do
        $id_t := StreAM(V', id_t, \{a,b\}, add)$ ;            /* process addition */
    end
    for all $\{a,b\} \in E_t^-$ do
        $id_t := StreAM(V', id_t, \{a,b\}, rem)$ ;            /* process removal */
    end
    return $id_t$ ;
end
```

Algorithm 2: *StreAM*$_B$: batch-based computation of the adjacency id

StreAM-T_g

For the design or redesign of aptamers it is crucial to provide experimental researchers informations about e.g. dynamics at the nulceotide level. To this end, *StreAM*-T_g combines every adajcency-based transition matrix, one nucleotide participates in, into a global model $T_g(a)$. This model can be derived for every nucleotide of the regarded RNA structure and contains all the structural transition of a nuclotide between the complete ensemble of remaining nucleotides. In order to do this, we present *StreAM*-T_g, an algorithm for the computation of global transition matrices, one particular vertex is participating in, given in Algorithm 3. A full computation with *StreAM*-T_g can be divided into the following steps. The first step is the computation of all possible Markov models that fulfill $V' \in \mathbb{P}(V) : |V'| = k$ with *StreAM* for a given k with $k \in [2, 10]$. This results in $\binom{|V|}{k} \cdot k! = \frac{|V|!}{(|V|-k)!}$ combinations. Afterwards, *StreAM*-T_g sorts the matrices by vertex *id* into different sets, each with the size of $\binom{|V|}{k-1} \cdot (k-1)!$. For each vertex a, *StreAM*-T_g combines the obtained $T(V')$ that fulfill $a \in V'$ in a transition tensor $C_a(V)$, which is normalized by $W(V')$ the global distribution of transition states a vertex is immersing in, taking the whole ensemble into account. $W(V')$ can be directly computed from $C_a(V)$ (e.g. "Dynamic graphs, their analysis, and Markovian dynamics")

```
Data: $T, a, k$
begin
    $C_a(V) := \{V' \in \mathbb{P}(V) : |V'| = k, a \in V'\}$ ;      /* $C_a$ vertex $a$ immersed in */
    $T_g(a) := 0_{|\mathscr{A}_k|,|\mathscr{A}_k|}$ ;                 /* initialize $T_g(a)$ */
    for all $V' \in C_a(V)$ do
        $T_g(a) := T_g(a) + W(V') \cdot T(V')$ ;                     /* sum up $T_g(a)$ */
    end
    return $T_g(a)$
end
```

Algorithm 3: *StreAM*-$T_g(a)$ for computing the global transition matrix $T_g(a)$

StreAM-T_g optimization using precomputed contact probability

The large computational demands for a full computation of the $\binom{|V|}{k} \cdot k! = \frac{|V|!}{(|V|-k)!}$ transition matrices to derive a set of $T_g(a)$, motivated us to implement an optimization: The number of Markov models can be reduced by considering only adjacencies including possible contacts between at least two vertices of $G_t = (V, E_t)$. This can be precomputed before the full computation by considering the contact probability $P(X, \vec{r}_i, \vec{r}_j)$ between vertices. To this end we only compute transition matrices forming a contact within the dynamic graph with $P(X, \vec{r}_i, \vec{r}_j) > 0$.

Evaluation

Objectives

As *StreAM-T_g* is intended to analyze large MD trajectories we first measure the speed of *StreAM* for computing a single $\mathcal{T}(V')$ to estimate overall computational resources. With this in mind, we benchmark different G_t with increasing adjacency size k (Table 1). Furthermore, we need to quantify the dependence of computational speed with respect to δ_t. Note, δ_t represents changes in conformations within G_t. For the full computation of $T_g(a)$, we want to measure computing time in order to benchmark *StreAM-T_g* by increasing network size $|V|$ and k for a given system due to exponentially increasing matrix dimensions $|\mathcal{A}_k| = 2^{\frac{k\cdot(k-1)}{2}}$ ($k = 3$ 8, $k = 4$ 64, $k = 5$ 1,024, $k = 6$ 32,768, $k = 7$ 2,097,152 size of matrix dimensions). We expect due to combinatorial complexity of matrix computation a linear relation between $|V|$ and speed and an exponential relation between increasing k and speed. To access robustness of influence of d robustness regarding the computation of $T_g(a)$ stationary distribution $\vec{\pi}$. We expect a strong linear correlation between derived stationary distributions. Details are shown in "Robustness against threshold". We compare Markovian dynamics between the native TC-Aptamer and the structure in complex with 7-cl-tc with experimental data. We discuss the details in "Workflow" and "Application to molecular synthetic biology". Furthermore, we want to illustrate the biological relevance by applying it to a riboswitch design problem; this is shown in detail in "Application to molecular synthetic biology". For the last part, we investigate the ligand binding of four different TC derivates using *StreAM-T_g* and compare them with a classical metric (e.g. RMSF) in "Comparison of tetracycline derivates".

Evaluation setup

All benchmarks were performed on a machine with four *Intel(R) Xeon(R) CPU E5-2687W v2* processors with 3.4GHz running a Debian operating system. We implemented *StreAM* in Java; all sources are available in a GitHub repository.[2] The final implementation *StreAM-T_g* is integrated in a `Julia` repository.[3] We created plots using the `AssayToolbox` library for `R` [39, 40]. We generate all random graphs using a generator for dynamic graphs[4] derived for vertex combination.

Runtime dependencies of StreAM on adjacency size

For every dynamic graph $G_t(V, E_t)$, we selected a total number of 100,000 snapshots to measure *StreAM* runtime performance. In order to perform benchmarks with increasing k, we chose randomly nodes $k \in [3, 10]$ and repeated this 500 times for different numbers of snapshots (every 10,000 steps). We determined the slope (speed $\frac{frames}{ms}$) of compute time vs. k for random and MD graphs with different parameters (Table 1).

Runtime dependence of StreAM on batch size

We measured runtime performance of *StreAM* for the computation of a set of all transitions $\mathcal{T}(V')$ with different adjacency sizes k as well as dynamic networks with increasing batch sizes. To test *StreAM* batch size dependencies, 35 random graphs were drawn with increasing batch size and constant numbers of vertex and edges. All graphs contained 100,000 snapshots and k is calculated from 500 random combinations of vertices.

Runtime dependencies of StreAM-T_g on network size

We benchmarked the full computation of $T_g(a)$ with different $k \in [3, 5]$ for increasing network sizes $|V|$. Therefore we performed a full computation with *StreAM*. *StreAM-T_g* sorts the obtained transition list, converts them into transition matrices and combines them into a global Markov model for each vertex.

Runtime evaluation

Figure 4b shows computational speeds for each dynamic graph. Speed decreases linearly with a small slope (Fig. 4a). While this is encouraging the computation of transition matrices for $k > 5$ is still prohibitively expensive due to the exponential increase of the matrix dimensions with $2^{\frac{k\cdot(k-1)}{2}}$. For G_t obtained from MD simulations, we observe fast speeds due to small batch sizes (Table 1).

Figure 4b reveals that T_{cpu} increases linearly with increasing $|V|$ and with k exponentially. We restrict the $T_g(a)$ full computation to $k < 5$. In Fig. 4c, speed decreases linearly with δ_t. As δ_t represents the changes between snapshots our observation has implications for

---

[2] https://github.com/BenjaminSchiller/Stream.

[3] http://www.cbs.tu-darmstadt.de/streAM-Tg.tar.gz.

[4] https://github.com/BenjaminSchiller/DNA.datasets

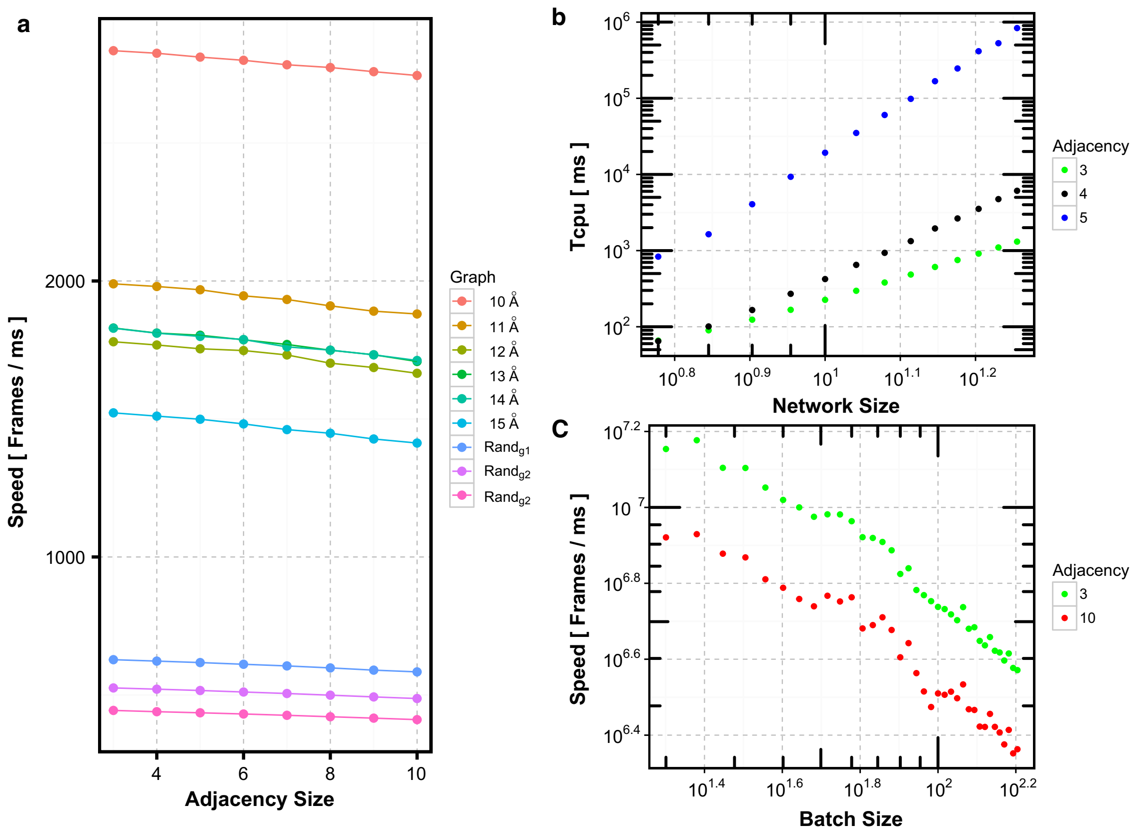


Fig. 4 Runtime performance of StreAM-T_g. **a** Speed of computing a set of $\mathcal{T}(V')$ using *StreAM*. **b** Performance of $T_g(a)$ full computation with increasing network size $|V|$ and different adjacency sizes $k = 3, 4, 5$. **c** Speed of *StreAM* with increasing batch size for $k = 3, 10$

the choice of MD integration step lengths as well as trajectory granularity.

Performance enhancing by precomputed contact probability

The exponential increase of transition matrix dimensions with $2^{\frac{k\cdot(k-1)}{2}}$ is an obvious disadvantage of the proposed method. However, there exist several $T(V')$ where every vertex is never in contact with another vertex from the set. These adjacencies remain only in one state during the whole simulation. To avoid the computation of the respective Markov models we precomputed $P(X, \vec{r}_i, \vec{r}_j)$ of all vertices. Thus only combinations are considered with $P(X, \vec{r}_i, \vec{r}_j) > 0$. This procedure leads to a large reduction of T_{cpu} due to fewer number of matrices to be computed to derive $T_g(a)$. To illustrate this reduction, we compute the number of adjacencies left after a precomputation of $P(X, \vec{r}_i, \vec{r}_j)$ as a function of d for the TC-Aptamer simulation without TC. The remaining number of transition matrices for adjacency sizes $k = 3, 4, 5$ are shown in Fig. 5b. For further illustration we show the graph of the RNA molecule obtained for a cut-off of $d = 15$ Å in Fig. 5a.

We can observe that using a precomputation of $P(X, \vec{r}_i, \vec{r}_j)$ to a full computation of $T_g(a)$ hardly depends on the Euclidean cut-off (d) for all considered adjacencies. The reduced computational costs in case of a full computation can be expressed by a significant smaller number of transition matrices left to compute for all considered adjacency sizes $k = 3, 4, 5$. For example if we use $k = 4$ and $d = 13$ Å we have to compute 16,248,960 transition matrices, if we use a precomputation of $P(X, \vec{r}_i, \vec{r}_j)$ we can reduce this value to 2,063,100, this roughly eightfold. Furthermore, in case of new contact formation due to an increased d the number of transition matrices can increase.

Robustness against threshold

Here, we investigate the influence of threshold d for the full computation of $T_g(a)$. To this end, we created dynamic graphs with different $d \in [11, 15]$ Å of the

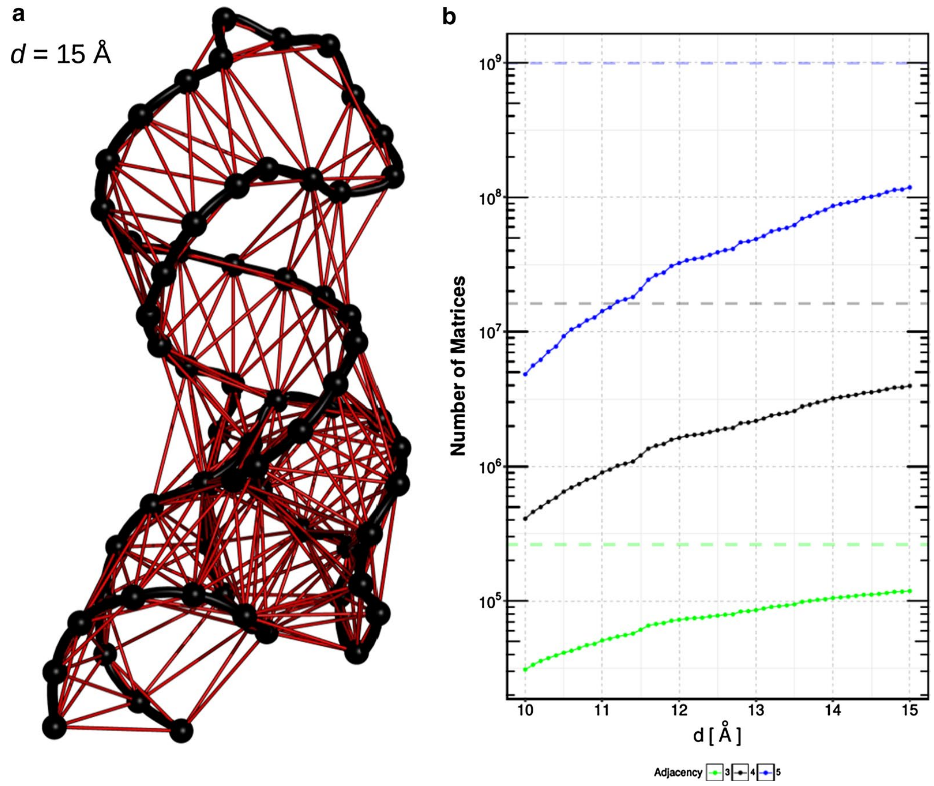


Fig. 5 Precomputation with different cut-offs. **a** Illustration of the the first frame of the TC-Aptamer simulation without TC th created with a cut-off of $d = 15$ Å. Vertices (representing nucleotides) are colored in *black* and edges (representing interactions) in *red*. The edges belonging to the backbone are furthermore highlighted in *black*. Graphics were created using `Pymol` and `R` [39, 47]. **b** Number of $\mathcal{T}(V')$ for a full computation of $T_g(a)$ after selection with contact probability as function of cut-off d for three different adjacency sizes ($k = 3, 4, 5$). The *dashed lines* show the number of matrices normally required for a full computation [$k = 3$, 262,080 matrices (*green*); $k = 4$, 16,248,960 matrices (*black*); $k = 5$, 991,186,560 matrices (*blue*)]

TC-Aptamer simulation without TC. Here, we focus on a simple model with an adjacency size of $k = 3$, thus with eight states. In particular, we focus on the local adjacency matrix of combination 52, 54 and 51 because these nucleotides are important for TC binding and stabilization of intermediates.

To access the overall robustness of a full computation of $T_g(a)$ we compute the stationary distribution for every $T_g(a)$ and afterwards we compare them with each other. For the comparison we use the Pearson product moment correlation (Pearson's r). Figure 6 illustrates the comparison of stationary distributions obtained from 65 $T_g(a)$ for unit sphere dynamic graphs with different d.

The obtained Pearson correlations r are also shown in Fig. 6 (a, upper triangle). We observed a high robustness expressed by an overall high correlation ($r = 0.938$ to $r = 0.98$) of the dynamic graphs created with different d. However transient states disappear with increasing threshold d (Fig. 6b). This observation stems from the fact that the obtained graph becomes more and more densely connected. One consequence of a high threshold d is that the adjacency remain in the same state.

Accuracy of StreAM

In this section we discuss the accuracy of *StreAM* for the computation of a set of all transitions $\mathcal{T}(V')$ on finite data samples. Our approach estimates the transition probabilities from a trajectory as frequencies of occurrences. It could be shown that uncertainties derived from a transition matrix (e.g derived from a molecular dynamics simulation) decreases with increasing simulation time [22]. Thus the error and bias in our estimator are driven by the available data set size to derive $\mathcal{T}(V')$. Additionally, there is an implicit influence of k on the accuracy since

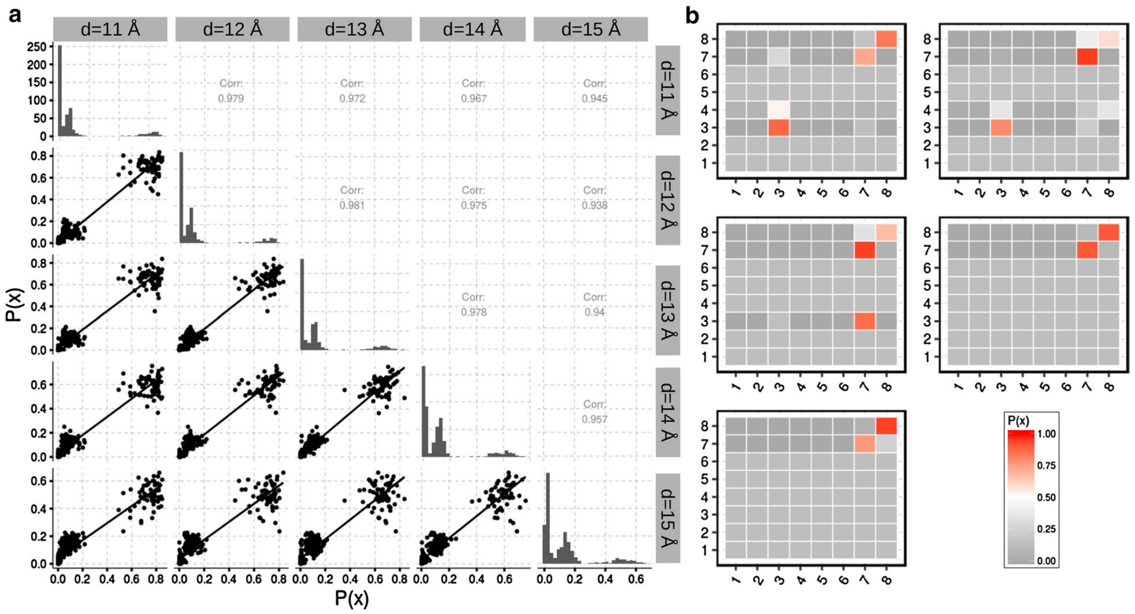


Fig. 6 Robustness for $T_g(a)$ of the native riboswitch. **a** Scatter plot matrix of computed $\vec{\pi}$ for each $T_g(a)$ at different d. The *lower triangle* includes the scatterplots obtained at different d. The diagonal includes the histogram of all 65 $\vec{\pi}$ and the *upper triangle* includes the Pearson product moment correlation of the corresonding scatterplots. **b** Illustration of single $T(V')$ derived for vertex combination 52, 54 and 51 for $d \in [11, 15]$ Å as heat maps

the number of k determines the transition matrix dimensions. Consequently, the available trajectory (system) data must be at least larger than the number of entries in the transition matrix to be estimated in order to use *StreAM*.

Application to molecular synthetic biology

This section is devoted to investigate possible changes in Markovian dynamics of the TC-Aptamer upon binding of 7-cl-tc. This particular antibiotic is part of the crystal structure of the TC-Aptamer thus structure of 7-cl-tc has the correct geometry and orientation of functional groups.

For both simulations of “Workflow”, we computed 16,248,960 transition matrices and combined them into 65 global models (one for each vertex of the riboswitch). To account for both the pair-interactions and potential stacking effects we focus on $k = 4$-vertex adjacencies and use dynamic RNA graphs with $d = 13$ Å. One global transition matrix contains all the transitions a single nucleotide participates in. The stationary distribution and the implied entropy (changes) help to understand the effects of ligand binding and potential improvements on this (the design problem at hand). The ΔH obtained are shown in Fig. 7.

A positive value of ΔH in Fig. 7 indicates a loss of conformational entropy upon ligand binding. Interestingly, the binding loop as well as complexing nucleotides gain entropy. This is due to the fact of rearrangements between the nucleotides in spatial proximity to the ligand because 70% of the accessible surface area of TC is buried within the binding pocket L3 [23]. Experiments confirmed that local rearrangement of the binding pocket are necessary to prevent a possible release of the ligand [41]. Furthermore crystallographic studies have revealed that the largest changes occur in L3 upon TC binding [23]. Furthermore, we observe the highest entropy difference for nucleotide G51. Experimental data reveals that G51 crosslinks to tetracycline when the complex is subjected to UV irradiation [42]. These findings suggest a strong interaction with TC and thus a dramatic, positive change in ΔH. Nucleotides A52 and U54 show a positive entropy difference inside L3. Interestingly, molecular probing experiments show that G51, A52, and U54 of L3 are—in the absence of the antibiotic—the most modified nucleotides [23, 34]. Clearly, they change their conformational flexibility upon ligand binding due they direct interaction with the solvent. U54 further interacts with A51,A52,A53 and A55 building the core of the riboswitch [23]. Taken

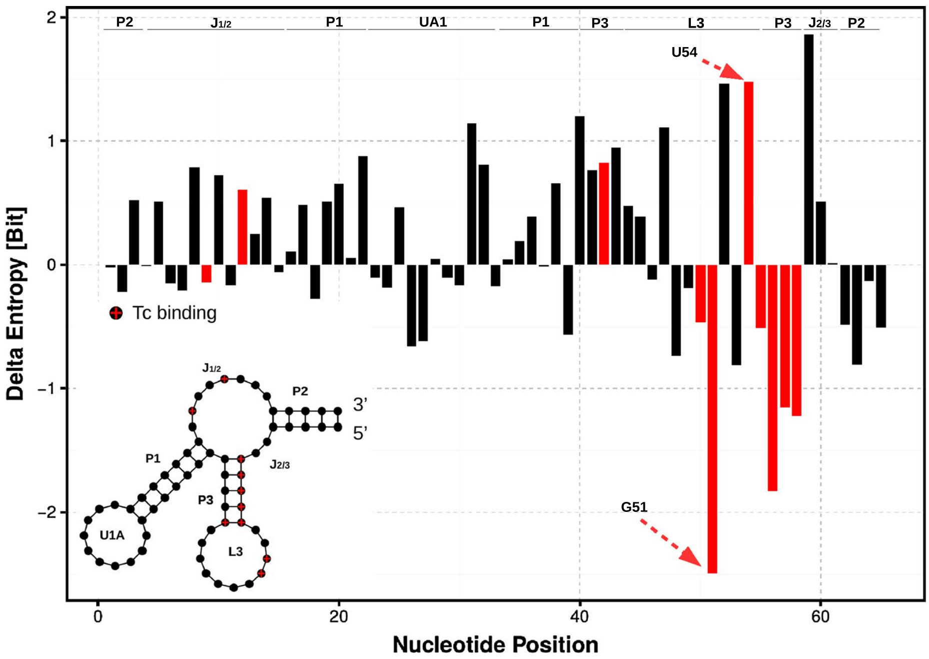


Fig. 7 ΔH (in bit) comparison for 7-cl-tc. ΔH for $T_g(a)$ of the native riboswitch and the one in complex with 7-cl-tc. Nucleotides with 7-cl-tc in complex are colored in *red*. At the *top*, we annotate the nucleotides with secondary structure information. A positive value of ΔH indicates a loss and a negative a gain of conformational entropy

together, these observations reveal that U54 is necessary for the stabilization of L3. A more flexible dynamics (ΔH) will change the configuration of the binding pocket and promotes TC release.

Comparison of tetracycline derivates

In this section, we want to investigate possible changes in configuration entropy by binding of different TC derivates. Moreover, we want to contrast *StreAM*-T_g to conventional metrics like RMSF (Eq. 5) using the entropy of the stationary distributions obtained from $T_g(a)$. Therefore, we simulated a set consisting of four different antibiotics (atc, dc, ddtc, tc) in complex with the riboswitch of "Workflow". The structures of all derivates, each with different functional groups and different chemical properties, are shown in Fig. 3. For this approach we use a precomputation of $P(X, \vec{r_i}, \vec{r_j})$ to reduce the number of transition matrices for a full computation of $T_g(a)$. Hence for all four simulations of TC derivates, we computed 1,763,208 (for tc), 1,534,488 (for atc), 2,685,816 (for dc) and 2,699,280 (for ddtc) transition matrices and combined them into 65 global models $T_g(a)$ each. Similar to "Application to molecular synthetic biology", we compute $\Delta H = H_{wt} - H_{complex}$ from the stationary distribution as well as $\Delta RMSF = RMSF_{wt} - RMSF_{complex}$ from individual RMSF computations. The results are shown in Fig. 8.

The $\Delta RMSF$ in Fig. 8b and in ΔH Fig. 8a shows a similar picture in terms of nucleotide dynamics. If we focus on atc we can observe a loss of conformational entropy upon ligand binding for almost every nucleotide. Considering this example the RMSF only detects a significant loss of nucleotide-based dynamics ranging from nucleotide 37–46. However, for dc, we observe the same effects like for dc. Contrary to this observation we detect, for ddtc, an increase in dynamic upon ligand binding as well as negative $\Delta RMSF$ values. For tc, we observe a similar picture as for 7-cl-tc ("Comparison of tetracycline derivates"). In a next step, we want to compare the obtained differences in stationary distribution with experimental values. To this end,we use an experimental metric: *xfold* values. A xfold value describes the efficiency of regulation in vivo and is given as the ratio of fluorescence without and with antibiotic in the experimental setup [43]. Unfortunately, atc reveals no experimental dynamics due to growth inhibition caused by the toxicity of the respective tc derivative [43]. In contrast to atc, dc and ddtc

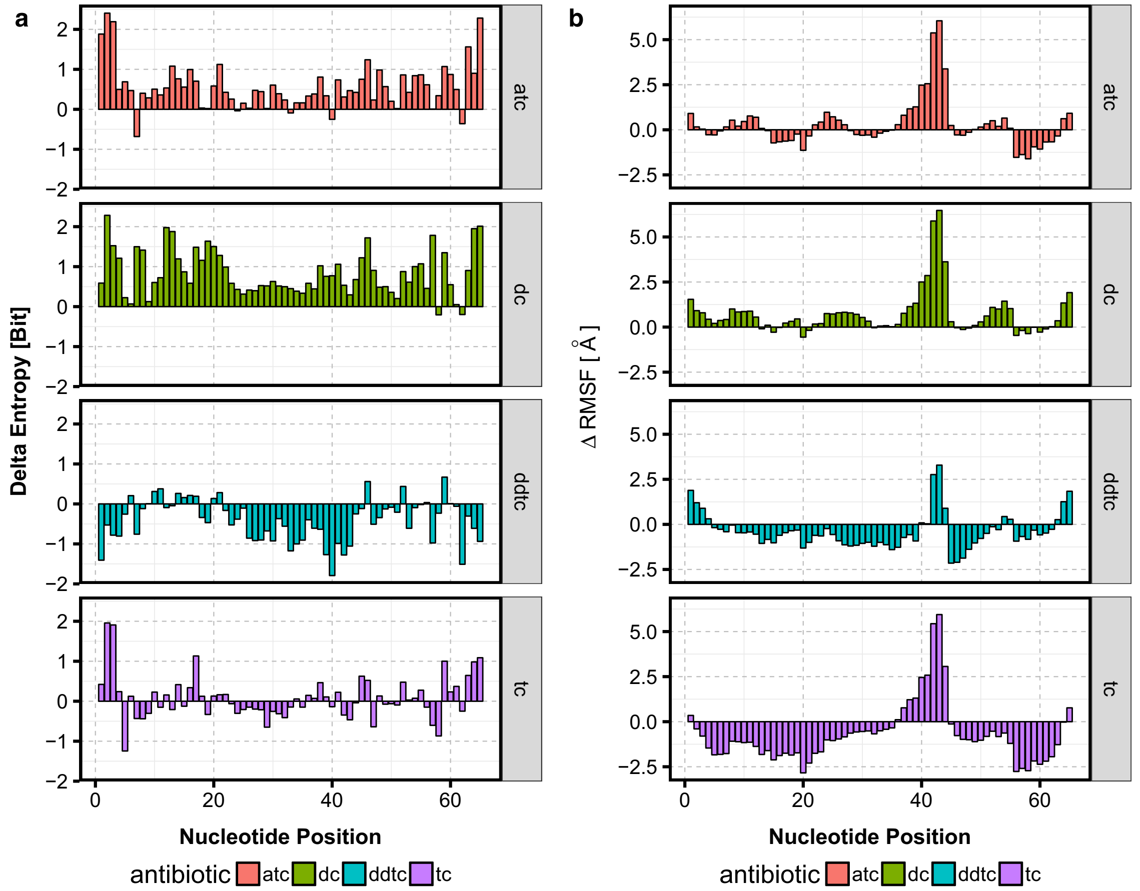


Fig. 8 Comparison of ΔH and $\Delta RMSF$. **a** ΔH for $T_g(a)$ between the native riboswitch and the complex with four different TC derivates. ΔH is plotted against nucleotide position as a bar plot. A positive value of ΔH indicates a loss and a negative a gain of conformational entropy. **b** $\Delta RMSF$ between the native riboswitch and the complex with four different TC derivates (antibiotic). A positive value of $\Delta RMSF$ indicates a loss and a negative an increase in fluctuations

show only a weak performance (xfold = 1.1) in comparison to tc (xfold = 5.8) and 7-cl-tc (xfold = 3.8) [43]. On the one hand, atc and dc appear overall too rigid and on the other hand ddtc too flexible to obtain a stable bound structure, implying insufficient riboswitch performance. For our design criterion of high xfold, we conclude that only certain nucleotides are allowed to be affected upon ligand binding. In particular, we need flexible nucleotides for the process of induced ligand binding (like nucleotide G51 Fig. 7) and stabilization of the complex intermediates ("Application to molecular synthetic biology"). Additionally, the switch needs rigidity for nucleotides building the stem region of the TC-Aptamer upon ligand binding (like nucleotides A51, A52 and A53 Fig. 7).

Summary, conclusion, and future work

Simulation tools to design and analyze functionally RNA based devices are nowadays very limited. In this study, we developed a new method *StreAM-*T_g to analyze structural transitions, based on a coarse grained representation of RNA MD simulations, in order to gain insights into RNA dynamics. We demonstrate that *StreAM-*T_g fulfills our demands for a method to extract the coarse-grained Markovian dynamics of motifs of a complex RNA molecule. Moreover *StreAM-*T_g provides valuable insights into nucleotide based RNA dynamics in comparison to conventional metrics like the RMSF.

The effects observed in a designable riboswitch can be related to known experimental facts, such as conformational altering caused by ligand binding. Hence *StreAM*-T_g derived Markov models in an abstract space of motif creation and destruction. This allows for the efficient analysis of large MD trajectories.

Thus we hope to elucidate molecular relaxation timescales, spectral analysis in relation to single-molecule studies, as well as transition path theory in the future. At present, we use it for the design of switchable synthetic RNA based circuits in living cells [2, 44].

To broaden the application areas of *StreAM*-T_g we will extend it to proteins as well as evolutionary graphs mimicking the dynamics of molecular evolution in sequence space [45].

Abbreviations

MD: molecular dynamics; RMSF: root-mean-square fluctuation; TC: tetracycline; dc: doxycycline; atc: anhydrotetracycline; ddtc: 6-deoxy-6-demythyltetracycline; 7-cl-tc: 7-chlorotetracycline.

Authors' contributions

Conceptualization of research work by SJ and KH. Data preparation by SJ, PB and MB. Implementation by SJ and BS. Benchmarks and Analysis of algorithm by SJ. Graphics by SJ, BS and PB. Writing of the manuscript by SJ, BS, KH, TS, MB and PB. Valuable suggestions to improve the manuscript by SJ, TS and KH. This Paper is an extended version of the research article: StreAM-T_g : Algorithms for Analyzing Coarse Grained RNA Dynamics Based on Markov Models of Connectivity-Graphs [46]. All authors read and approved the final manuscript.

Author details

[1] Department of Biology, TU Darmstadt, Schnittspahnstr. 2, 64283 Darmstadt, Germany. [2] Department of Computer Science, TU Dresden, Nöthnitzer Str. 46, 01187 Dresden, Germany. [3] Department of Biology, Department of Computer Science, Department of Physics, TU Darmstadt, Schnittspahnstr. 2, 64283 Darmstadt, Germany.

Acknowledgements

Not applicable.

Competing interests

The authors declares that they have no competing interests.

Availability of data and materials

Generator for dynamic graphs: https://github.com/BenjaminSchiller/DNA.datasets Implementation of *StreAM*-T_g: http://www.cbs.tu-darmstadt.de/streAM-Tg.tar.gz Implementation of *StreAM* and *StreAM*$_B$: https://github.com/BenjaminSchiller/Stream

Funding

The Authors gratefully acknowledge financial support by the LOEWE project CompuGene of the Hessen State Ministry of Higher Education, Research and the Arts. Parts of this work have also been supported by the DFG, through the Cluster of Excellence cfaed as well as the CRC HAEC.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 31 December 2016 Accepted: 16 May 2017
Published online: 30 May 2017

References

1. Schlick T. Mathematical and biological scientists assess the state of the art in Rna science at an Ima workshop, Rna in biology, bioengineering, and biotechnology. Int J Multiscale Comput Eng. 2010;8(4):369–78.
2. Carothers JM, Goler JA, Juminaga D, Keasling JD. Model-driven engineering of RNA devices to quantitatively program gene expression. Science. 2011;334(6063):1716–9.
3. Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP. Computational approaches for RNA energy parameter estimation. RNA. 2010;16(12):2304–18.
4. Laing C, Schlick T. Computational approaches to RNA structure prediction, analysis, and design. Curr Opin Struct Biol. 2011;21(3):306–18.
5. III TEC. Simulation and modeling of nucleic acid structure, dynamics and interactions. Curr Opin Struct Biol. 2004;14(3):360–7.
6. Zhang M, Perelson AS, Tung C-S. RNA Structural Motifs. eLS. 2011;1–4. doi:10.1002/9780470015902.a0003132.pub2.
7. Parisien M, Major F. The MC-fold and MC-Sym pipeline infers RNA structure from sequence data. Nature. 2008;452(7183):51–5.
8. Alder BJ, Wainwright TE. Studies in molecular dynamics. J Chem Phys. 1959;30:459–66.
9. Huang X, Yao Y, Bowman GR, Sun J, Guibas LJ, Carlsson G, Pande VS. Constructing multi-resolution markov state models (msms) to elucidate rna hairpin folding mechanisms. In: Biocomputing 2010. World Scientific. 2012. p. 228–39. doi: 10.1142/9789814295291 0025.
10. Gu C, Chang H-W, Maibaum L, Pande VS, Carlsson GE, Guibas LJ. Building Markov state models with solvent dynamics. BMC Bioinform. 2013;14(2):8. doi:10.1186/1471-2105-14-S2-S8.
11. Pinamonti G, Zhao J, Condon DE, Paul F, Noé F, Turner DH, Bussi G. Predicting the kinetics of RNA oligonucleotides using Markov state models. J Chem Theory Comput. 2017;13(2):926–34. doi:10.1021/acs.jctc.6b00982.
12. Warfield BM, Anderson PC. Molecular simulations and Markov state modeling reveal the structural diversity and dynamics of a theophylline-binding RNA aptamer in its unbound state. PLoS ONE. 2017;12:e0176229. doi:10.1371/journal.pone.0176229.
13. Bottaro S, Gil-Ley A, Bussi G. RNA folding pathways in stop motion. Nucleic Acids Res. 2016;44(12):5883–91. doi:10.1093/nar/gkw239.
14. Shapiro BA, Yingling YG, Kasprzak W, Bindewald E. Bridging the gap in RNA structure prediction. Curr Opin Struct Biol. 2007;17(2):157–65.
15. Chodera JD, Noé F. Markov state models of biomolecular conformational dynamics. Curr Opin Struct Biol. 2014;25:135–44.
16. Schiller B, Jager S, Hamacher K, Strufe T. Stream—a stream-based algorithm for counting motifs in dynamic graphs. Berlin: Springer LNCS; 2015. p. 53–67.
17. Jonikas MA, Radmer RJ, Laederach A, Das R, Pearlman S, Herschlag D, Altman RB. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. RNA. 2009;15(2):189–99.
18. Zeng X, Zhang L, Xiao X, Jiang Y, Guo Y, Yu X, Pu X, Li M. Unfolding mechanism of thrombin-binding aptamer revealed by molecular dynamics simulation and Markov State Model. Sci Rep. 2016;6:24065. doi:10.1038/srep24065.
19. Manzourolajdad A, Arnold J. Secondary structural entropy in RNA switch (Riboswitch) identification. BMC Bioinform. 2015;16(1):133.
20. Gan HH, Pasquali S, Schlick T. Exploring the repertoire of RNA secondary motifs using graph theory; implications for RNA design. Nuc Acids Res. 2003;31(11):2926–43.
21. Stombaugh J, Zirbel CL, Westhof E, Leontis NB. Frequency and isostericity of RNA base pairs. Nucleic Acids Res. 2009;37(7):2294–312.
22. Metzner P, Noé F, Schütte C. Estimating the sampling error: distribution of transition matrices and functions of transition matrices for given trajectory data. Phys Rev E Stat Nonlin Soft Matter Phys. 2009;80(2):1–33. doi:10.1103/PhysRevE.80.021106.
23. Xiao H, Edwards TE, Ferré-D'Amaré AR. Structural basis for specific, high-affinity tetracycline binding by an in vitro evolved aptamer and artificial riboswitch. Chem Biol. 2008;15(10):1125–37.
24. Wunnicke D, Strohbach D, Weigand JE, Appel B, Feresin E, Suess B, Muller S, Steinhoff HJ. Ligand-induced conformational capture of a synthetic tetracycline riboswitch revealed by pulse EPR. RNA. 2011;17(1):182–8.

25. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. Charmm: a program for macromolecular energy, minimization, and dynamics calculations. J Comp Chem. 1983;4(2):187–217.
26. Aleksandrov A, Simonson T. Molecular mechanics models for tetracycline analogs. J Comp Chem. 2009;30(2):243–55. doi:10.1002/jcc.21040.
27. Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. J Chem Phys. 2007;126(1):014101. doi:10.1063/1.2408420.
28. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR. Molecular dynamics with coupling to an external bath. J Chem Phys. 1984;81(8):3684. doi:10.1063/1.448118.
29. Nosé S. A molecular dynamics method for simulations in the canonical ensemble. Mol Phys. 1984;52(2):255–68. doi:10.1080/00268978400101201.
30. Hoover WG. Canonical dynamics: equilibrium phase-space distributions. Phys Rev A. 1985;31(3):1695–7. doi:10.1103/PhysRevA.31.1695.
31. Parrinello M, Rahman A. Polymorphic transitions in single crystals: a new molecular dynamics method. J Appl Phys. 1981;52(12):7182. doi:10.1063/1.328693.
32. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: a linear constraint solver for molecular simulations. J Comput Chem. 1997;18(12):1463–72.
33. Hanwell MD, Curtis DE, Lonie DC, Vandermeersch T, Zurek E, Hutchison GR. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. J Cheminform. 2012;4(1):17. doi:10.1186/1758-2946-4-17.
34. Hanson S, Bauer G, Fink B, Suess B. Molecular analysis of a synthetic tetracycline-binding riboswitch. RNA. 2005;11:503–11.
35. Deigan KE, Li TW, Mathews DH, Weeks KM. Accurate SHAPE-directed RNA structure determination. PNAS. 2009;106(1):97–102.
36. Hamacher K, Trylska J, McCammon JA. Dependency map of proteins in the small ribosomal subunit. PLoS Comput Biol. 2006;2(2):1–8.
37. Senne M, Trendelkamp-schroer B, Mey ASJS, Schütte C, Noe F. EMMA: a software package for Markov model building and analysis. Theory Comput J Chem. 2012;8(7):2223–38. doi:10.1021/ct300274u.
38. Spedicato GA. Markovchain: discrete time Markov chains made easy. (2015). R package version 0.4.3
39. Team RDC. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2008.
40. Buß O, Jager S, Dold SM, Zimmermann S, Hamacher K, Schmitz K, Rudat J. Statistical evaluation of HTS assays for enzymatic hydrolysis of β-Keto esters. PloS ONE. 2016;11(1):e146104.
41. Reuss A, Vogel M, Weigand J, Suess B, Wachtveitl J. Tetracycline determines the conformation of its aptamer at physiological magnesium concentrations. Biophys J. 2014;107(12):2962–71.
42. Berens C, Thain A, Schroeder R. A tetracycline-binding rna aptamer. Bioorg Med Chem. 2001;9(10):2549–56.
43. Müller M, Weigand JE, Weichenrieder O, Suess B. Thermodynamic characterization of an engineered tetracycline-binding riboswitch. Nucleic Acids Res. 2006;34(9):2607. doi:10.1093/nar/gkl347.
44. Cameron DE, Bashor CJ, Collins JJ. A brief history of synthetic biology. Nature reviews. Microbiology. 2014;12(5):381–90.
45. Lenz O, Keul F, Bremm S, Hamacher K, von Landesberger T. Visual analysis of patterns in multiple amino acid mutation graphs. In: IEEE conference on visual analytics science and technology (VAST). 2014. p. 93–102.
46. Jager S, Schiller B, Strufe T, Hamacher K. Stream-T_g : algorithms for analyzing coarse grained RNA dynamics based on markov models of connectivity-graphs. Berlin: Springer; 2016.
47. Schrödinger L. The PyMOL molecular graphics system, Version 1.8. 2015.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

6.4 StreaMD: Advanced analysis of Molecular Dynamics using R

This manuscript deals with a software package which provides an interface for the `stream` library and the *Gromacs* `XDRfile` as well as the `tng` library. Both libraries contain compression and parser routines for MD trajectories. The current manuscript is under revision:

- Dombrowsky, M.J.*, **Jager, S.***, Schiller, B., Mayer, B.E., Stammler, S., Hamacher, K. (2017) StreaMD: Advanced analysis of Molecular Dynamics using R, Journal of Computational Chemistry JCC (in revision)

With this package, the above named motif-based methods in combination with MD are made available to a wide range of users. The interface was evaluated in terms of run-time and memory requirements. As an example of an application, the 3D structure of the Tetracycline aptamer (cf. Section 6.3) was used and analyzed with the help of StreaM$_k$ as well as classical MD analysis like RMSD and RDF.

Contributions The idea as well as the conception of the work are contributions from myself. I also supervised my student Max Dombrowsky in the writing and submission process of the publishing. The evaluation process of the MD simulations was carried out by Max Dombrowsky, the creation of the dynamic graphs and the motif-based analysis was done by myself. Moreover, I helped to write and motivate the paper. Both Dombrowsky and I implemented the package. Benjamin Mayer and Sebastian Stammler gave valuable hints for the implementation and helped to write the paper. Kay Hamacher helped to motivate it and improved it in its logical structure. Benjamin Schiller implemented and helped to interface the `stream` library for motif counting in dynamic graphs. In this article I am shared first author.

StreaMD: Advanced analysis of Molecular Dynamics using R

Maximilian J. Dombrowsky, Sven Jager, Benjamin Schiller, Benjamin E. Mayer,
Sebastian Stammler, Kay Hamacher

January 9, 2018

Abstract

`Gromacs` is one of the most popular molecular simulation suites currently available. In this contribution we present `streaMD`, the first interface between `Gromacs` trajectory files and the statistical language `R`.
The amount of data created due to ever increasing computational power renders fast and efficient analysis of trajectories into a challenge. Especially as standard approaches such as root-mean square fluctuations and the like provide only limited physical insight. In our `streaMD` package integration of the `Gromacs` I/O libraries with advanced, graph-based analysis methods as the `java` library `Stream` leads to both: improved speed and analysis depth. We benchmark our results and highlight the applicability of the package by an interesting problem in RNA design, namely the interaction of tetracycline with an aptamer.

Keywords: Molecular Dynamics, Dynamic Graphs, R, Rcpp, rJava , Gromacs ■

1 Introduction

Understanding the structure and dynamics of (bio)molecular systems is a major challenge for computational chemistry and synthetic biology[1]. *Molecular dynamics* (MD) simulation is one strategy to investigate the behavior of biomolecules in their natural environment. MD is applied extensively to study proteins, nucleic acids, and their interactions in different solvents[2–4]. Insights into thermodynamics of conformational changes are especially important for the rational design of proteins/enzymes[5].

Due to the exponential increase of computational resources it is now possible to describe large biochemical systems like channel proteins in cell membranes on an atomistic scale[6,7]. Protein interaction mechanisms as well as experimental quantities like transition energies and other thermodynamic properties can be obtained[8,9]. As computational power increases even further, MD simulations will grow in size and reach longer timescales, possibly as much as a million fold greater than they are today[10]. However, the analysis of long(er) simulation time series is expensive and time consuming, especially if performed at high resolutions. Thus, the development of efficient methods to analyze extensive MD simulations becomes more important.

Beyond the computational and storage demands, one question arises: Whether basic analysis of spatial 3D coordinates of particles is the best semantic level to describe dynamics of complex biomolecules or is a new paradigm to analyze trajectories more promising?

The derivation of (time-dependent) graph properties constitutes a possible solution. To this end, biomolecular configurations and their dynamics can be represented as dynamic graphs and evaluated using a motif-based analysis. Motifs, the basic building blocks of most networks, help to understand their complexity and high-order organization[11,12]. These provide insight into the relations between the components on a more abstract semantic level than spatial coordinates alone.

In previous work we were able to show that graphs can be formulated based on spatial proximity of constituents, e.g., amino acids. It turned out that four-vertex motif counts can be used to annotate secondary structures and determine essential dynamics for protein simulations[13]. Furthermore, the approach is applicable to other biomolecular systems as well, e.g., nucleic acid based graphs in RNAs[14,15]. In this context we devolped `StreaM`$_k$, an efficent algorithm that counts motifs as well as `StreAM` which creates motif based Markov state models from dynamic graphs modelled from MD simulations.

At the same time, several MD simulation suites are available, where `Gromacs`[16,17] is one of the most commonly used. It provides high performance and support for several specialized MD concepts[2]. Until now, `Gromacs` I/O integration in `R`[18] is hindered by severe technical difficulties as most libraries either are more specialized on sequence analysis (e.g., `BioPhysConnector`[19]), characterization of topological knots regarding static molecular structures (`Rkonts`[20]), visualization of `Gromacs` output files (`MDplot`[21]) or focus on not compatible file formats (e.g., `bio3d`[22] can only parse `CHARMM`'s dcd files[23]).

Nevertheless, the `R` eco system provides a broad range of diverse packages, methods (e.g., machine learning, network analysis, sequence annotation, ...), and I/O capabilities to augment MD analysis (e.g., fasta, pdb, fcs, xlsx, ...) rendering the use of `R` highly desirable. At the moment, such advanced analysis and parsing programs are only available for `Python` (e.g., `MDtraj`[24]) or `Julia`[25,26]. Thus, our package `streaMD`, aims to provide all functions to directly parse `Gromacs` output files into the statistical language `R`, perform graph based analysis, and employ popular analytical methods like root mean square deviation (RMSD) and fluctuation (RMSF), as well as radial distribution function (RDF) calculations. To ensure high performance we implemented these functions in `C++` and `Java`. Additionally, we provide functions to interface our methods and formats with already established `R` libraries like `bio3d`[22] and `BioPhysConnectoR`[19].

2 Computational Methods

2.1 Implementation Details

In order to parse `Gromacs`[16,17] trajectory files we used the `libxdrfile`[27] and `libtngfile`[28] APIs. `Rcpp`[29] has been employed to make `C` and `C++` functionalities accessible to `R`[18] while `RcppArmadillo`[30] provides the matrix data type for trajectory storage: trajectory files are stored as lists of $N \times 3$ matrices where each list element represents one time step and each matrix row contains (x, y, z) coordinates of one particle $n \in [1, \dots, N]$ with N as the number of particles. This approach enables straightforward parallelization using the `parallel`[18] package for `R`.

Gromacs gro structure files are read and written via `C` functions, enabling fast and format specific I/O. Content of gro files is represented in `R` as a list of one data frame containing all atom informations, and one matrix containing only the spatial coordinates.

RMSD, RMSF, and RDFs (see Secs. 2.6, 2.7, and 2.8, respectively) are implemented in `C++` using `RcppArmadillo`[30]. While `fast_rmsd` and `fast_rmsf` return a `R` vector, `rdf` returns a `RcppArmadillo` matrix containing the calculated histogram. `streaMD` object manipulation functions like `trjselect` (fig. 1) are `R` native and employ vector based functionalities in the spirit of `lapply`[18].

The package includes the implementation of the `streAM` algorithm for the construction of Markov state models[14] as well as generic graph based analytics, e.g., the computation of degree distributions. Additionally, `streaMD` provides a random generator for dynamic graphs[13]. This generator is usefull to explore `streaMD` functionalities by generating example graphs.

Fig. 1 gives an overview of `streaMD`'s architecture. After input generation by `Gromacs`, trajectories can be parsed into `R` using `loadxtc/loadtng`. This creates a `streaMD` trajectory object. Subsequently, one can employ several trajectory manipulation functions like `trjselect` or `CoC` (see below). Classic MD evaluation methods like RMSD or RDF calculations are available for the `streaMD` trajectory format. All these functions return `R base` objects which can be analyzed further within the `R` eco system. Complementing these classic evaluation methods, time dependent graph analysis is available via an `streaMD` object. Finally, the lightweighted `stream` trajectory format contains a reduced representation of changes in graph composition for each time step and can be stored on the user's hard drive. It can be created from xtc files or `streaMD` style trajectories via the `xtc_to_stream` function.

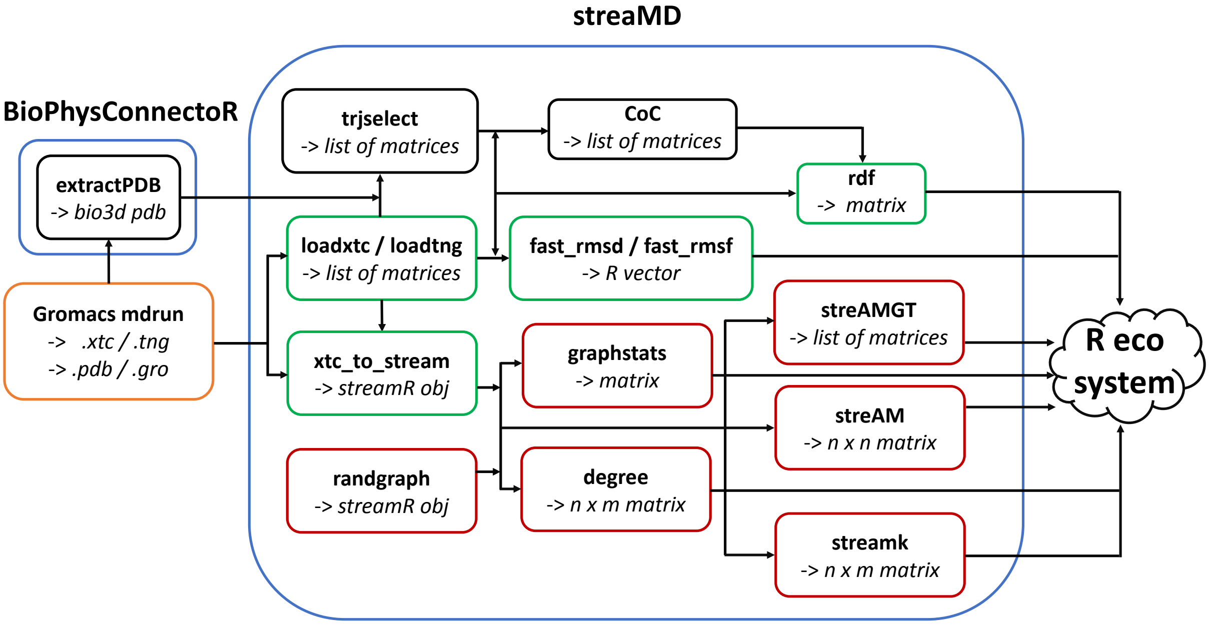


Figure 1: Overview of `streaMD` components. Blue boxes describe `R` packages. Command names are written in bold font and the corresponding output format is written in italic. We mix several languages (green: `C++`; red: `java`; black: pure `R`).

2.2 Datasets

2.2.1 Benchmarks

All benchmarks were performed on a workstation containing two *Intel Xeon CPUs X5482*, resulting in 8 physical cores with up to 3.2 GHz running an Arch Linux operating system. We employed `R` version 3.3.2, `Rcpp` version 0.12.8 and `RcppArmadillo` version 0.7.500.0.0.

In order to evaluate the performance of our xtc file interface it was compared to several other packages. Since no other xtc interface for `R`[18] is available to the best of our knowledge, we evaluated against the `loadxtc()` function for `MatLab`[31] provided within the `gro2mat`[32] package. In contrast to our high level API approach the `gro2mat` interface strategy is based on `VMD`'s[33] xtc parser. In here, only the most basic `libxdrfile` subroutines are used while tasks like memory allocation and object creation are explicitly programmed.

Additionally, the `read_gmx()` function published within the `MolecularDynamics`[25] package for `julia`[26] was used in our benchmark. The `MolecularDynamics` approach is quite similar to ours in that it is based on the `libxdrfile` high level API.

One up to 16,000 frames of a `Gromacs`[16,17] written xtc trajectory containing 6427 atoms were each loaded 1,000 times while the runtime was measured using either built-in methods like `tic()` (`MatLab`, `julia`) or the `microbenchmark` package[34] for `R`.

To compare writing time and compression between `streaMD`'s `writextc()` function and the `R base` function `save()` we prepared square matrices ranging from 1×1 up to 50000×50000 entries. These were filled with uniform distributed random `64 bit floats` between 0 and 1.

The computational time was measured using `microbenchmark`[34]. The analysis was repeated 1,000 times. Additionally the used diskspace was determined using `bash`'s base function `du`. Relative speed-up $\mathcal{S}$ is defined as $\mathcal{S} := {}^{t^{\text{o}}_{\text{CPU}}}/_{t^{\text{s}}_{\text{CPU}}}$ in which $t^{\text{o}}_{\text{CPU}}$ represents the computational time of the already available function and $t^{\text{s}}_{\text{CPU}}$ the computational time of the `streaMD` function. Relative memory saving $\mathcal{MS}$ is defined as $\mathcal{MS} := {}^{\mathcal{D}^{\text{o}}}/_{\mathcal{D}^{\text{s}}}$ accordingly.

2.2.2 Usecase

We analyzed two tetracycline aptamer MD simulations, containing 160,000 frames of 1,000 ns simulation time. While one simulation was performed in presence of 7-chlorotetracycline (wtc) the other simulation was in absence of it (wotc). The analysis was conducted using `streaMD` and `bio3d 2.3-1`[22]. Graphics were plotted using `ggplot2 2.2.0`[35] and molecular representations were visualized using `PyMol 1.8.4.0`[36]. `PDB` files were parsed into `R` using `extractPDB` distributed with `BioPhysConnectoR`[19].

2.2.3 MD Simulation Setup

All simulations were executed in Gromacs using the CHARMM27 force field with parameters for the synthetic tetracycline derivative 7-chlorotetracycline[37–39] from Aleksandrov and Simonson. Water was represented via the TIP3P explicit water model[40]. Both simulations were performed at constant temperature (300 K) for a time frame of 1,000 ns of explicit all-atom MD simulations. The systems were initially energy minimized and equilibrated for 1 ns simulation time in the NVT-ensemble at a temperature of 300 K and for 10 ns in the NpT-ensemble at a pressure of 1 bar. During the equilibration, temperature coupling was achieved through the Berendsen thermostat[41] while pressure was contained via the Berendsen barostat[41]. During the 1,000 ns production, temperature was controlled using the velocity-rescale thermostat[42] while pressure was preserved via the Parrinello-Rahman barostat[43]. During all simulations Lennard-Jones nonbonded interactions were evaluated with a cutoff distance of 1.2 nm and integration step-size was set to 1.5 fs. Both MD simulation production runs resulted in 160,000 snapshots.

2.3 Motifs, Graphs and Dynamic Graphs

A *graph* $G = (V, E)$ is an ordered pair, consisting of a set of *vertices* $V = \{v_1, v_2, \dots v_{|V|}\}$ and a set set of *edges* E. In this work, we only consider *undirected graphs without self-loops*, i.e., $E \subseteq \{\{v, w\} : v, w \in V, v \neq w\}$. A graph is called *connected* in case any two vertices $v, w \in V$ are connected. Accordingly ,a *dynamic graph* $G_t(V_t, E_t)$ can be defined as a list of graphs (e.g., $G_{t0}, G_{t1}, G_{t2}, \dots$). Two graphs $G = (V, E)$ and $G' = (V', E')$ are called *isomorphic* if they contain the same number of vertices, i.e., $|V| = |V'|$, and there exists a so-called edge-preserving bijection $f : V \to V'$ such that $\{v, w\} \in V \iff \{f(v), f(w)\} \in V'$.
A motif m is a subset of a graph G with a defined numbers of vertices $|V|$ and edges $|E|$. Motifs are sorted in classes $\mathcal{M}$, containing all motifs m of same the same $|V|$.

2.3.1 Network Transformation

We generated a dynamic graph of the 160,000 simulation snapshots (F_t) with 65 vertices $|V|$, each modeling a nucleic *3C'*. Such type of resolution has proven to be an efficient approach for representing small secondary structure as well as large quarteric RNA complexes[14,44,45]. We created two graphs with undirected edges E between two vertices V in case their spatial distance cut-off (d) is shorter than $d \in [1.1, \ldots, 1.5]\ nm$. Statistics of the obtained graphs are show in Tab. 1. $|V|$ and $|E|$ were defined to be the amount of vertices or edges respectively. Here, *add state* refers to the average number of modification operations that occur in the dynamic graph between one frame and the next. Accordingly, *total removals* depicts the sum of all edge removals and *total additions* denotes for the sum of all edge additions in the dynamic graph.

Table 1: `graphstat` statistics of obtained graphs.

Type	d	$\|V\|$	$\|E\|$	$F_t \cdot 1000$	total additions	total removals	add state	remove state
+ligand	1.1	65	118	160	1 246 057	1 245 929	7	7
+ligand	1.2	65	189	160	1 557 759	1 557 581	9	9
+ligand	1.3	65	222	160	1 440 117	1 439 893	9	8
+ligand	1.4	65	282	160	1 569 254	1 568 974	9	9
+ligand	1.5	65	332	160	1 907 673	1 907 345	11	11
−ligand	1.1	65	116	160	1 213 819	1 213 695	7	7
−ligand	1.2	65	186	160	1 458 533	1 458 350	9	9
−ligand	1.3	65	224	160	1 371 554	1 371 318	8	8
−ligand	1.4	65	279	160	1 456 476	1 456 202	9	9
−ligand	1.5	65	327	160	1 869 136	1 868 805	11	11

2.4 xtc and tng Compression Strategy

The xtc file format requires several processing steps: Initially, all floating point number coordinates are mapped to an integer after multiplication with a precision factor. This reduces the amount of bits necessary to store these numbers. Instead of writing each x, y and z value, only the differences dx, dy and dz are written if these are small enough, furthermore reducing the number of required bits. Subsequently, the differences dx, dy and dz are combined into one integer by calculating $(x \cdot y_{max} + y) \cdot z_{max} + z$ requiring less bits then the three separated integers[27].

To cope with series of small differences, the number of coordinates in such a sequence is stored, removing the need for storing every and each coordinate. If several sequences of equal length are following each other, this will be indicated by one bit. This is especially useful for water molecules which are represented by three coordinates with only small differences[27].

2.5 Stream Library

The `Stream` Library is a dynamic graph analysis library by Schiller `et al.`[13] This library includes several algorithms for dynamic graph based analysis. Furthermore, it provides a transferable analytic method with a light weight format (stream) which is perfectly suited to work with large MD trajectories. One of its algorithms is $\texttt{StreaM}_k$, an extention to count the occurrences of k-vertex motifs in dynamic graphs. Furthermore, `streAM` and $\texttt{streAM-}T_g$ are included for the construction of motif based Markov models from dynamic graphs[14][15].

2.6 RMSD

The Root Mean Square Deviation (RMSD) describes the average displacement of a structure $\mathcal{A} = (x^N_\mathcal{A}, y^N_\mathcal{A}, z^N_\mathcal{A})$ with regard to a reference structure $\mathcal{R} = (x^N_\mathcal{R}, y^N_\mathcal{R}, z^N_\mathcal{R})$. Both structures contain N particles with x, y and z constituting their spatial coordinates. It is defined as the sum over the Euclidean distances of all particles N as stated in Eq. 1.

$$RMSD(\mathcal{A}, \mathcal{R}) := \sqrt{\frac{1}{N}\sum_{n=1}^{N} (x^n_\mathcal{A} - x^n_\mathcal{R})^2 + (y^n_\mathcal{A} - y^n_\mathcal{R})^2 + (z^n_\mathcal{A} - z^n_\mathcal{R})^2} \tag{1}$$

2.7 RMSF

The Root Mean Square Fluctuation (RMSF) describes the time averaged displacement of a selected particle n at time $\tau \in [1, \ldots, \mathrm{T}]$ with respect to its time mean value $(\langle x_n\rangle, \langle y_n\rangle, \langle z_n\rangle)$:

$$RMSF(n) := \sqrt{\frac{1}{\mathrm{T}}\sum_{\tau=1}^{\mathrm{T}} (x^\tau_n - \langle x_n\rangle)^2 + (y^\tau_n - \langle y_n\rangle)^2 + (z^\tau_n - \langle z_n\rangle)^2} \tag{2}$$

2.8 RDF

The Radial Distribution Function (RDF) $g_{\mathcal{AB}}$ is a measure of density of particles $\mathcal{B}$ in the distance interval δ around particles $\mathcal{A}$. Usually $g_{\mathcal{AB}}$ is defined as weighted histogram H over the set of distances $\mathcal{D}$ with $\mathcal{D} := \{|a - b| : a \in \mathcal{A}, b \in \mathcal{B}\}$. One histogram value of a δ wide bin around c is defined as $h_\delta(c) := |\{d \in \mathcal{D} : (c - {}^{\delta}\!/_{2}) \leq d < (c + {}^{\delta}\!/_{2})\}|$ and represents the particle distribution of $\mathcal{B}$ around $\mathcal{A}$. After normalization to the shell volume $\Delta V = {}^{4}\!/_{3} \cdot \pi\, [(c + {}^{\delta}\!/_{2})^3 - (c - {}^{\delta}\!/_{2})^3]$, the overall density ρ and the number of particles $|\mathcal{A}|$, the radial distribution $g_{\mathcal{AB}}(c)$ at distance c can be retrieved.

$$g_{\mathcal{AB}}(c) = \frac{h_\delta(c)}{(4\delta c^2 + {}^{1}\!/_{3}\, \delta^3)\ \pi\ \rho\ |\mathcal{A}|} \tag{3}$$

2.9 Entropy derived from Motif Counts

Later on, our goal will be the assessment of ligand binding from our use case. To this end, we compute the motif-based entropy (changes). However, to measure changes in motif based conformational entropy we defined the frequency $f(m|_\mathcal{M}) = C(m|_\mathcal{M})/T$ as the count C of a given motif m and its corresponding motif class $\mathcal{M}$ divided by the simulation Time T. For illustration we use five vertex motifs $V = 5$. Using Shannon's definition of entropy H_S

in Eq. 4 with the number of motifs M, we determined the change of entropy upon ligand binding to be $\Delta H_S(m|_{\mathcal{M}}) := H_{S,\ -ligand}(m|_{\mathcal{M}}) - H_{S,\ +ligand}(m|_{\mathcal{M}})$.

$$H_S(m|_{\mathcal{M}}) := -\sum_{m}^{M} f(m|_{\mathcal{M}}) \cdot \log f(m|_{\mathcal{M}}) \tag{4}$$

3 Results & Discussion

3.1 Benchmarks

Figs. 2 A & B show the parsing time with regard to xtc size. Parsing time of `streaMD`'s `loadxtc()` is in some cases higher for xtc sizes under 100 frames than `gro2mat`'s `parseXtc()` but performs better for xtc sizes over 1,000 frames in comparison to `MolecularDynamics`'s `read_gmx()`. Thus our implementation is favorable in comparison to `gro2mat`'s. Since our implementation is similar to the `read_gmx()` function we can ascribe our speedup to the `Rcpp` interface and `RcppArmadillo`'s fast matrix class.

Figure 2 C shows writing time with regard to matrix entries. Writing time of `writextc()` is 31 times higher than for `R`'s `save()` if more than 1,000 matrix entries are written. Memory usage of written xtc files is up to 4 times smaller than for Rdata files that range nearer to their corresponding RAM usage(fig.2 D).

In summary, our implementation of the `libxdrfile` API is advantageous towards all compared algorithms. We provide easy to implement functions, enabling the user to save `R` objects significantly faster while requiring 4 times less hard drive space. Additionally it is now possible to write `Gromacs` trajectory files directly from `R` which allows straightforward modification of xtc files. As our `libtngfile` implementation is the only in an interpreted language we were not able to state conclusive comparisons.

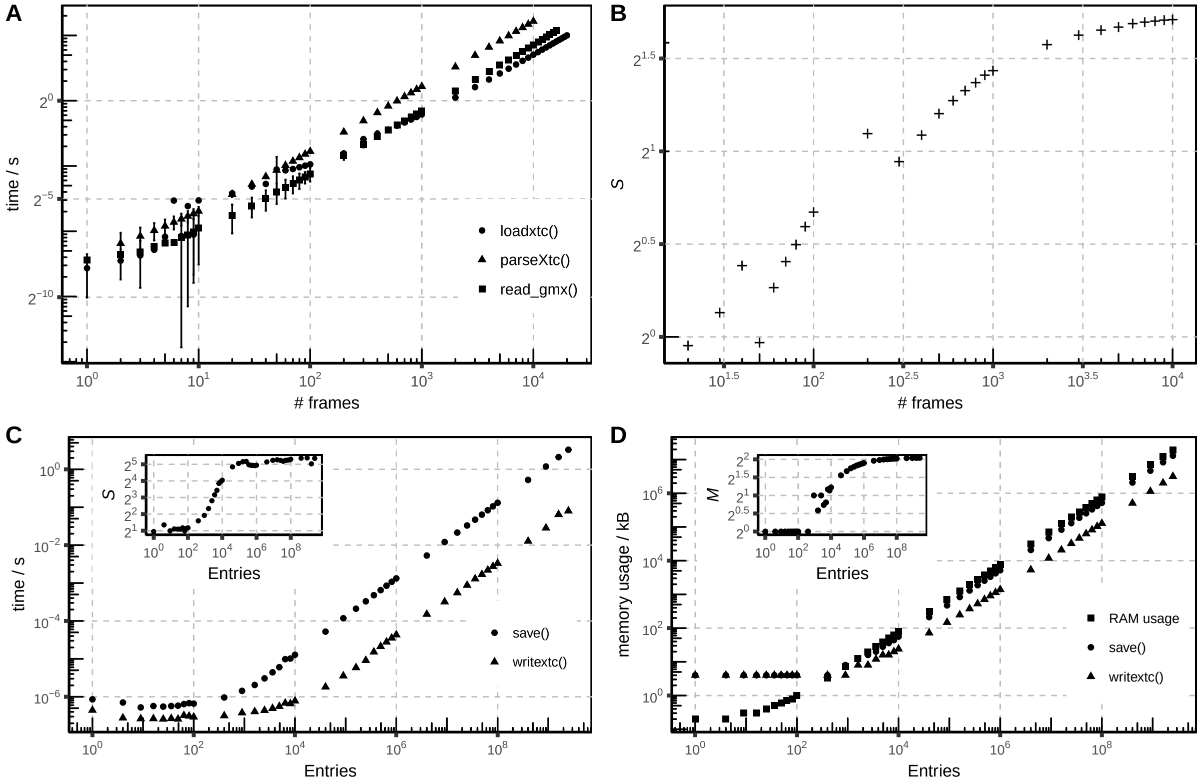


Figure 2: Performance comparison between `streaMD` xtc functionalities and `R base` or other xtc tools. A) Loading time of up to 16,000 frames xtc trajectory containing 6,427 atoms. B) Relative speedup of `loadxtc()` vs. `parseXtc()`. C) Writing time of matrices with up to $50,000 \times 50,000$ random values. C inlet) Relative speedup of `writextc()` vs. `R base`'s `save()`. D) Memory usage of written random matrices with up to $50,000 \times 50,000$ entries. D inlet) Relative memory saving of `writextc()` vs. `R base`'s `save()`.

3.2 Usecase

3.2.1 Conventional Analysis

To demonstrate the efficiency of `streaMD` functions in non-graph based analysis we computed the RMSD (eq. 1), the particle distribution (eq. 3) and the RMSF (eq. 2) of two molecular dynamic simulations both containing only the tetracycline binding aptamer (wotc) and one additionally containing tetracycline (wtc) in its binding pocket.

In order to compute RMSD and RMSF we applied a coarse-graining approach where each nucleotide is represented by its C3' atom[14]. To achieve this we first loaded the xtc-trajectory file and its corresponding pdb file. We identified all C3' atoms using `bio3d`[22] and adjusted our trajectory file accordingly. The resulting model is now storable as a xtc-file for later usage.

```
pdb <- extractPDB("structure.pdb")
trajectory <- loadxtc("trajectory.xtc", 1, 160000)
select <- atom.select(pdb, elety = "C3'")$atoms
cg.model <- trjselect( trajectory, select)
writextc("my_coarse-grained_trajectoy.xtc", cg.model)
```

Following the coarse-graining, we calculated the RMSD with regard to the first snapshot. The trajectory organization as matrices in a large list enables us to use the `R-base lapply()` method which can also be parallelized easily, using the `parallel`[18] package (Supplementary Material 2). The resulting RMSD is shown in figure 3 A.

```
RMSD <- lapply(cg.model, fast_rmsd)
RMSD <- unlist(RMSD)
```

Similar to the `fast_rmsd()` function, the radial distribution function can be computed using `lapply()` from the `parallel` package. To study the distance distribution between tetracycline and its aptamer we calculated the center of coordinates of all tetracycline atoms using `CoC` and calculated the radial particle distribution via `rdf()`. Subsequently, the resulting list of histograms was combined using `post_rdf()` that merges the histograms of `rdf()` into one.

```
select.tetra <- atom.select(pdb, resno = 67)$atoms
cg.rdf <- CoC(cg.model,select.tetra)
RDF <- lapply(cg.rdf,rdf,
              n_bins = 200,
              Box_size = c(,,),
              atoms = c(1),
              atoms_compare = c(2),
              pbc = TRUE,
              absolute = FALSE)
rdf.hist<-post_rdf(RDF,FALSE)
```

Finally, the RMSF was computed via `bio3d`. For simple transformations between `streaMD` and `bio3d` format the `streaMD_to_bio3d` procedure is available. The resulting fluctuations are shown in figure 3 B.

```
trj.mat <- streaMD_to_bio3d(cg.model)
RMSF <- rmsf(trj.mat)
```

Figure 4 displays the absolute particle distribution of the tetracycline RNA complex. This representation highlights that the complex inherits three major states in the analyzed MD simulation. The first and most frequented state at a distance around 0.65 nm represents the bound state. The second state inherits a distance around 1.46 nm representing a semi bound state in which tetracycline is partially bound to the RNA but its binding pocket is not fully formed. The last state at a distance between 4.9 and 6.0 nm represents the unbound tetracycline. This distribution reveals a large under sampling in the simulation, as there is no obvious reason why the free ligand remains around 5.4 nm. While no direct claims regarding the thermodynamic properties of the system can be made, the sampling efficiency of the simulation is evaluable, revealing e.g., distribution biases.

Assuming a completely converged system, it is possible to calculate the binding free energy

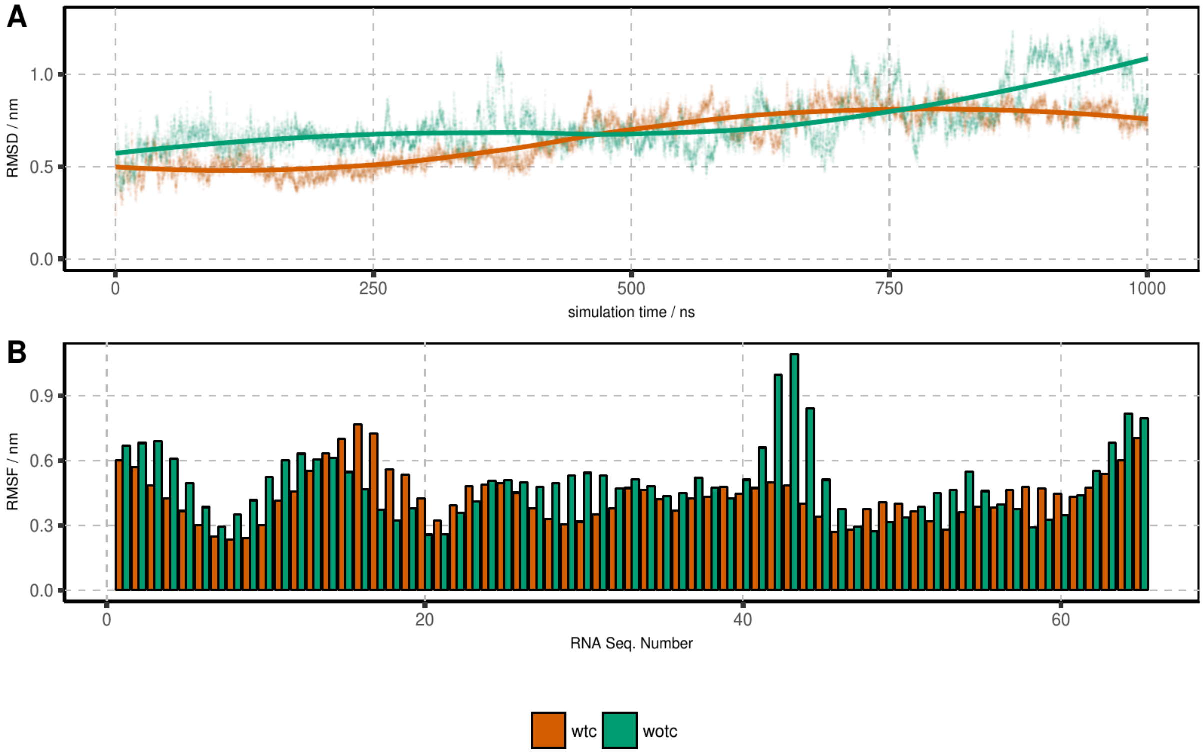


Figure 3: A) RMSD in reference to the first coordinate setting of C3' atoms of the complex (red) and single (green) simulation. B) RMSF of C3' atoms of the complex (red) and single (green) simulation.

ΔG via Boltzmann inversion[46], allowing to link simulations to laboratory experiments and therefore evaluate their validity. It is however necessary to stress, that multiple binding events should occur in one's simulation and several possible binding paths have to be sampled to allow a meaningful energy computation.

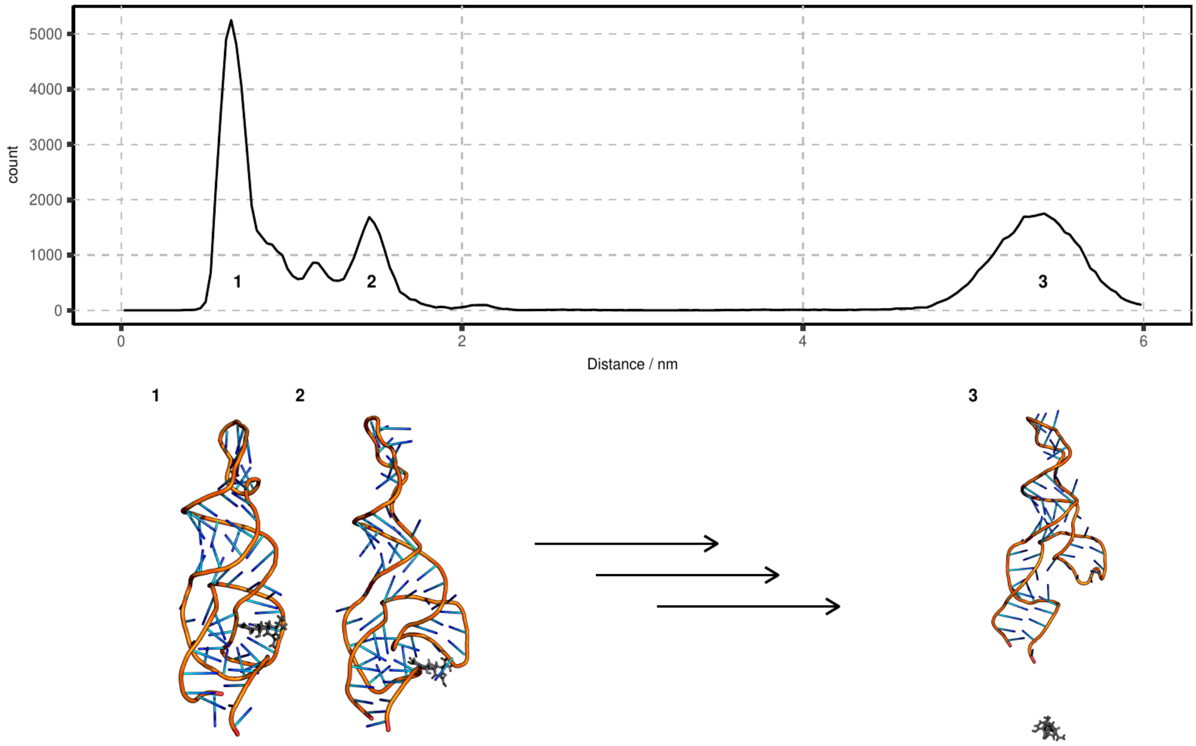


Figure 4: Particle distances between the center of coordinates of tetracycline and the center of coordinates of the binding pocket in the complex simulation added up over all coordinate settings. 1 - 3) Exemplary conformation at distance 0.65 nm (1), 1.46 nm (2) and 5.3 nm (3).

3.2.2 Graph-based Analysis

Here, we introduce the-graph based analysis approach with a generic example: The computation of backbone dynamics using 5-vertex motif counts.

It is well known that nucleotide-based molecular interactions take place between more than one partner[47]. For this reason, interactions exist for several edges observable in the dynamic graph (obtained via an Euclidean distance cut-off d) of $C3'$ coordinates at a given time-step. The resulting edges represent strong local interactions such as Watson-Crick pairing, Hoogsteen base pairing, or $\pi - \pi$-stacking of the respective RNA[14]. Thus, motifs consisting of nucleotides as vertices, might be a better semantic to describe RNA structure and dynamic rather than simple distance metrics.

Using `streaMD`, MD simulations are easily convertible to dynamic graphs. This takes advantage of a smart analytical approach, transferring your simulation into a so-called "*motif space*" and describing their respective time-dependent configuration. Initialization with a random graph, or converting a xtc trajectory into stream file format are both possible.

The next example shows the creation of a random dynamic graph where both, the initial layout and edge exchange are randomized (used for the frame computation). First, we convert xtc trajectories of both simulations to a stream file. The dynamic graph contains

1,500 edges, 1,000 vertices and 10 frames. These graphs can easily be used for testing or the creation of toy models. Additionally, `streaMD` provides a statistics module that analyzes generated stream files. The `graphstatistics()` function returns characteristics from the generated dynamic graph.

```
# creating a random graph
set.seed(1)
opath1 <- tempfile()

random <- randgraph(opath=opath1,
                    layout = "Random",
                    len=10,
                    batchType = "RandomEdgeExchange",
                    vertices = 1000,
                    edges=1500)

graphstats(opath1,"StatsOnly")

#[,1]              [,2]
#[1,] "nodes"              "1000"
#[2,] "initial edges"      "1500"
#[3,] "states"             "11"
#[4,] "first timestamp"    "0"
#[5,] "last timestamp"     "10"
#[6,] "total additions"    "1700"
#[7,] "total removals"     "200"
#[8,] "add per state"      "154.55"
#[9,] "remove per state"   "18.18"
```

One key feature of the package is the conversion of `gromacs` xtc trajectories to the stream dynamic graph file format. This particular function can convert an xtc trajectory into an unit-sphere model with a given cut-off radius d. To speed up computation `xtc_to_stream()` conversion uses the `parallel` package in `R`. The dynamic graphs are afterwards stored in a stream file. In the next example, we create unit-sphere graphs from a riboswitch trajectory. The xtc file used here contains only the $C3'$ atom coordinates of the riboswitch. Additionally, we have the possibility to take other coarse-graining schemes, e.g., reduced tree representation (structure element as a vertex)[48], use of SPQR - SPlit and conQueR (one nucleotide as two vertices)[49] or even the use of entire structural segments as vertex[50].

```

xtc.file <- "xtc.trajectory"
stream.ofile <- tempfile()

for (i in seq(1.1,1.5,0.1)){
        stream.ofile <- tempfile()
        xtc_to_stream(path=xtc.file,
                      opath = stream.ofile,
                      nframes = 160000,
                      cores = 4,
                      cutoff = i)
        all.wo <- c(all.wo,stream.ofile)
}
```

After a successful conversion, calculating the degree distribution for every vertex returns the connecting distribution for every vertex. The result is a matrix where every row stands for a timestep and the columns denote the respective degree. If the stream files are combined in a vector or list you can easily compute several functions via `lapply()`.

```
# getting degree
wo_degree <- lapply(X = all.wo,FUN=degree)
w_degree  <- lapply(X= all.tc,FUN=degree)

# simple lapply exec
wo_k <- lapply(X = all_wo,FUN=streamk,mframes=160000,motif="5")
w_k  <- lapply(X= all_tc,FUN=streamk, mframes=160000,motif="5")
```

Now, we start the motif-based analysis, using the `streamk()` algorithm[13] of the dynamic graph modeled from both MD trajectories (wotc, wtc). For this, we applied the unit-sphere model with a distance cut-off $d = 1.3$ nm. This cut-off is appropriate to measure conformational dynamics in coarse-grained models of RNA[14]. The `streamk()` algorithm efficiently counts motifs in a given dynamic graph.

$\mathcal{M}_5(9)$ is an excellent motif for this kind of analysis. In this motif all nodes are connected to the next neighbor with one edge. This type of cross-linking can also be found in the backbone of RNA. In addition, there are two cross-linked connections, each of which is connected to the second closest neighbor of a node. Due to this topology, we expect $\mathcal{M}_5(9)$ to be found increasingly in a weak cross-linked structure.

If one derives the motif counts for the MD trajectory as a function of time, its fluctuations with a motif-based entropy are obtained (eq. 4). These are interpreted as a configurational dynamic inside the motif space.

Figure 5 A, displays the degree distribution of different dynamic graphs obtained with increasing d. It can be observed, that the distributions start to shift right. This is caused by a higher threshold d leading to more undirected edges per vertex. Figure 5 B, shows the motif counts for both dynamic graphs obtained from MD simulations. In our case, tetracycline triggers upon binding a slight change in configuration of the RNA. Here it can be observed that both simulations are clearly different. More Motif $\mathcal{M}_5(9)$ is counted in the graph without, rather than with tetracycline. Since this motif, due to its topology, is found increasingly in weakly cross-linked structures, it indicates that the presence of tetracycline maximizes intramolecular interactions.

Moreover, we observe differences in motif dynamic obtained from their counts over time expressed by a motif frequency. Employing a motif-based entropy computation, it is also possible to extract certain dynamics of the atomic ensemble. Here we have used Shannon entropy H_S in detail to investigate the distribution of motif count frequencies. In Fig. 5 C we use the motif-based fluctuations from the time series of Figure 5 B. In case of our example, we determined the fluctuations of different motif counts. In this case, the differences in Shannon entropy ΔH_S are nearly all positive. Thus, our motif-based approach detects fluctuations rather than simple metrics like the RMSF. Additionally, these results suggest that the increase of motif fluctuations is proportional with structural density and complexity.

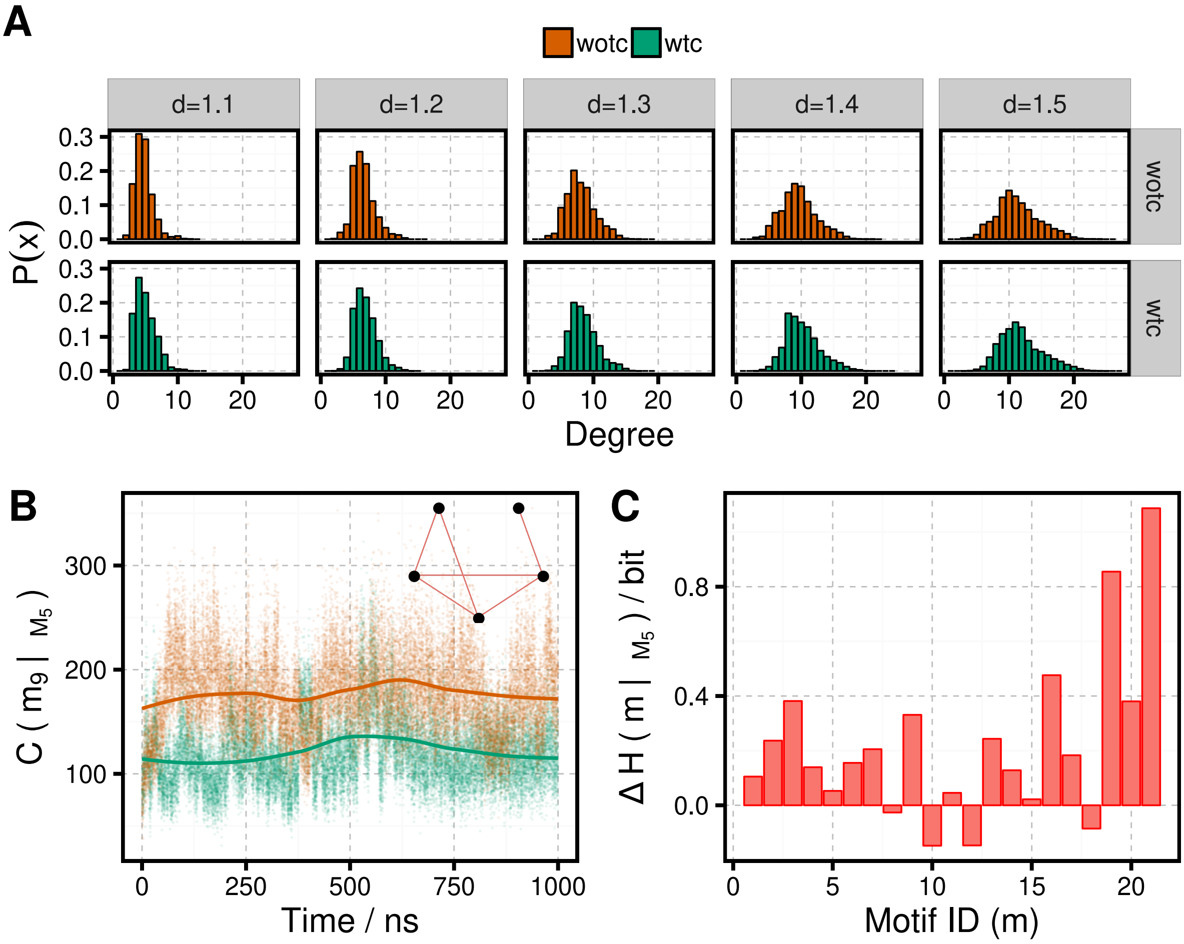


Figure 5: A) Degree Distribution for five different d of both simulations (wotc,wtc) B) Motif counts $C(m_9|_{\mathcal{M}_5})$ of a five vertex motif with the motif id 9 as a function of time. The inset shows the topology of this motif in circular layout. C) ΔH_S motif-based Shannon entropy derived from every time series counting a different five vertex motif $C\left(m|_{\mathcal{M}_5}\right)$.

4 Summary & Outlook

The `streaMD` package provides powerful tools for analysing molecular dynamics and enables researchers to use `Gromacs` output files directly in combination with `R`. Moreover, `streaMD` enables the simple exchange between MD trajectory and dynamic graph format. Users are enabled to write xtc files that can be used in combination with molecular viewers to visualize coarse-grained models in MD trajectories. Furthermore, by using `streaMD` it is now possible to combine different coarse graining methods in combination with a dynamic graph-based analysis approach. Our implementation provides at least four times parsing time and 31 times writing time speedups towards competing software packages. Furthermore, abstract graph-based analysis tools are implemented for the first time in `R` which are applicable to obtain motif-based analysis or even Markovian dynamics in motif space[14].

Project home page: `http://www.cbs.tu-darmstadt.de/streaMD.tar.gz`

5 Acknowledgments

The authors thank the LOEWE projects CompuGene and iNAPO funded by the Ministry of Higher Education, Research and the Arts (HMWK) of the Hessen state. BEM and KH acknowledge financial support by the Deutsche Forschungsgemeinschaft via the Graduate School (GRK 1657) "Radiation Biology", project 1A.

References

1. D. E. Cameron, C. J. Bashor, and J. J. Collins, Nature reviews. Microbiology **12**, 381 (2014), ISSN 1740-1534, URL `http://www.ncbi.nlm.nih.gov/pubmed/24686414`.
2. B. J. Alder and T. E. Wainwright, The Journal of Chemical Physics **31**, 459 (1959).
3. J. D. Chodera, W. C. Swope, J. W. Pitera, and K. A. Dill, Multiscale Modeling and Simulation **5**, 1214 (2006), `http://dx.doi.org/10.1137/06065146X`, URL `http://dx.doi.org/10.1137/06065146X`.
4. Q. Cui, R. K. Tan, S. C. Harvey, and D. A. Case, Multiscale Modeling and Simulation **5**, 1248 (2006), `http://dx.doi.org/10.1137/05064850X`, URL `http://dx.doi.org/10.1137/05064850X`.
5. C. Gross, K. Hamacher, K. Schmitz, and S. Jager, Journal of Chemical Information and Modeling **57**, 243 (2017), pMID: 28128951, `http://dx.doi.org/10.1021/acs.jcim.6b00556`, URL `http://dx.doi.org/10.1021/acs.jcim.6b00556`.
6. I. Callebaut, B. Hoffmann, P. Lehn, and J.-P. Mornon, Cellular and Molecular Life Sciences **74**, 3 (2017), ISSN 1420-9071, URL `http://dx.doi.org/10.1007/s00018-016-2385-9`.
7. R. Vianello, C. Domene, and J. Mavri, Frontiers in Neuroscience **10**, 327 (2016), ISSN 1662-453X, URL `http://journal.frontiersin.org/article/10.3389/fnins.2016.00327`.
8. A. de Ruiter and C. Oostenbrink, Current Opinion in Chemical Biology **15**, 547 (2011), ISSN 1367-5931, next Generation Therapeutics, URL `http://www.sciencedirect.com/science/article/pii/S1367593111000901`.
9. J. J. Montalvo-Acosta and M. Cecchini, Molecular Informatics **35**, 555 (2016), ISSN 1868-1751, URL `http://dx.doi.org/10.1002/minf.201600052`.
10. D. W. Borhani and D. E. Shaw, Journal of Computer-Aided Molecular Design **26**, 15 (2012), ISSN 1573-4951, URL `http://dx.doi.org/10.1007/s10822-011-9517-y`.

11. A. R. Benson, D. F. Gleich, and J. Leskovec, Science **353**, 163 (2016), ISSN 0036-8075, `http://science.sciencemag.org/content/353/6295/163.full.pdf`, URL `http://science.sciencemag.org/content/353/6295/163`.

12. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science **298**, 824 (2002), ISSN 0036-8075, `http://science.sciencemag.org/content/298/5594/824.full.pdf`, URL `http://science.sciencemag.org/content/298/5594/824`.

13. B. Schiller, S. Jager, K. Hamacher, and T. Strufe, in *Algorithms for Computational Biology - Second International Conference, AlCoB 2015, Mexico City, Mexico, August 4-5, 2015, Proceedings*, edited by A. H. Dediu, F. H. Quiroz, C. Martín-Vide, and D. A. Rosenblueth (Springer, 2015), vol. 9199 of *Lecture Notes in Computer Science*, pp. 53–67, ISBN 978-3-319-21232-6, URL `http://dx.doi.org/10.1007/978-3-319-21233-3{_}5`.

14. S. Jager, B. Schiller, T. Strufe, and K. Hamacher, in *Algorithms for Bioinformatics (WABI)* (Springer, 2016), vol. 9838 of *Lecture Notes in Computer Science.*

15. S. Jager, B. Schiller, P. Babel, M. Blumenroth, T. Strufe, and K. Hamacher, Algorithms for Molecular Biology **12**, 15 (2017), ISSN 1748-7188, URL `http://almob.biomedcentral.com/articles/10.1186/s13015-017-0105-0`.

16. H. Berendsen, D. van der Spoel, and R. van Drunen, Computer Physics Communications **91**, 43 (1995), ISSN 0010-4655, URL `http://www.sciencedirect.com/science/article/pii/001046559500042E`.

17. E. Lindahl, B. Hess, and D. van der Spoel, Molecular modeling annual **7**, 306 (2001), ISSN 0948-5023, URL `http://dx.doi.org/10.1007/s008940100045`.

18. R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2008), ISBN 3-900051-07-0, URL `http://www.R-project.org`.

19. F. Hoffgaard, P. Weil, and K. Hamacher, BMC bioinformatics **11**, 199 (2010), ISSN 1471-2105.

20. F. Comoglio and M. Rinaldi, Bioinformatics **28**, 1400 (2012), URL `+http://dx.doi.org/10.1093/bioinformatics/bts160`.

21. C. Margreitter and C. Oostenbrink, The R Journal **9**, 164 (2017), URL `https://journal.r-project.org/archive/2017/RJ-2017-007/index.html`.

22. B. Grant, A. Rodrigues, K. ElSawy, J. McCammon, and L. Caves, Bioinformatics **22**, 2695 (2006).

23. B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, et al., Journal of Computational Chemistry **30**, 1545 (2009), ISSN 1096-987X, URL `http://dx.doi.org/10.1002/jcc.21287`.

24. R. T. McGibbon, K. A. Beauchamp, M. P. Harrigan, C. Klein, J. M. Swails, C. X. Hernández, C. R. Schwantes, L.-P. Wang, T. J. Lane, and V. S. Pande, Biophysical Journal **109**, 1528 (2015).

25. J. P. Warnett, *Moleculardynamics*, `https://github.com/wesbarnett/MolecularDynamics.jl/releases` (2014).

26. J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, CoRR **abs/1411.1607** (2014), URL `http://arxiv.org/abs/1411.1607`.

27. D. G. Green, K. E. Meacham, and M. S. ans F. van H. J. C. Hoesel, *Parallelization of molecular dynamics code: GROMOS87 parallelized for distributed memory architectures, in 'Methods and Techniques in Computational Chemistry*, Lecture Notes in Computer Science (STEF, Cagliari, 1995).

28. M. Lundborg, R. Apostolov, D. Spngberg, A. Grdens, D. van der Spoel, and E. Lindahl, Journal of Computational Chemistry **35**, 260 (2014), ISSN 1096-987X, URL `http://dx.doi.org/10.1002/jcc.23495`.

29. D. Eddelbuettel and R. Francois, Journal of Statistical Software **40**, 1 (2011), URL `http://www.jstatsoft.org/v40/i08/`.

30. D. Eddelbuettel and C. Sanderson, Computational Statistics and Data Analysis **71**, 1054 (2014), URL `http://dx.doi.org/10.1016/j.csda.2013.02.005`.

31. H. Dien, C. M. Deane, and B. Knapp, Journal of Computational Chemistry **35**, 1528 (2014), ISSN 1096-987X, URL `http://dx.doi.org/10.1002/jcc.23650`.

32. H. Dien, C. M. Deane, and B. Knapp, Journal of Computational Chemistry **35**, 1528 (2014), ISSN 1096-987X, URL `http://dx.doi.org/10.1002/jcc.23650`.

33. W. Humphrey, A. Dalke, and K. Schulten, Journal of Molecular Graphics **14**, 33 (1996).

34. O. Mersmann, *microbenchmark: Accurate Timing Functions* (2015), r package version 1.4-2.1, URL `https://CRAN.R-project.org/package=microbenchmark`.

35. H. Wickham, *ggplot2: elegant graphics for data analysis* (2009), URL `http://had.co.nz/ggplot2/book`.

36. Schrödinger, LLC (2015).

37. A. Aleksandrov and T. Simonson, J. Comp. Chem. **30**, 243 (2009), ISSN 1096-987X.

38. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, Journal of Computational Chemistry **4**, 187 (1983), ISSN 1096-987X, URL `http://dx.doi.org/10.1002/jcc.540040211`.

39. S. Pronk, S. Pll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, et al., Bioinformatics **29**, 845 (2013), `http://bioinformatics.oxfordjournals.org/content/29/7/845.full.pdf+html`, URL `http://bioinformatics.oxfordjournals.org/content/29/7/845.abstract`.

40. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, The Journal of Chemical Physics **79**, 926 (1983), `http://dx.doi.org/10.1063/1.445869`, URL `http://dx.doi.org/10.1063/1.445869`.

41. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, The Journal of Chemical Physics **81**, 3684 (1984).

42. G. Bussi, D. Donadio, and M. Parrinello, The Journal of chemical physics **126**, 1 (2007).

43. M. Parrinello and A. Rahman, Journal of Applied Physics **52**, 7182 (1981).

44. K. Hamacher and J. A. McCammon, Journal of Chemical Theory and Computation **2**, 873 (2006), ISSN 15499618.

45. K. E. Deigan, T. W. Li, D. H. Mathews, and K. M. Weeks, PNAS **106**, 97 (2009), ISSN 0027-8424.

46. S. Pape, F. Hoffgaard, M. Dr, and K. Hamacher, Journal of Computational Chemistry **34**, 10 (2013), ISSN 1096-987X, URL `http://dx.doi.org/10.1002/jcc.23099`.

47. J. Stombaugh, C. L. Zirbel, E. Westhof, and N. B. Leontis, Nucleic Acids Research **37**, 2294 (2009), ISSN 03051048.

48. I. Hofacker, W. Fontana, P. Stadler, L. Bonhoeffer, M. Tacker, and P. Schuster, Monatshefte für Chemie/Chemical Monthly **125**, 167 (1994).

49. S. Poblete, S. Bottaro, and G. Bussi, Nucleic Acids Research p. gkx1269 (2017), URL `http://dx.doi.org/10.1093/nar/gkx1269`.

50. G. Nagy and C. Oostenbrink, Journal of Chemical Information and Modeling **54**, 278 (2014), pMID: 24364355, `http://dx.doi.org/10.1021/ci400542n`, URL `http://dx.doi.org/10.1021/ci400542n`.

A Supplementary

A.1 Motif Class

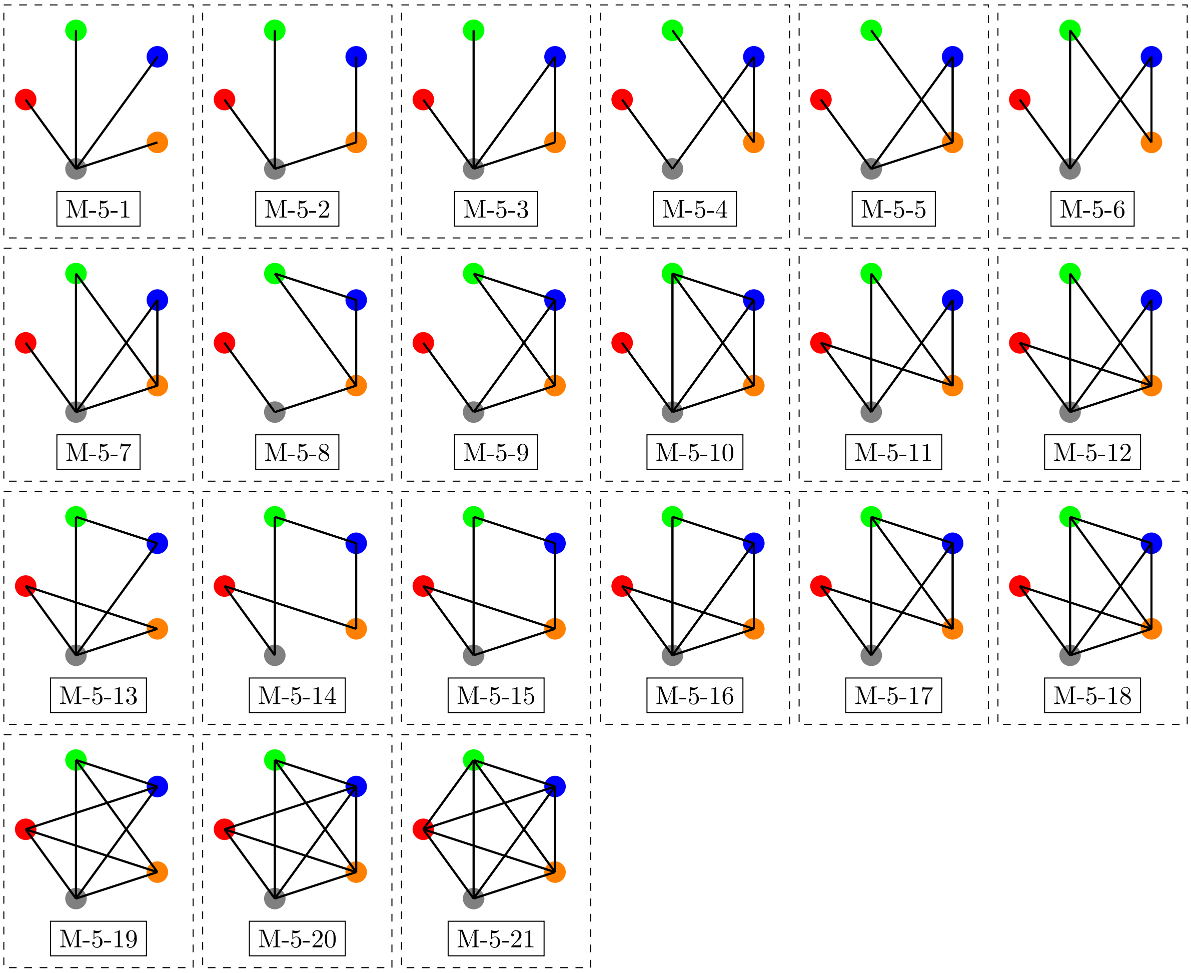


Figure 1: All motifs with five vertices. The Motif ID can be found below the circular layout of the given motif.

A.2 Parallelization Benchmark

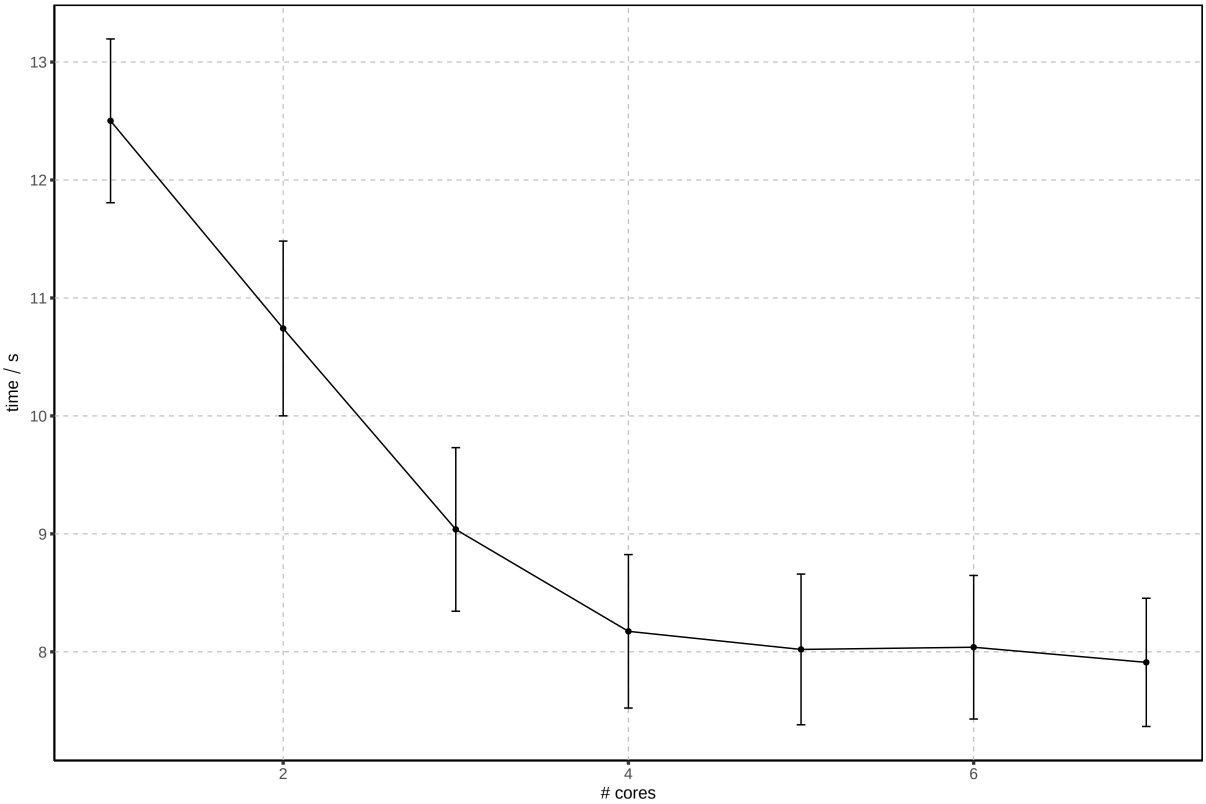


Figure 2: Runtime of `fast_rmsd()` against number of processor cores. The benchmark was performed 500 times using 150,000 frames with 5232 atoms each. All runtimes were measured on 4 *AMD Opteron 6274* processors with 2.2 GHz.

7 Graph-based Analysis of HTS-SELEX

7.1 Riboswitching with ciprofloxacin - Development and characterization of a novel RNA regulator

This section deals with the contribution in the following work:

- Groher, F., Bofill-Bosch, C., Schneider, C., Braun, J., **Jager, S.**, Geißler, K., Hamacher, K., Suess, B., (2018) Riboswitching with ciprofloxacin – Development and characterization of a novel RNA regulator. Nucleic Acids Research

The paper describes an successful HTS approach of an *ciprofloxacin* (CFX)-Riboswitch using SELEX in Combination with NGS. For this purpose, an existing aptamer domain, which recognizes CFX as a building block for the SELEX process, was employed. CFX is a well-known fluoroquinolone antibiotic, FDA approved and has a favorable toxicity profile for many organisms. Furthermore, cellular uptake is granted in both, lower and higher eukaryotes.

Contributions The CFX-SELEX dataset was analyzed with support of Florian Groher. My task was to calculate the structural/sequential diversity of the whole SELEX dataset. Especially the structure prediction provides a particularly meaningful example, since a small change in the sequence can cause a significant change in the resulting MFE structure. Moreover, this information is helpful since it shows in which round the structural diversity decreases. Specially, in these rounds certain species (motifs, structural) have become prevalent and the success is high that these structural motifs can also be used in riboswitches. Identifying these rounds helps to optimize the screening, because the further processing (e.g. measurements, cloning) is time and material intensive.

On account of the very large datasets the LD was used in this work. The upper triangle matrix was computed for all sequences and structures for each SELEX round one respectively. The *Kolomogrov Smirnov* (KS) test was used to determine how distant the distributions are from the initial randomized pool during the SELEX process.

Nucleic Acids Research, 2018 **1–12**
doi: 10.1093/nar/gkx1319

Riboswitching with ciprofloxacin—development and characterization of a novel RNA regulator

Florian Groher[1], Cristina Bofill-Bosch[1], Christopher Schneider[1], Johannes Braun[1], Sven Jager[2], Katharina Geißler[1], Kay Hamacher[2,3] and Beatrix Suess[1,*]

[1]Synthetic Genetic Circuits, Dept. of Biology, TU Darmstadt, Darmstadt, Germany, [2]Computational Biology and Simulation, Dept. of Biology, TU Darmstadt, Darmstadt, Germany and [3]Dept. of Physics, Dept. of Computer Science, TU Darmstadt, Darmstadt, Germany

Received November 03, 2017; Revised December 22, 2017; Editorial Decision December 27, 2017; Accepted December 28, 2017

ABSTRACT

RNA molecules play important and diverse regulatory roles in the cell. Inspired by this natural versatility, RNA devices are increasingly important for many synthetic biology applications, e.g. optimizing engineered metabolic pathways, gene therapeutics or building up complex logical units. A major advantage of RNA is the possibility of *de novo* design of RNA-based sensing domains via an *in vitro* selection process (SELEX). Here, we describe development of a novel ciprofloxacin-responsive riboswitch by *in vitro* selection and next-generation sequencing-guided cellular screening. The riboswitch recognizes the small molecule drug ciprofloxacin with a K_D in the low nanomolar range and adopts a pseudoknot fold stabilized by ligand binding. It efficiently interferes with gene expression both in lower and higher eukaryotes. By controlling an auxotrophy marker and a resistance gene, respectively, we demonstrate efficient, scalable and programmable control of cellular survival in yeast. The applied strategy for the development of the ciprofloxacin riboswitch is easily transferrable to any small molecule target of choice and will thus broaden the spectrum of RNA regulators considerably.

INTRODUCTION

RNA devices became a key focus of synthetic biology in recent years. They have been used to implement genetic circuits and networks based on small regulatory RNAs, e.g. toehold-switches (1) or STARs (2), synthetic riboswitches (3) and allosterically controlled ribozymes (4–7). The fast progression of this development resulted in the transition from simple proof of concept to sophisticated and useful applications targeting complex problems (8). RNA devices are unique due to their modular nature that allows the simple and straightforward linkage of different domains, e.g. between a sensor and an actuator. Thus, a whole range of different functions may be united in one RNA molecule. Although it is very common to select natural regulatory domains and adapt them for different purposes, there is also the option of *de novo* generation of RNA sensor domains. In principle, the powerful *in vitro* selection method (SELEX) (9,10) allows the selection of a suitable sensor RNA for any desired target molecule. Sensor RNAs thus designed recognize their target with great affinity and specificity. However, despite the fact that several dozen small molecule-binding aptamers have been generated to date (11), only a handful of those are suitable for the design of RNA devices (12). Theophylline- and tetracycline-binding aptamers have been the most successful here (13). They allow the construction of synthetic riboswitches that may be used for control of transcription termination (14), or translation initiation (15), mRNA splicing (16) or control of mRNA stability (17). In contrast to natural riboswitches that are mainly found in bacteria (18), synthetic riboswitches could be developed for all three domains of life (3).

A wealth of adroit proof-of-concept studies demonstrating the application of synthetic riboswitches are available nowadays. However, all studies to date are exclusively limited to the theophylline or tetracycline aptamer systems, which effectively prevents a wider application of synthetic riboswitches. To remedy this shortage of applications and to stimulate and invigorate the field of synthetic riboswitch development, the repertoire of aptamers suitable and available for riboswitch design needs to be extended. First and foremost, methodology for the identification of such aptamer domains is required.

The main reason for the limited suitability of most aptamers is that both excellent binding properties and conformational switching are essential, yet the latter is a feature not addressed during the process of *in vitro* selection (3). To find aptamers that combine superior binding properties and the ability to undergo conformational switching, cellular screening after *in vitro* selection is required. Such screen-

---

[*]To whom correspondence should be addressed. Tel: +49 6151 1622000; Fax: +49 6151 1622003; Email: bsuess@bio.tu-darmstadt.de

© The Author(s) 2018. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

ing systems have been established (19–21) and their functionality has been successfully demonstrated, e.g. for the neomycin aptamer (22). Now, we have extended the method to include next-generation sequencing (NGS). NGS has revolutionized not only aptamer selection, but proved the proverbial game changer for research across most disciplines of the life sciences (23). The application of NGS allows the collection of detailed information for the individual selection rounds. Thus, it was possible to choose selection rounds that showed a certain degree of enrichment, yet maintained maximum diversity. We assume that this approach will allow a substantial acceleration of the transition between *in vitro* and *in vivo,* while simultaneously reducing screening efforts.

In the present study, we demonstrate the approach described above for ciprofloxacin (CFX) as it is a well-known fluoroquinolone antibiotic, FDA-approved (24) and with a favorable toxicity profile for many organisms (25,26). Thus, it guarantees portability and broad applicability. Furthermore, cellular uptake is granted in both lower and higher eukaryotes (27). We were able to generate a CFX-binding aptamer with riboswitching properties. Structural probing revealed a pseudoknot structure that enfolded and essentially sealed the binding pocket. It showed functionality both in yeast and a human cell line and is sufficiently efficient to control cell fate by blocking pyrimidine metabolism or a resistance gene.

In sum, we demonstrate here the *de novo* development of a novel small molecule-dependent synthetic riboswitch. We characterized a robust procedure for development that may be used as a template for application with any other ligand. Thus, our findings present an ideal basis and a springboard to jumpstart the wide application of aptamer-based riboregulators.

MATERIALS AND METHODS

Pool preparation

For *in vitro* selection experiments, we used a 1:1 mixture of a completely randomized and a partially-structured library (28). In short, the completely randomized library consists of 64 nucleotides (nt) whereas the partially structured library contained a 12-nt long fixed sequence (5′-CTG CTT CGG CAG-3′) flanked by 26 random nt on each side. Both libraries are flanked by constant regions (5′ constant: 5′-GGG AGA CGC AAC TGA ATG AA-3′/3′ constant: 5′-TCC GTA ACT AGT CGC GTC AC-3′) for amplification using the oligonucleotides Pool_fwd 5′-GTA TAA TAC GAC TCA CTA TAG GGA GAC GCA ACT GAA TGA A-3′ and Pool_rev 5′-GTG ACG CGA CTA GTT ACG GA-3′). Both pools were amplified using the following PCR conditions: 10 mM Tris–Cl (pH 9.0), 50 mM KCl, 1.5 mM $MgCl_2$, 0.1% Triton X-100, 0.2 mM dNTPs (each), 30 nM pool template, 2 μM Pool_fwd, 2 μM Pool_rev, 50 U/ml Taq DNA Polymerase (NEB). 10^{15} pool template molecules were amplified in a 60 mL PCR reaction for only 7 cycles to reduce PCR-induced bias. PCR efficiency was calculated according to Hall *et al.* (29).

After large-scale amplification, DNA pool template was ethanol-precipitated, dissolved in MQ-H_2O [de-ionized water purified with ion exchange resin and filtered through a Biofilter (ELGA) to remove possible RNase contamination] and subsequently phenol:chloroform extracted (30). The purified DNA template was transcribed using T7 RNA polymerase as described previously (31). Afterwards, the transcribed RNA was ethanol-precipitated, dissolved in formamide containing 25 mM EDTA and loaded on a 6% denaturing polyacrylamide gel (8 M urea). The RNA was visualized by UV shadowing, sliced out and eluted overnight in 300 mM Na-acetate (pH 6.5). Hereafter, eluted RNA was ethanol-precipitated, the pellet was redissolved in a suitable amount of water and molarity was calculated.

CFX immobilization

CFX was immobilized on Profinity™ Epoxide Resin (Bio-Rad). For this, 2 g dry resin was swollen in MQ-H_2O, twice washed with MQ-H_2O and vacuum-filtered. After a second wash with coupling buffer (50 mM KCl, 132 mM NaOH, pH 13.0), the resin was mixed 1:2 with 5 mM CFX solution in coupling buffer. The reaction was protected from light and incubated over night at room temperature (RT) on an H5600 rotator (Labnet). Afterwards, the resin was washed with MQ-H_2O, vacuum-filtered and remaining active groups were blocked by incubation with 1 M ethanolamine (MEA) for 4 h. Finally, the CFX-coupled resin was washed according to the supplier's instructions with alternating buffer change from pH 4.0 (100 mM acetate, 500 mM NaCl) to pH 8.0 (100 mM phosphate, 500 mM NaCl). Lastly, the resin was washed with MQ-H_2O and stored in 0.02% (w/v) NaN_3 at 4°C in the dark for up to 3 months.

In vitro selection

For the first round of *in vitro* selection, 1.2×10^{15} RNA molecules from the initial pool (1:1 mixture of completely randomized and pre-structured pool, see above) were spiked with ∼250 kCPM of 5′ ^{32}P-labeled RNA pool in MQ-H_2O. RNA folding was performed by heating the mixture to 95°C for 5 min and subsequently placed on ice water for additional 5 min. After the folding step, yeast tRNA was added to a final concentration of 1 mg/ml and the volume was adjusted to 1 column volume (CV, 500 μl) with 1× binding buffer (40 mM HEPES pH 7.4, 125 mM KCl, 5 mM $MgCl_2$, 5% DMSO), respectively. For depletion of RNAs able to bind the affinity matrix, the RNA library was first incubated for 30 min with 1 CV of a non-derivatized column (mock). The mock column consisted of Profinity™ Epoxide Resin that had been treated only with MEA instead of CFX. After negative selection, unbound RNAs were added to 1 CV CFX-coupled resin and incubated for 30 min at RT. Next, the column was washed with 10 CV binding buffer and bound RNAs were eluted with either 4 CV 20 mM EDTA (round 1–5) or 4 CV 1 mM CFX (round 6–10) in 1× binding buffer.

Eluted RNA was ethanol-precipitated with Na-acetate in the presence of 15 μg GlycoBlue™ Coprecipitant (Ambion) and washed twice with 70% (v/v) ethanol. The air-dried pellets were dissolved in a total volume of 50 μl MQ-H_2O and reverse-transcribed and amplified (RT-PCR). For RT-PCR, 50 μl eluted RNA was mixed with 1× PCR buffer

(10 mM Tris–Cl pH 9.0, 50 mM KCl, 0.1% Triton X-100), 1× first strand buffer (Invitrogen), 2 mM DTT (Roche), 1 μM Pool_fwd, 1 μM Pool_rev, 1.5 mM $MgCl_2$ and 0.3 mM dNTPs (each). The reaction was heated to 65°C for 5 min and then quickly placed on ice. After that, 5 U *Taq* DNA Polymerase (NEB) and 200 U SuperScript™ II (Thermo Fisher Scientific) were added to the reaction and RNA was reverse-transcribed and amplified (54°C for 10 min followed by 6–10 cycles of 95°C for 1 min, 58°C for 1 min and 72°C for 1 min). Product formation was monitored on a 3% agarose gel.

For the following rounds, RNA was transcribed following (31). In short, 10 μl of RT-PCR product was mixed with 40 mM Tris–Cl (pH 8.0), 5 mM DTT, 2.5 mM NTPs (each), 15 mM $MgCl_2$, 100 U T7 RNA Polymerase (NEB), 40 U ribonuclease inhibitor (moloX) and 33 nM ^{32}P-α-UTP (Hartmann analytics) in a total volume of 100 μl. Transcription was carried out at 37°C for 1 h. Afterwards, transcription was precipitated with NH_4-acetate/ethanol, washed twice with 70% EtOH and the pellet was dissolved in a suitable amount of water. Five hundred kCPM RNA was folded, diluted in 1× binding buffer and subsequently loaded onto the column for the next round of SELEX.

Plasmid cloning and doped pool generation

All plasmids and oligonucleotides used in this study are listed in Supplementary Table S1 and Supplementary Table S2, respectively. For cloning, two 30 bp overlapping oligonucleotides were designed and amplified using Q5® High-Fidelity DNA polymerase (NEB) according to the supplier's instructions. The resulting PCR product was purified (QIAquick PCR Purification Kit, Qiagen), digested with AgeI-HF and NheI-HF (NEB) and ligated into equally digested pWHE601* with T4 DNA Ligase (NEB).

Doped pools were generated using the oligonucleotides AgeI_doped_fwd and NheI_[3.0/4.5/9.0/30.0]_doped_rev (Supplementary Table S3, Microsynth AG), respectively, with the construct ΔATG as template and amplified using Q5® High-Fidelity DNA polymerase (NEB) according to the supplier's instructions. Again, digestion and ligation into pWHE601* followed. Transformation of the ligation mixture was done after butanol precipitation into NEB® 10-beta Competent *Escherichia coli* (High Efficiency) according to the supplier's protocol. This ensured that the number of different plasmids yielded by this process were >50,000.

Cultivation of yeast and GFP measurements

The *Saccharomyces cerevisiae* strain RS453α (*MATα ade2-1 trp1-1 can1-100 leu2-3 his3-1 ura3-52*) (32) was transformed using Frozen-EZ Yeast Transformation II Kit (Zymo Research). Transformed cells were plated on SCD-ura plates [0.2% YNB w/o AA (Difco), 0.55% ammonium sulfate (Roth), 2% glucose (Roth), 12 μg/ml adenine (SIGMA), 1× MEM amino acids (SIGMA), 2% Agar (Oxoid)] and incubated at 30°C for 3 days in a humidified incubator. Single colonies were picked and cultured in 1.5 ml SCD-ura for 24 h (450 rpm, 30°C, 24-well plates) before they were diluted 1:1000 in fresh media with and without 1 mM CFX. Again, after 24 h incubation cells were washed twice with 1× PBS and diluted to an OD_{600} of ∼0.4 for fluorescence measurements.

Fluorescence measurements were performed on a Fluorolog FL3-22 (Horiba Jobin Yvon) with an excitation wavelength set to 474 nm (slit 2 nm) and an emission wavelength of 509 nm (slit 2 nm). The integration time was set to 0.5 sec and temperature was adjusted to 28°C. Afterward, OD_{600} for each culture was determined and fluorescence intensity was normalized to it. As negative control, pWHE601* (21) was analyzed in parallel as a blank and its value was subtracted from all data. Yeast cells containing pWHE601* are referred to as GFP– cells, whereas cells expressing GFP (transformed with pWHE601 (33)) are referred to as GFP+ cells. Both controls are treated equally as the riboswitch-controlled constructs. Each experiment was done in duplicates and reproduced at least three times.

In vivo screening

Library preparation and *in vivo* screening was performed according to the established protocol by Suess *et al.* (34) with modifications described in Schneider *et al.* (21). In short, libraries for *in vivo* screening were cloned by homologues recombination in yeast. For that, RT-PCR product of a defined round was amplified with CFX_HR_fwd (5′-CAA GCT ATA CCA AGC ATA CAA TCA ACT CCA AGC TAG ATC TAC CGG TGG GAG ACG CAA CTG AAT GAA-3′) and CFX_HR_rev (5′-CAA GAA TTG GGA CAA CTC CAG TGA AAA GTT CTT CTC CTT TGC TAG CGT GAC GCG ACT AGT TAC GGA-3′) to attach 46 bp overhang for recombination into pWHE601*. Target vector pWHE601* was digested using AgeI-HF and NheI-HF (both NEB) and transformed into RS463α with a 10-molar excess of insert using Frozen-EZ Yeast Transformation II Kit (Zymo Research). Transformed cells were spread on several SCD-ura plates in a way that assures a moderate colony density which simplifies picking clones. For the first screening round, cells were selected under the fluorescence binocular and checked for GFP expression. Clones with low and moderate fluorescence were picked and transferred to a 96-well plate with 200 μL SCD-ura. After sealing the plates with BREATHseal™ (Greiner Bio-One), cells were incubated for 24 h at 30°C on the plate shaker Titramax 1000 (Heidolph instruments) at 450 rpm. Next, cells were split 1:10 in fresh SCD-ura containing no ligand (control) or 1 mM CFX [preliminary experiments showed no influence on cell growth or GFP signal intensity with CFX concentrations up to 10 mM (Supplementary Figure S1)]. Again, after 24 h incubation cells were diluted 1:10 with 1x PBS and GFP fluorescence and OD_{600} was measured. As positive control pWHE601 (GFP+) and as negative control pWHE601* (GFP-) were measured in parallel and used for normalization (pWHE601) and for subtracting autofluorescence of yeast and media (pWHE601*). Positive hits were streaked out on SCD-ura plates and incubated. From this, four independent colonies were picked and screened again with the protocol above. Verified hits were taken for plasmid preparation using QIAprep Spin Miniprep Kit (Qiagen) and the user-developed protocol from Michael Jones (protocol PR04 'Isolation of plasmid DNA from yeast'). Plasmids were passaged trough *E. coli* DH5α and sequenced with GFP_rev (5′-CCA CTG ACA GAA AAT TTG TGC-3′). Unique candidates were then transformed back into yeast and GFP fluorescence was measured again.

NGS library preparation and data analysis

Barcodes were attached to all selection rounds by PCR using the oligonucleotides Seq_IL_fwd and Seq_IL_rev[0-10] (Supplementary Table S4). Forward and reverse oligonucleotides hybridizes at the 5′ and 3′ constant regions, respectively, thus the sequence of the T7 polymerase promoter was removed. The oligonucleotides Seq_IL_rev[0-10] introduced a 4-mer barcode to assign each sequence to the specific round after sequencing. After amplification, the samples were Gel-purified (Zymoclean Gel DNA Recovery Kit, Zymo Research) and mixed in equimolar amounts for Illumina sequencing reaction (GenXPro GmbH).

To monitor the enrichment process of single sequences and to characterize the SELEX process, we computed for each round of selection a Levenshtein distance distribution from sequence and structural data. The Levenshtein distance measures the difference between two sequences by calculating the smallest number of insertions, deletions, or substitutions necessary to transform one character string—such as a biomolecular sequence, or a RNA secondary structure—into another (35). We computed the Levenshtein distance between every sequence within each selection round. To compare RNA structures, all sequences were folded with RNAfold 2.3.4 (36) at 300 K with the thermodynamic parameter set from Andronescu *et al.* (37). Here, the Levenshtein distance was computed between the respective dot-bracket annotations of the RNA molecules. Afterward, the Levenshtein distance was normalized by their respective reads per million (RPM) value for both sequences. For every round, a histogram was generated, followed by calculating its cumulative frequency distribution (CFD) followed by normalization by the number of data to cumulative probabilities [$P(x)$]. To assess the distances between these Levenshtein distance distribution, we computed the Kolmogorov–Smirnoff statistic (38), CDF $D := |\ F_n - F_0| = sup_x\ |F_n(x) - F_0(x)|$, where F_n is a cumulative distribution function (CDF) derived from cumulative probabilities of the respective Levenshtein distance distributions obtained from the nth SELEX round. Accordingly, F_0 is the CDF obtained from the first round and *sup* is defined as the supremum of the set of distances.

RNA synthesis for *in vitro* analysis

For *in vitro* analysis (in-line probing, fluorescence titration experiments and ITC measurements), RNA was transcribed from PCR-generated templates, all containing at least one 5′-terminal guanosyl residue to facilitate transcription *in vitro* using T7 RNA polymerase. For this, two oligonucleotides were designed with an overlap of 30 bp (Supplementary Table S5) and amplified using Q5® High-Fidelity DNA polymerase (NEB) according to the supplier's instructions. After ethanol precipitation, the DNA template was used for *in vitro* transcription with T7 RNA polymerase (NEB) as reported previously (31). The RNA was gel purified (39) and molarity was determined by spectrophotometric measurement using NanoDrop 1000 Spectrophotometer (Thermo Scientific).

In-line probing experiments

For in-line probing, RNA was dephosphorylated and 5′ ^{32}P-labeled as previously described (40). After PAGE purification, 35 kcpm of each 5′ ^{32}P-labeled RNA were incubated for 68 h at 22°C in in-line reaction buffer (10 mM Tris–Cl pH 8.3 @ 20°C, 10 mM $MgCl_2$, 100 mM KCl). To generate a size marker, the 5′ ^{32}P-labeled RNAs were subjected to alkaline hydroxylation by incubation for 3 min at 96°C in 50 mM Na_2CO_3 (pH 9.0), or incubated for 3 min at 55°C with 20 U RNase T1 at denaturing conditions to identify guanines (41). After in-line reaction, alkaline hydroxylation or RNase T1 treatment, reactions were ethanol precipitated and the pellet was dissolved in 5 M urea. All reactions were separated by denaturing polyacrylamide gel electrophoresis. Afterward, gels were dried and analyzed using phosphoimaging (GE Healthcare).

Fluorescence titration experiments

Dissociation constants (K_D) for RNA-CFX complexes were determined by measuring the fluorescence quenching as a function of RNA concentration in the presence of a fixed CFX concentration. Fluorescence intensities were measured on a Fluorolog FL3-22 (Horiba Jobin Yvon) with an excitation wavelength set to 335 nm (slit 5 nm) and an emission wavelength of 420 nm (slit 5 nm). The integration time was set to 0.5 s and temperature was adjusted to 25°C. In between, the addition of RNA, the reaction was stirred for 1 min and equilibrated for an extra minute. For the titration experiments, 50 nM CFX in 1× binding puffer (F_0) was mixed with increasing amounts of gel-purified RNA and fluorescence intensity was measured (F). Prior to the titration experiment, RNA solutions were heated to 95°C for 5 min and snap-cooled on ice for 5 min (RNA folding step). After that, binding buffer was added to a final concentration of 40 mM HEPES, 125 mM KCl, 5 mM $MgCl_2$, pH 7.4.

Curve fitting was done using Prism (GraphPad Software) and nonlinear regression analysis with following equation by least squares fitting: $Y = B_{max} * X^h/({K_D}^h + X^h)$, with B_{max} = maximum binding, h = hill slope, X = concentration of RNA.

Isothermal titration calorimetry

RNA folding and buffer compositions were chosen according to the fluorescence titration experiments. 100×10^{-6} M CFX solution were prepared in the same buffer. ITC experiments were carried out with an MicroCal PEAQ-ITC (Malvern Instruments) with the sample cell (200 μl) containing 10×10^{-6} M RNA and 100×10^{-6} M CFX solution in the injector syringe (40 μl). After thermal equilibration at 25°C, an initial 150 s delay and one initial 0.4 μl injection, 12 serial injections of 3.0 μl at intervals of 150 s and at a stirring speed of 750 rpm were performed. Raw data were recorded as power (μcal s^{-1}) over time (min). The heat associated with each titration peak was integrated and plotted against the corresponding molar ratio of CFX and RNA. The dissociation constant (K_D) was extracted from a curve fit of the corrected data by use of the one-site binding model provided by MicroCal PEAQ-ITC Analysis Software 1.1.0.1262. Measurements were repeated at least twice.

Serial dilution growth assay

Overnight cultures were either grown in YPD [1% yeast extract (Oxoid), 2% peptone (BD), 2% glucose (Roth), 2% Agar] supplemented with 0.5 mg/ml G418 or in SCD-ura. Both YPD and SCD-ura were supplemented with CFX to a final concentration of 1 mM to condition the cells to the OFF-state. After overnight incubation, cultures were 1- to 5-fold diluted in fresh media and grown to an OD_{600} of 1–2. Cells were washed with 1× PBS and diluted to an OD_{600} of 10.0 followed by 6-fold 1:10 serial dilution in 1× PBS (denoted as 0.6 respectively). From the diluted cultures, 5 μl were spotted onto SCD-ura plates supplemented with 0.5 mg/ml G418 in the absence (control) or presence of 1 mM CFX. Growth differences were recorded following incubation of the plates for 2–3 days at 30°C.

Dual luciferase assay

One day before transfection, HeLa cells were transferred to a 24-well plate (40 000 cells/well in 1 ml DMEM). According to the manufacturer's protocol, 1 μl Lipofectamine 2000 (Invitrogen) and 250 ng pDNA was used for transfection. After 2 h, transfection medium (Opti-MEM) was replaced by fresh medium supplemented with or without 100 μM CFX (Sigma-Aldrich). Luminescence was measured 24 h post transfection using the Dual-Glo Luciferase Assay System according to the manufacturer's instructions (Promega). Luminescence was detected using an Infinite M200 Microplate Reader (Tecan). The ratio between firefly and *Renilla* luciferase activity was calculated for each well to normalize for transfection efficiency. Mean values and standard deviations were calculated from triplicates and normalized to the values of the corresponding vector without riboswitch. Each experiment was repeated at least three times.

RESULTS AND DISCUSSION

Identification of CFX-binding aptamers by *in vitro* selection (SELEX)

To select aptamers that recognize CFX, we immobilized CFX directly to an epoxy-activated, solid polyacrylamide support (Figure 1A). The reaction conditions were adjusted to a slight molar excess of CFX compared to accessible reactive epoxy groups on the column. In consequence, we assume that under these alkaline conditions, the epoxy group mainly reacts with the secondary amino group of the piperazinyl residue, exposing CFX to the solvent (42).

The RNA library with a starting diversity of 1.2 × 10^{15} RNA molecules included a 64 nt-long random region with half of it containing a small stem loop in the middle, a library composition already established (28). It was discussed that preformed stem-loops provide favourable conditions for aptamer selection by acting as nucleation sites for RNA structure formation (43–45). We had no *a priori* knowledge of the nature of the aptamer we were exploring, including both a completely randomized region and a preformed stem loop gave us the full scope to unrestrainedly select for the best fit.

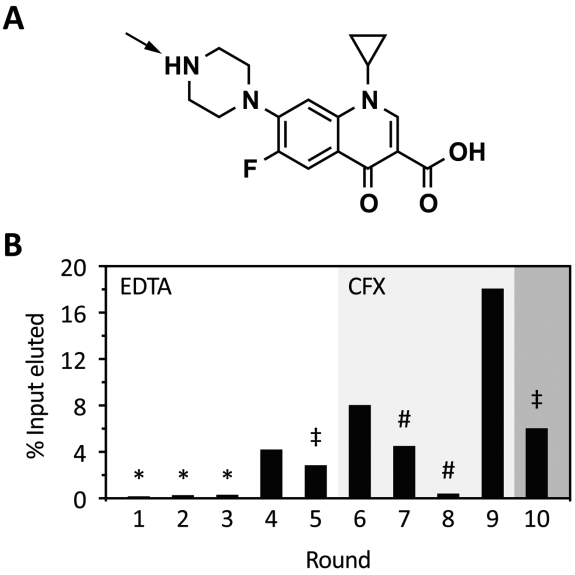


Figure 1. Progress of *in vitro* selection. (**A**) Chemical structure of CFX. The arrow indicates the most likely attachment site to the epoxy-activated PAA-matrix. (**B**) Shown is the fraction of loaded RNA that could be eluted from CFX-derivatized columns after each selection round. RNA was eluted by either 20 mM EDTA (round 1–5) or 1 mM CFX (round 6–10). In the first three rounds, a negative selection was performed (*). In round 5 and 10, stringency was increased by doubling the number of column washes or a decrease in the concentration of immobilized CFX to one-tenth, respectively (‡). Pre-elution steps were performed in round seven and eight (#) (for further details see also Supplementary Table S6).

In the first five rounds, we eluted unspecifically with EDTA to ensure elution of every RNA molecule, neglecting their binding properties. This approach should guarantee that the first enrichment of the pool introduces no bias toward low affinity aptamers because of the mild selection conditions. Furthermore, in the first three rounds, a negative selection step was carried out to remove RNA molecules that recognize the solid support. The amount of eluted RNA in these early rounds was as expected to be very low (details in Supplementary Table S6) since most of the RNA molecules of the randomized pools do not recognize the ligand. After a first enrichment in round 4 (Figure 1B), we increased stringency by increasing the number of washing steps (round 5) or switching to specific elution with CFX in round 6. Despite increased stringency, more RNA was eluted from the column. In consequence, we decided to implement a pre-elution step in rounds 7 and 8 to eliminate RNA species with fast K_{off} rates (11). Additionally, we reduced the amount of immobilized ligands to one-tenth in the last round of selection. A detailed summary of the selection process can be found in Supplementary Table S6.

For a first glimpse of the selection progress, we sequenced 23 candidates of round 10. We found 13 different sequences, eight of them were unique and about half of them contained the predefined stem loop, but no shared motifs could be found (Supplementary Table S7). Binding capacities of all individual candidates were analyzed by their interaction with the CFX-derivatized column (Supplementary Figure S2A). Most of the aptamers showed a strong interaction with the column and could be specifically eluted with CFX. Some candidates included in the analysis did not perform as expected, e.g. R10K3 showed a weak interaction and R10K9 an interaction similar to the entire pool binding capacity in round 10. However, R10K6, R10K7 and R10K4 showed an elution profile up to 4-fold improved. Four candidates were selected for quantification of CFX binding by fluorescence titration experiments. Here, the intrinsic fluorescence of CFX and the respective quenching upon RNA-binding were used to determine the dissociation constant (K_D) of the respective aptamers. For all tested RNAs, the K_D was below 100 nM (Supplementary Figure S2B). Our analysis suggested that the enrichment process yielded aptamers of the desired high binding affinity. However, we simultaneously managed to maintain a sufficient diversity of candidates with adequate structural flexibility, i.e. ideal conditions for a detailed examination of the sequence composition and subsequent riboswitch screening.

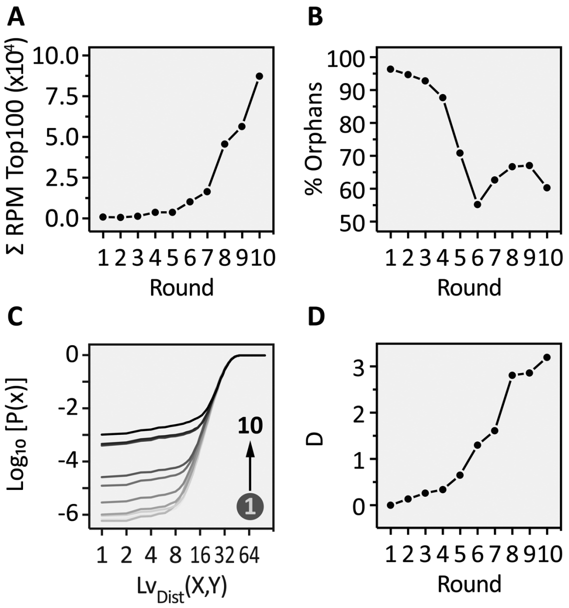


Figure 2. NGS analysis of CFX *in vitro* selection. (**A**) Over the course of the SELEX experiment, the most abundant sequences (Top100 of each round) were enriched in an exponential fashion. (**B**) Increased stringency over time reduced the amount of background ('Orphans') continuously till a plateau was reached (round 6–10). (**C**) The cumulative distribution function (CDF) for each round based on calculated Levenshtein distances on MFE structures is plotted for each round, resulting in an increased P(x) over the selection experiment. Shown are the results for the Top1000. (**D**) Based on CDF, D was derived and its logarithm is plotted against the selection rounds for the Top1000. Here, D is computed between the first round and all remaining.

Deep sequencing of aptamer selection populations

Selection experiments aim to enrich aptamers with high binding affinity and specificity for their respective ligands. However, our practical experience in recent years clearly demonstrates that superior binding affinity of aptamers alone is insufficient for successful development into riboswitches. On the contrary, a subsequent screening step has proved indispensable (12,22,46). Implementing NGS into our workflow was instrumental in the identification of selection rounds that were best-suited for the laborious screening for switching aptamers, which considered both sequence and structure diversity in conjunction with library enrichment.

By Illumina sequencing, we obtained a total number of 4.2 million reads for all investigated rounds, of which 92% could be sorted according to their corresponding barcoding (Supplementary Table S4). Next, identical sequences were summed up and the total read count was normalized for each round to reads per million (RPM).

A statistical analysis was performed for the NGS data. We determined the enrichment of the 100 most abundant sequences (Top100, Figure 2A) and the proportion of the background or so-called orphans (Figure 2B). For the most abundant sequences, we see a clear exponential enrichment throughout the SELEX process, which is in line with our expectations. On the other hand, the number of orphans drops constantly up to round 6 and remains at a low level to the end of selection. In addition to the sequence-based analysis, we also performed a structural evaluation. For this, we predicted MFE structures of all sequences. We compared similarities of the dot-bracket annotation by calculating the Levenshtein distance $Lv_{Dist}(X, Y)$ (35) of all predictions within each round with each other. In our experience, sequence-based motif search will only yield useful results if the pool is enriched to a certain degree. Since we evaluated all rounds of selection, structural similarity rather than motifs will allow a better estimation of pools diversity (47). The histogram of the $Lv_{Dist}(X, Y)$ distribution was converted to a cumulative distribution function (CDF), so that the diversity of the pool of each round can be easily assessed (Figure 2C). As a distance measure between the obtained CDFs, we used Kolmogorv–Smirnoff's D (ks-test) which computes the supremum $(D := | F_n - F_0| \; = sup_x \; |F_n(x) - F_0(x)|)$ between two CDFs (Figure 2D).

In contrast to the cumulative RPM of the Top100 enriched sequences, we observed neither a gradual nor an exponential increase in enrichment (compare Figure 2A with D). Rather, we found a more uneven distribution of enrichment. These findings can be correlated to the SELEX procedure (Figure 1B). The differences between the rounds can be assessed by looking at the distance measure D (Figure 2D). We observed the highest increase in enrichment between rounds 5/6, 7/8 and 9/10. All of these large enrichment steps can be correlated to the experimental conditions applied in the corresponding rounds. In round 6, we switched from EDTA to CFX elution, in round 8 we applied the pre-elution step twice and drastically reduced the amount of eluted RNA and in round 10 we reduced the amount of immobilized CFX to one-tenth. Interestingly and counterintuitively, in round 9 we found nearly the same sequence and structure distribution compared to round 8 and the change in distance between 8 and 9 is almost zero. Although we could elute the highest amount of RNA from the column, the removal of the stringency (no pre-elution) led to an amplification-only round, where no further selection took place. These results suggest that selection pressure should be kept constant or increased over the experiment, but not omitted. Otherwise other factors can influence pool sequence distribution, such as RT-PCR or *in vitro* transcription, which may introduce bias. In sum, NGS is able to determine structural diversity and by doing so building a foundation for selection the best-suited rounds for *in vivo* screening. With respect to diversity, we have chosen round 6 and round 10 as the libraries to start screening. In round 6, we observed a prominent structural enrichment for the first time, whereas we consider round 10 the most enriched library.

NGS guided *in vivo* screening and riboswitch engineering

Aptamer sequences from round 6 and 10 were cloned into the 5′ UTR of a constitutively driven *gfp*+ gene. The pools were integrated into *S. cerevisiae* by homologous recombination and analyzed for CFX-dependent changes in fluorescence (21). We screened 6000 colonies in total for both rounds and discarded candidates with a fluorescence signal considered too low. Here, inserted aptamers were either already too structured or the insertion into the vector, which by default lacks a start codon, failed. Based on empirical knowledge, we also discarded candidates with very high fluorescence indicating the absence of a structured RNA. After this initial elimination procedure, 17% of the total that initially showed a GFP fluorescence in the desired range remained. For 599 and 435 clones for round 6 and round 10, respectively, fluorescence was measured in the absence and presence of CFX and the regulatory activity was calculated (Figure 3A). In round 6, no candidates showed any changes in fluorescence. In round 10, two candidates were identified with 1.7- and 1.5-fold decrease in GFP expression, respectively.

We decided to continue with candidate 10A and partially randomized each position of the 103-nt long sequence to different degrees to identify aptamer mutants with an improved phenotype. Before randomization, we deleted an upstream start codon within 10A (ΔAUG) to prevent premature translation initiation. We started with 30.0% and 9.0% randomization and analyzed around 2000 clones. With 30.0% randomization, we completely lost any regulation, whereas 13 clones with improved phenotype could be identified within the pool with 9.0% randomization (Figure 3A). The detailed sequence analysis of these clones revealed a maximum of only up to four nucleotide exchanges. Based on this, we repeated the analysis with two new doped libraries with 4.5% and 3.0% randomization, respectively. We screened about 1000 clones, out of which about 100 clones fell into the gain-of-function (GOF) group. Sequencing 100 clones from both GOF and also from the loss-of-function (LOF) group revealed that two regions are nearly invariant (nts 26–40 and 63–103), whereas two regions can acquire mutations (nts 1–25 and 41–62). Furthermore, we identified seven mutation hot spots (Figure 3B). The hot spots were defined not only based on the overall mutation rate, but also on the fact that these point mutations were directed into one specific base. By combining all directed GOF mutations (G1U, A11C, A25C, U47C, C51U, A56C, U61G), we considerably improved regulation. GFP measurements revealed an increased *in vivo* activity of 7.5-fold (Figure 3C). For the resulting construct, we chose the term CFX riboswitch. Investigating the impact of each single mutation, U61G can be highlighted as one of the mutations with the highest contribution to the enhanced switching property (Figure 3D). However, U61G alone is not fully responsible for the enhanced phenotype.

Taken together, initial screening, subsequent partial randomization and the combination of beneficial mutations led to the new synthetic CFX riboswitch. With 7.5-fold regulatory activity, the dynamic window is comparable to other synthetic riboswitches, e.g. the tetracycline or neomycin riboswitch (22,33) and other RNA-based devices that control gene expression in eukaryotes (48,49).

Secondary structure analysis of the CFX riboswitch by structural probing and mutational analysis revealed a pseudoknot structure

Next, we endeavoured to gain insight into the secondary structure of the CFX riboswitch. The RNA was *in vitro* transcribed, radiolabeled and subjected to an in-line probing analysis (41). The cleavage pattern is shown in Figure 4A. Interestingly, the riboswitch consists to a large extent of non-cleaved nucleotides, implicating a high degree of structured regions. RNA folding prediction with programs based on the Zuker algorithm (50), however, did not result in any structure that fit to the observed probing data. On the other hand, the assumption of a pseudoknot fold resolved all mismatches to the in-line probing pattern and resulted in a secondary structure prediction illustrated in Figure 4B. To prove the assumed pseudoknot, we mutated it and analyzed respective rescue mutations (M2/M2R) (Figure 4B). In addition, we introduced a mutation and its respective rescue into the closing stem P1 (M1/M1R) (Figure 4B). For both regions, disrupted base paring completely diminished regulation. Functionality could be restored by introduction of compensatory mutations. Only two regions showed significant flexibility, the first 25 nucleotides (nt 1–25) and the L2 region (nucleotides 44–59, blue in Figure 4B). Interestingly, gain-of-function mutants that improved regulation were exclusively found in these two regions (position 1, 11 and 25, U61G removes a mismatch within P2, and U47C, C51U and A56C located within the L2 region). There is no indication that nts 1–25 were involved in the aptamer structure, the-gain of-function may be attributed to context dependencies. The L2 region, however, seems to be important for regulation since three gain-of-function mutants were located in this region. In contrast, no gain-of-function mutant was identified in the central part formed by the pseudoknot, P3 and P1. This region harboured the only two positions with significant CFX-dependent changes in the in-line probing pattern indicating a role of U37 and G72, respectively, for ligand binding. We determined a dissoci-

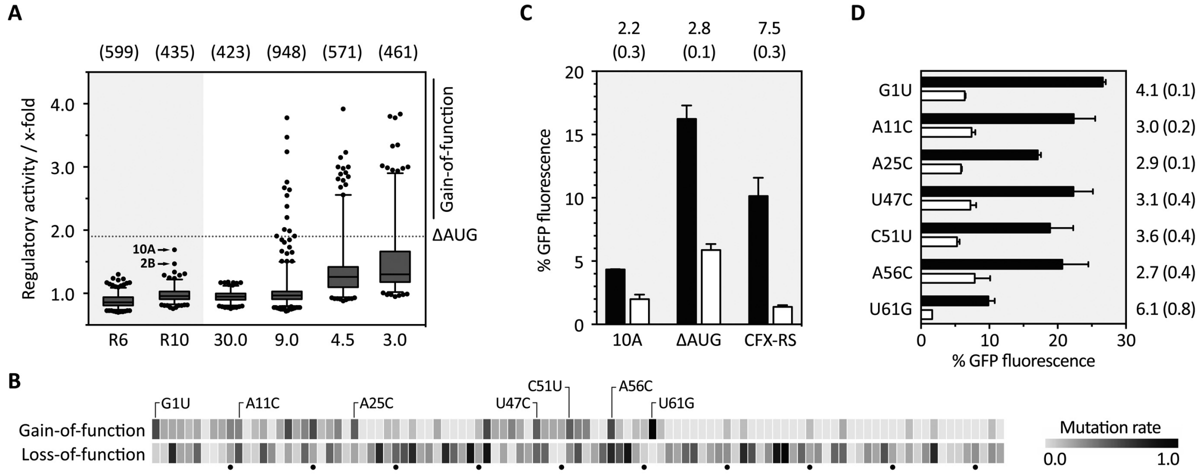


Figure 3. *In vivo* screening for CFX riboswitches and refinement of the aptamer domain. (**A**) The boxplot summarizes the *in vivo* screening in *S. cerevisiae* for round 6 (R6) and round 10 (R10) and the screened doped pools with different degrees of randomization, starting with 30.0% (30.0) down to 3.0% (3.0). For each investigated clone, the regulatory activity was calculated as ratio of GFP fluorescence with and without 1 mM CFX (x-fold). The two regulatory active sequences 2B and 10A are highlighted. Based on 10A, a mutant was derived with a mutated AUG that was upstream of the original start codon (ΔAUG) and used for synthesizing four doped libraries (30.0–3.0%). Clones that showed better switching properties than ΔAUG (dotted line) were sorted into the gain-of-function group. The numbers written above the boxplot indicate the number of clones used for the particular box. (**B**) Heatmap based on sequencing the gain- and loss-of-function group. For each group, mutation rate was normalized to 1.0 and plotted as a heatmap as per-nucleotide function. Highlighted point mutations in the gain-of-function group are considered as directed mutations. All single data can be found in Supplementary Table S8. (**C**) Fluorescence measurements of the originally found candidate 10A, the mutated one ΔAUG and the CFX riboswitch (CFX-RS). Shown are the fluorescence values without (black bars) and with 1 mM CFX (white bars). Above each construct, the regulatory activity is written with standard deviation (SD) in brackets. (**D**) Investigation of the impact of each single mutation. As in C, the fluorescence values without and with 1 mM CFX are plotted and the regulatory activity and SD in brackets is shown on the right, respectively.

ation constant for the CFX riboswitch of 60 nM by fluorescence titration spectroscopy and isothermal calorimetry (ITC) (Figure 4C and D). The analysis of the binding affinity of the mutants U37A and G72C, respectively, resulted in a dramatically reduced binding affinity. Simultaneously, both mutations lead to a complete loss of regulation *in vivo* (Figure 4E). We speculate that this region (the pseudoknot, P3 and the upper base pair of P1) may constitute the CFX binding pocket. Interestingly, the binding constant of the CFX riboswitch is similar to the initial candidates 10A and 10AΔAUG, although considerably enhanced in *in vivo* activity. It indicates that the improvement of regulation targeted the switching potential of the riboswitch rather than its ligand binding.

In sum, the CFX riboswitch presented here once more exemplifies the essential requirement for tight ligand binding, which is in line with previous work (3,22,33). Nucleotides involved in ligand binding were clustered around the pseudoknot fold, which supports the idea of the formation of a binding pocket that allows an initial binding of CFX to the aptamer. During a second binding step, the binding pocket is then closed through the P2/L2 region. We have demonstrated a similar two-step binding model for the tetracycline aptamer (51).

Ligand binding and recognition

We determined the binding affinities of seven different fluoroquinolones to further characterize the structure-function relationship of the CFX riboswitch. We analyzed ligand binding by fluorescence titration spectroscopy and determined their switching potential (Figure 5). The titration experiments revealed two side groups to be important for CFX binding: the carboxyl group on C3 and the fluorine group on C6. Here, decarboxy CFX (dCFX), pipedimic acid (PA) and 6-hydroxy-6-defluoro CFX (hCFX) showed a reduction in binding affinity of about 6- to 14-fold, respectively. Less relevant for binding is the cyclopropyl residue on N1, showing a 3-fold higher dissociation constant for norfloxacin (NFX) compared to CFX. Danofloxacin (DFX) and enrofloxacin (EFX), which have modifications on the piperazinyl residue, showed no significant change in affinity to the CFX riboswitch. This supports the assumption that immobilization of CFX for *in vitro* selection was most probably achieved by coupling the secondary amine to the activated epoxy group of the solid support (indicated by an arrow in Figure 1A on CFX).

The *in vivo* activity is roughly related to the binding affinity of each ligand to the riboswitch, with the exception of EFX caused by its toxicity (reduction up to one tenth in yeast growth [data not shown]) and hCFX. hCFX may be directly converted into an intermediate of the CFX degradation pathway which gives an explanation for the missing *in vivo* activity (52,53).

Scalable and programmable control of cellular survival

The CFX riboswitch is capable to regulate GFP expression up to 7.5-fold in a nearly binary fashion due to the low OFF state. Taking this into account, we aimed to prove the po-

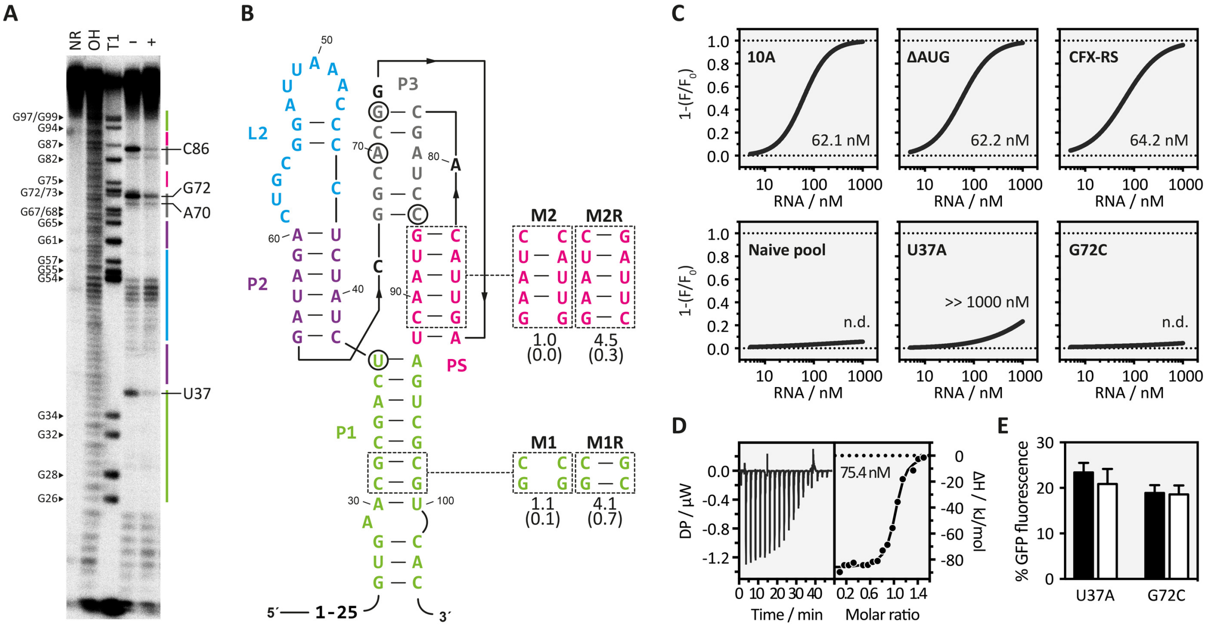


Figure 4. Structure determination of the CFX riboswitch. (**A**) In-line probing experiment for CFX riboswitch. Shown is the cleavage pattern in the absence (–) and presence (+) of 10 μM CFX under alkaline conditions. As references and for nucleotide position assignment, non-reacted RNA (NR), hydroxyl reaction (OH) and nuclease T1 digestion (T1) were loaded onto the gel. G nucleotides and nucleotides that showed a change upon ligand addition are highlighted. Colour coding for identified stem and loop regions follows the coding for the proposed secondary structure in B. (**B**) Proposed secondary structure of the CFX riboswitch including three stems (P1, P2, P3), a loop-region (L2) and a pseudoknot fold (PS). Nucleotides with changes in the probing pattern are encircled. Mutations introduced to study structure–function relationships are indicated. (**C**) Fluorescence titration experiment data for the indicated RNAs. (**D**) Verification the binding constant of CFX-RS with ITC. Left panel: power required to maintain the temperature of the RNA solution recorded over the time until saturation was reached (baseline-corrected). Right panel: integrated heats of interaction plotted against the molar ratio of ligand over RNA and fitted to a single binding site model (MicroCal PEAQ-ITC Analysis Software 1.1.0). (**E**) *In vivo* data for point-mutated nucleotides U37 to A and G72 to C.

tential of the CFX riboswitch to control survival by regulating genes necessary for cellular growth. For this purpose, we exchanged *gfp+* with either the *kanR* or the *URA3* gene. *KanR* codes for the aminoglycoside 3′-phosphotransferase that allows growth in the presence of the toxic compound geneticin (G418) (54). The *URA3* gene encodes the orotidine 5′-phosphate decarboxylase that allows growth on synthetic drop-out media without uracil (55).

Both genes were cloned under the control the CFX riboswitch and the precursor 10AΔAUG. In addition, the respective positive control without the insertion of the riboswitch was analyzed (*kanR+* or *URA3+*). As negative control, the start codon of the respective genes (*kanR-* or *URA3-*) was removed preventing gene expression and consequently cell growth.

Yeast strains with the respective plasmids were grown in the absence and in the presence of 1 mM CFX on appropriate plates and cell growth was analyzed by serial dilution growth assays. In the absence of CFX, no negative effect on cell growth could be detected upon riboswitch insertion controlling *kanR* (Figure 6A). This is interesting, since decreased expression level was observed in the GFP reporter gene assay compared to the control without riboswitch (Figure 3C). For controlling the *URA3* gene, we detected a slight reduction of cellular growth upon introduction of the CFX riboswitch (Figure 6B). In the presence of 1 mM CFX, expression of both genes could be significantly reduced. For *kanR*, growth was reduced over three orders of magnitude and for *URA3,* hardly any growth could be detected after two days on plate.

In sum, the growth of yeast carrying essential genes under the control of the CFX riboswitch can be effectively and quantitatively controlled through addition of CFX. Furthermore, the effect of aptamer insertion on gene expression in a physiological context was absent or negligible, although expression levels in reporter gene assays responded to aptamer insertion.

The CFX riboswitch controls gene expression in HeLa cells

To prove CFX riboswitch functionality not only in yeast, but also in higher eukaryotes we exploited a dual luciferase system that expresses firefly and *Renilla* luciferase in HeLa cells. We cloned the CFX riboswitch in front of the start codon of the firefly luciferase in the pDLP vector system (56), analogous to the GFP variants in yeast. Previous work on tetracycline and neomycin riboswitches carried out in our lab indicated that the transfer of a yeast-optimized variant to higher eukaryotes can be challenging (3). One conceivable explanation could be the increased helicase activity of the ribosome in higher eukaryotes compared to yeast

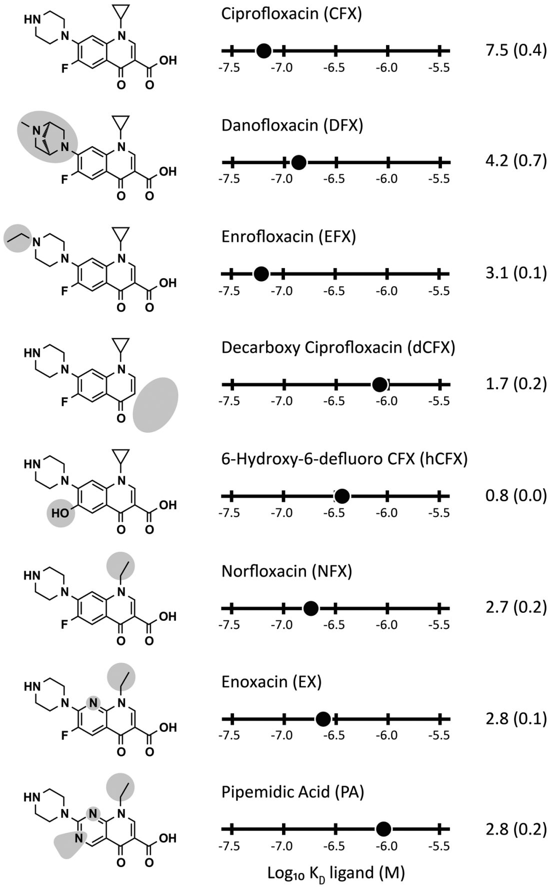


Figure 5. Specificity of molecular recognition of the CFX riboswitch and their impact on riboswitching. Left: Chemical structure of ciprofloxacin and seven selected fluoroquinolones that show binding to the CFX riboswitch. Chemical changes relative to CFX are shaded in light grey. Middle: Plot of $\log_{10} K_D$ values of the different fluoroquinolones and their activity *in vivo* (right), respectively (SD in brackets). Data shown in this graphical overview are summarized in Supplementary Table S9.

(57,58). Consequently, the underlying strategy was to stabilize the CFX riboswitch. Therefore, we exchanged loop L2 into a stable GAAA tetraloop (59). This stabilization of the riboswitch led to a nearly 2-fold regulation in HeLa cells (Figure 7). By application of the same partial randomization strategy outlined above, further improvement may be possible. Thus, we demonstrate for the first time to our knowledge that a riboswitch controlling translation initiation is portable between different species without major changes in sequence composition.

CONCLUSION

Riboswitches are associated with a range of advantages. These switches are often very simple as they consist of only one genetic element, RNA. As such, they are independent of transcription factors or other regulatory pro-

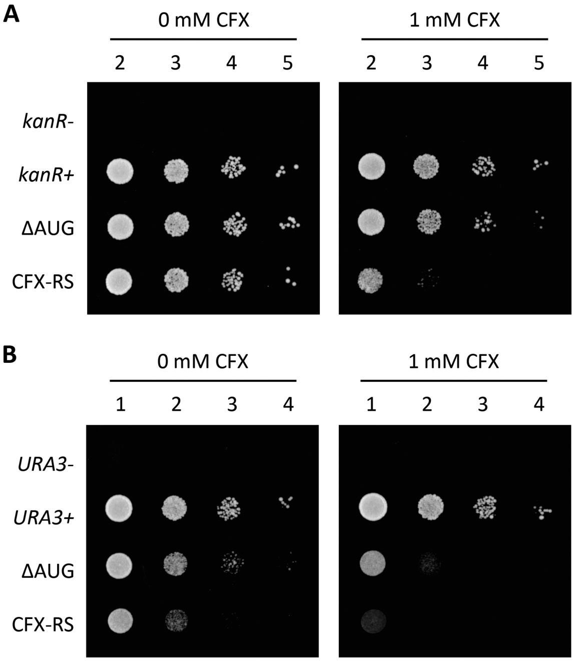


Figure 6. CFX riboswitch controls yeast growth. (**A**) Serial dilution growth assay were performed for both, 10AΔAUG (ΔAUG) and the engineered CFX riboswitch (CFX-RS). Additionally, two controls expressing the aminoglycoside 3′-phosphotransferase (*kanR+*) and a mutant without start codon ATG (*kanR-*) were spotted on SCD-ura plates (supplemented with G148) in the absence or presence of 1 mM CFX. Ten-fold serial dilutions were spotted from left to right (numbering above). Cells were grown for two days at 30°C. (**B**) Similar to the experiments with *kanR*, an analogous approach was performed by exchanging the *kanR* gene with *URA3*. Selection marker for plasmid maintenance were swapped.

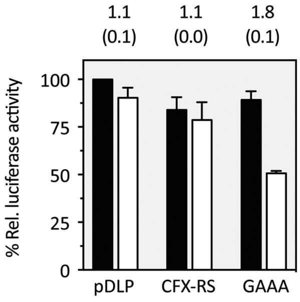


Figure 7. Application of the CFX riboswitch in a mammalian cell line. Dual luciferase assay of the CFX riboswitch (CFX-RS) and the stabilized GAAA mutant. The stabilized GAAA mutant, where the flexible L2 loop region was exchanged with a stable GAAA tetraloop, showed a significant reduction in *firefly* luciferase activity upon addition of 250 μM CFX. Black bars = w/o ligand, White bars = 250 μM CFX. pDLP is the vector without riboswitch. The experiments were repeated at least three times. SD are reported in brackets.

teins. Leakiness, a phenomenon often described for transcriptional regulation due to position effects, does not affect riboswitches. Moreover, RNA-based sensor domains may in theory be selected for any desired ligand. The riboswitch field is presently on the cusp of the transition from proof-of-concept studies to the development of robust and applicable tools and switches (13).

The main obstacle limiting functional synthetic riboswitch development may currently be found in the fact that only a fraction of the selected aptamers are suitable for riboswitch design. Fit-for-purpose aptamers require not only excellent binding properties, but also the conformational flexibility essential for a switch. Only the first can be addressed by SELEX; for the latter, however, laborious screening is necessary.

Through the process of *in vitro* selection, subsequent NGS-guided *in vivo* screening and further optimization, we could identify and engineer a novel, CFX-responsive riboswitch with a dynamic range sufficient to allow for complete control of cellular behavior. Moreover, we created a riboswitch that is active both in lower and higher eukaryotes. Due to the digital behavior of the ON and OFF state, we predict a high performance in biological circuits.

In essence, our findings present the first evidence for the *de novo* design of a riboswitch in the last 10 years. We are convinced that our optimized protocol will allow the discovery of a multitude of new riboswitches. Further advancement of our methodology, e.g. optimized selection procedures that combine selection for binding performance and conformational switching ability or the application of microfluidic high-throughput screening, is under way. These developments open up novel and unforeseen perspectives, e.g. for control of gene expression. However, riboswitches could also find application as biosensors, e.g. for monitoring of cell metabolite concentrations and optimization of metabolic pathways or measurements of metabolic flux rates. Moreover, riboswitches would be effective as highly sensitive sensors for detection of pollutants. Thus, this novel RNA device has the potential to energize and inspire efforts for synthetic riboswitch development.

In essence, synthetic riboswitches have received relatively little attention so far, and they may have been slightly overlooked due to the fact that they have yet to reach a critical mass. They are not entirely uncontroversial, mainly due to the low number of functional switches available to date. However, we take the opposite view and propose that the novel CFX riboswitch and the associated tools developed in our study will be a turning point for the synthetic riboswitch field. In all likelihood, ongoing work in our group and other laboratories will lead to a breakthrough and widespread use of riboswitches in the near future.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

FUNDING

Deutsche Forschungsgemeinschaft [SFB902/A2]; ERASynBio 'Intensify', LOEWE CompuGene and H2020 ITN MetaRNA. Funding for open access charge: Deutsche Forschungsgemeinschaft.

Conflict of interest statement. None declared.

REFERENCES

1. Green,A.A., Silver,P.A., Collins,J.J. and Yin,P. (2014) Toehold switches: de-novo-designed regulators of gene expression. *Cell*, **159**, 925–939.
2. Chappell,J., Takahashi,M.K. and Lucks,J.B. (2015) Creating small transcription activating RNAs. *Nat. Chem. Biol.*, **11**, 214–220.
3. Berens,C., Groher,F. and Suess,B. (2015) RNA aptamers as genetic control devices: the potential of riboswitches as synthetic elements for regulating gene expression. *Biotechnol. J.*, **10**, 246–257.
4. Win,M.N. and Smolke,C.D. (2008) Higher-order cellular information processing with synthetic RNA devices. *Science*, **322**, 456–460.
5. Felletti,M., Stifel,J., Wurmthaler,L.A., Geiger,S. and Hartig,J.S. (2016) Twister ribozymes as highly versatile expression platforms for artificial riboswitches. *Nat. Commun.*, **7**, 12834.
6. Wittmann,A. and Suess,B. (2011) Selection of tetracycline inducible self-cleaving ribozymes as synthetic devices for gene regulation in yeast. *Mol. Biosyst.*, **7**, 2419–2427.
7. Gammage,P.A., Gaude,E., Van Haute,L., Rebelo-Guiomar,P., Jackson,C.B., Rorbach,J., Pekalski,M.L., Robinson,A.J., Charpentier,M., Concordet,J.-P. *et al.* (2016) Near-complete elimination of mutant mtDNA by iterative or dynamic dose-controlled treatment with mtZFNs. *Nucleic Acids Res.*, **44**, 7804–7816.
8. Khalil,A.S. and Collins,J.J. (2010) Synthetic biology: applications come of age. *Nat. Rev. Genet.*, **11**, 367–379.
9. Ellington,A.D. and Szostak,J.W. (1990) In vitro selection of RNA molecules that bind specific ligands. *Nature*, **346**, 818–822.
10. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.
11. McKeague,M., McConnell,E.M., Cruz-Toledo,J., Bernard,E.D., Pach,A., Mastronardi,E., Zhang,X., Beking,M., Francis,T., Giamberardino,A. *et al.* (2015) Analysis of in vitro aptamer selection parameters. *J. Mol. Evol.*, **81**, 150–161.
12. McKeague,M., Wong,R.S. and Smolke,C.D. (2016) Opportunities in the design and application of RNA for gene expression control. *Nucleic Acids Res.*, **44**, 2987–2999.
13. Berens,C. and Suess,B. (2015) Riboswitch engineering—making the all-important second and third steps. *Curr. Opin. Biotechnol.*, **31**, 10–15.
14. Wachsmuth,M., Findeiß,S., Weissheimer,N., Stadler,P.F. and Mörl,M. (2013) De novo design of a synthetic riboswitch that regulates transcription termination. *Nucleic Acids Res.*, **41**, 2541–2551.
15. Lynch,S.A., Desai,S.K., Sajja,H.K. and Gallivan,J.P. (2007) A high-throughput screen for synthetic riboswitches reveals mechanistic insights into their function. *Chem. Biol.*, **14**, 173–184.
16. Weigand,J.E. and Suess,B. (2007) Tetracycline aptamer-controlled regulation of pre-mRNA splicing in yeast. *Nucleic Acids Res.*, **35**, 4179–4185.
17. Beilstein,K., Wittmann,A., Grez,M. and Suess,B. (2015) Conditional control of mammalian gene expression by tetracycline-dependent hammerhead ribozymes. *ACS Synth. Biol.*, **4**, 526–534.
18. Breaker,R.R. (2012) *Riboswitches and the RNA World*. CSHL Press.
19. Lynch,S.A. and Gallivan,J.P. (2009) A flow cytometry-based screen for synthetic riboswitches. *Nucleic Acids Res.*, **37**, 184–192.
20. Townshend,B., Kennedy,A.B., Xiang,J.S. and Smolke,C.D. (2015) High-throughput cellular RNA device engineering. *Nat. Methods*, **12**, 989–994.
21. Schneider,C. and Suess,B. (2016) Identification of RNA aptamers with riboswitching properties. *Methods*, **97**, 44–50.
22. Weigand,J.E., Sanchez,M., Gunnesch,E.-B., Zeiher,S., Schroeder,R. and Suess,B. (2008) Screening for engineered neomycin riboswitches that control translation initiation. *RNA*, **14**, 89–97.
23. Levy,S.E. and Myers,R.M. (2016) Advancements in next-generation sequencing. *Annu. Rev. Genom. Hum. Genet.*, **17**, 95–115.
24. Ronald,A.R. and Low,D. (2003) *Fluoroquinolone Antibiotics*. Springer Science & Business Media.
25. Suto,M.J., Domagala,J.M., Roland,G.E., Mailloux,G.B. and Cohen,M.A. (1992) Fluoroquinolones: relationships between

structural variations, mammalian cell cytotoxicity, and antimicrobial activity. *J. Med. Chem.*, **35**, 4745–4750.

26. Azéma,J., Guidetti,B., Dewelle,J., Le Calve,B., Mijatovic,T., Korolyov,A., Vaysse,J., Malet-Martino,M., Martino,R. and Kiss,R. (2009) 7-((4-Substituted)piperazin-1-yl) derivatives of ciprofloxacin: synthesis and in vitro biological evaluation as potential antitumor agents. *Bioorg. Med. Chem.*, **17**, 5396–5407.
27. Yang,Q., Nakkula,R.J. and Walters,J.D. (2002) Accumulation of ciprofloxacin and minocycline by cultured human gingival fibroblasts. *J. Dent. Res.*, **81**, 836–840.
28. Paige,J.S., Wu,K.Y. and Jaffrey,S.R. (2011) RNA mimics of green fluorescent protein. *Science*, **333**, 642–646.
29. Hall,B., Micheletti,J.M., Satya,P., Ogle,K., Pollard,J. and Ellington,A.D. (2009) Design, synthesis, and amplification of DNA pools for in vitro selection. *Curr. Protoc. Mol. Biol.*, ***88***, 24.2.1–24.2.27.
30. Vogel,M. and Suess,B. (2016) Label-free determination of the dissociation constant of small molecule-aptamer interaction by isothermal titration calorimetry. *Methods Mol. Biol.*, **1380**, 113–125.
31. Groher,F. and Suess,B. (2016) In vitro selection of antibiotic-binding aptamers. *Methods*, **106**, 42–50.
32. Kötter,P., Weigand,J.E., Meyer,B., Entian,K.-D. and Suess,B. (2009) A fast and efficient translational control system for conditional expression of yeast genes. *Nucleic Acids Res.*, **37**, e120.
33. Suess,B., Hanson,S., Berens,C., Fink,B., Schroeder,R. and Hillen,W. (2003) Conditional gene expression by controlling translation with tetracycline-binding aptamers. *Nucleic Acids Res.*, **31**, 1853–1858.
34. Suess,B. and Weigand,J.E. (2009) Aptamers as artificial gene regulation elements. *Methods Mol. Biol.*, **535**, 201–208.
35. Levenshtein,V.I. (1966) Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.*, **10**, 707.
36. Lorenz,R., Bernhart,S.H., Höner Zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
37. Andronescu,M., Condon,A., Hoos,H.H., Mathews,D.H. and Murphy,K.P. (2010) Computational approaches for RNA energy parameter estimation. *RNA*, **16**, 2304–2318.
38. Buß,O., Jager,S., Dold,S.-M., Zimmermann,S., Hamacher,K., Schmitz,K. and Rudat,J. (2016) Statistical evaluation of HTS assays for enzymatic hydrolysis of β-keto esters. *PLoS ONE*, **11**, e0146104.
39. Rio,D.C. (2011) *RNA: A Laboratory Manual*. CSH Press.
40. Seetharaman,S., Zivarts,M., Sudarsan,N. and Breaker,R.R. (2001) Immobilized RNA switches for the analysis of complex chemical and biological mixtures. *Nat. Biotechnol*, **19**, 336–341.
41. Regulski,E.E. and Breaker,R.R. (2008) In-line probing analysis of riboswitches. *Methods Mol. Biol.*, **419**, 53–67.
42. Pascault,J.-P. and Williams,R.J.J. (2009) *Epoxy Polymers*. John Wiley & Sons.
43. Davis,J.H. and Szostak,J.W. (2002) Isolation of high-affinity GTP aptamers from partially structured RNA libraries. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 11616–11621.
44. Nutiu,R. and Li,Y. (2005) In vitro selection of structure-switching signaling aptamers. *Angew. Chem. Int. Ed. Engl.*, **44**, 1061–1065.
45. Uhlenbeck,O.C. (1990) Tetraloops and RNA folding. *Nature*, **346**, 613–614.
46. Hunsicker,A., Steber,M., Mayer,G., Meitert,J., Klotzsche,M., Blind,M., Hillen,W., Berens,C. and Suess,B. (2009) An RNA aptamer that induces transcription. *Chem. Biol.*, **16**, 173–180.
47. Nguyen Quang,N., Perret,G. and Ducongé,F. (2016) Applications of high-throughput sequencing for in vitro selection and characterization of aptamers. *Pharmaceuticals (Basel)*, **9**, 76.
48. Culler,S.J., Hoff,K.G. and Smolke,C.D. (2010) Reprogramming cellular behavior with RNA controllers responsive to endogenous proteins. *Science*, **330**, 1251–1255.
49. Wieland,M. and Fussenegger,M. (2010) Ligand-dependent regulatory RNA parts for synthetic biology in eukaryotes. *Curr. Opin. Biotechnol.*, **21**, 760–765.
50. Zuker,M. and Stiegler,P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
51. Förster,U., Weigand,J.E., Trojanowski,P., Suess,B. and Wachtveitl,J. (2012) Conformational dynamics of the tetracycline-binding aptamer. *Nucleic Acids Res.*, **40**, 1807–1817.
52. Singh,A. and Ward,O.P. (2014) *Biodegradation and Bioremediation XVII*. Springer Science & Business Media.
53. Liao,X., Li,B., Zou,R., Dai,Y., Xie,S. and Yuan,B. (2016) Biodegradation of antibiotic ciprofloxacin: pathways, influential factors, and bacterial community structure. *Environ. Sci. Pollut. Res.*, **23**, 7911–7918.
54. Webster,T.D. and Dickson,R.C. (1983) Direct selection of Saccharomyces cerevisiae resistant to the antibiotic G418 following transformation with a DNA vector carrying the kanamycin-resistance gene of Tn903. *Gene*, **26**, 243–252.
55. Lacroute,F. (1968) Regulation of pyrimidine biosynthesis in Saccharomyces cerevisiae. *J. Bacteriol.*, **95**, 824–832.
56. Kemmerer,K. and Weigand,J.E. (2014) Hypoxia reduces MAX expression in endothelial cells by unproductive splicing. *FEBS Lett.*, **588**, 4784–4790.
57. McCarthy,J.E. (1998) Posttranscriptional control of gene expression in yeast. *Microbiol. Mol. Biol. Rev.*, **62**, 1492–1553.
58. Babendure,J.R., Babendure,J.L., Ding,J.-H. and Tsien,R.Y. (2006) Control of mammalian translation by mRNA structure near caps. *RNA*, **12**, 851–861.
59. Bottaro,S. and Lindorff-Larsen,K. (2017) Mapping the universe of RNA tetraloop folds. *Biophys. J.*, **113**, 257–267.

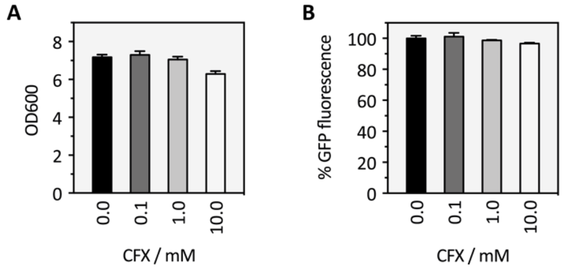
Supplementary material

Riboswitching with ciprofloxacin – Development and characterization of a novel RNA regulator

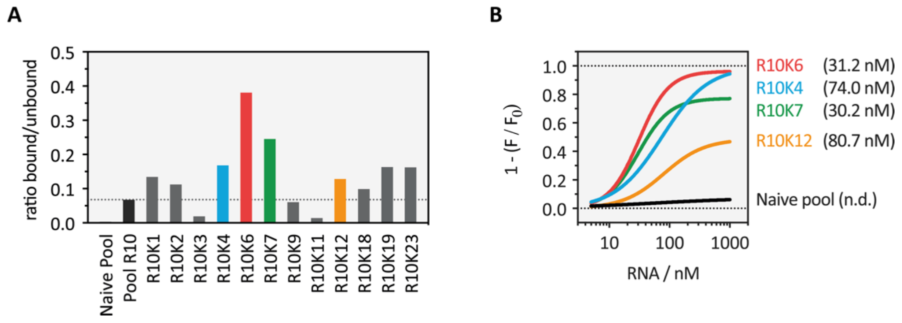
Florian Groher [1], Cristina Bofill-Bosch [1], Christopher Schneider [1], Johannes Braun [1], Sven Jager [2], Katharina Geißler [1], Kay Hamacher [2,3], Beatrix Suess [1,*]

[1] Synthetic Genetic Circuits, Dept. of Biology, TU Darmstadt, Darmstadt, Germany
[2] Computational Biology and Simulation, Dept. of Biology, TU Darmstadt, Darmstadt, Germany
[3] Dept. of Physics, Dept. of Computer Science, TU Darmstadt, Darmstadt, Germany

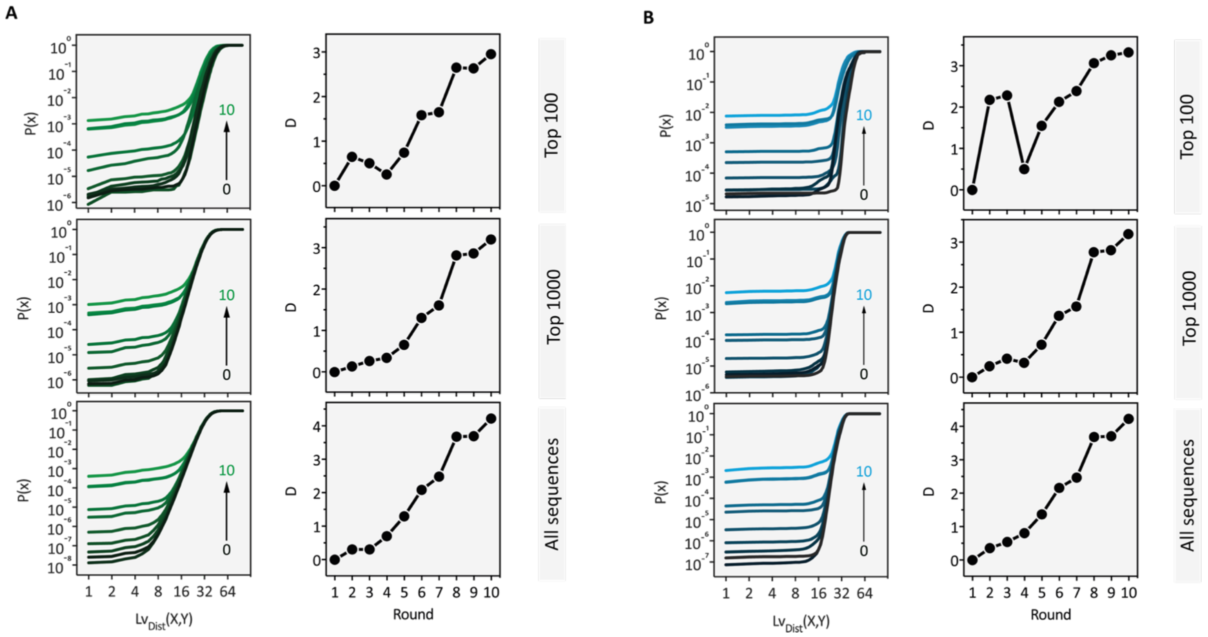
* To whom correspondence should be addressed. Tel: +49 6151 1622000; Fax: +49 6151 1622003; Email: bsuess@bio.tu-darmstadt.de



Supplementary Figure S1. Influence of CFX on yeast growth and GFP expression. A OD600 of yeast cultures grown overnight in media supplemented with the respective CFX concentration. **B** Relative GFP fluorescence of yeast cultures supplemented with the respective CFX concentration compared to untreated cells. Measurements were repeated three times with technical replicates.



Supplementary Figure S2. Analysis of single clone binding from round 10. A Ratios of bound vs. unbound RNA for different clones from selection round 10 are displayed. As references, the naive pool and the pool from round 10 are depicted. According to the SELEX procedure, RNA was transcribed and 500 kcpm were loaded onto the CFX-derivatized column. After 10 wash steps with 1 CV binding buffer each, the RNA was eluted by 4 wash steps with 1 mM CFX in solution. Afterwards each fraction was measured on a scintillation counter. Measured radioactivity in the fractions flow through and wash steps were summed up (unbound) and also for elution fractions (bound). The ratio of bound to unbound gives a direct qualitative feedback of the binding capacity of the tested clones. **B** Determination of binding affinity of the selected aptamer candidates by fluorescence titration spectroscopy. Measurements were repeated at least twice. Standard deviations and individual data points were omitted for clarity. K_D values are written in brackets.



Supplementary Figure S3. Next generation sequencing analysis Displayed are the cumulative distribution function (CDF) and Kolmogorv Smirnoff's ks test (D) for Top100, Top1000 and all sequences. **A** Results based on calculated minimal free energy (MFE) secondary structure. **B** Results based on sequence. The CDF for each round based on calculated Levenshtein distances on MFE structures is plotted for each round (left in A and B), resulting in an increased P(x) over the selection experiment. Based on CDF, D was derived and its logarithm is plotted against the selection rounds for Top100, Top1000 and all sequences (right panel in A and B**)**. Here, D is computed between the first round and all remaining.

One mayor drawback is the computational time that it takes to compute a Lv_{Dist} (X,Y) distribution where we compare every sequence with every other (often 10^12 single computations). Due to this, advanced computational resources as well as efficient software and memory management is required. However, the data suggests that calculating all levenshtein distances for each sequence and each round is not necessary and it is sufficient to look at the Top1000 enriched sequences to draw conclusions (at least in this SELEX experiment). This fact reduced the calculation efforts required by several orders of magnitude. We can conclude that comparing Top1000 vs all sequences by its levenshtein distance can improve the process of SELEX round selection for future work. Additionally, using only the Top1000 made the computation feasible on a desktop computer by reducing the computational time by several orders of magnitude.

Supplementary Table S1. Plasmids used in this study

Name	Description	Reference
pWHE601	2µ plasmid with constitutively expression of *gfp*+ from an adh promoter	(1)
pWHE601*	Derived from pWHE601 with deletion of AUG in gfp+ / AflII --> AgeI	(2)
10A	Active riboswitch found in initial *in vivo* screening	This work
ΔAUG	Deletion of AUG within the sequence of 10A	This work
GOF	Introduction of 7 point mutations in ΔAUG	This work
G1U	Investigation of the named point mutation within ΔAUG	This work
A11C	Investigation of the named point mutation within ΔAUG	This work
A25C	Investigation of the named point mutation within ΔAUG	This work
U47C	Investigation of the named point mutation within ΔAUG	This work
C51U	Investigation of the named point mutation within ΔAUG	This work
A56C	Investigation of the named point mutation within ΔAUG	This work
U61G	Investigation of the named point mutation within ΔAUG	This work
A35G	Investigation of the named point mutation within ΔAUG	This work
U41G	Investigation of the named point mutation within ΔAUG	This work
A50G	Investigation of the named point mutation within ΔAUG	This work
U92G	Investigation of the named point mutation within ΔAUG	This work
A102G	Investigation of the named point mutation within ΔAUG	This work
M1	Mutation of the C31 and G32 to G and C within GOF, respectively	This work
M1R	Compensatory point mutations for M1 to restore function	This work
U37A	Investigation of the named point mutation within GOF	This work
G72C	Investigation of the named point mutation within GOF	This work
M2	Mutation of GUU75 to CAA within GOF	This work
M2R	Compensatory mutations for M2 to restore pseudoknot and function	This work
M3	Mutation of G75C and C79G within GOF	This work
M3R	Compensatory mutations for M3 to restore pseudoknot and function	This work
COMP	Complementary sequence of GOF for investigation of basal expression	This work

Corresponding oligonucleotides for cloning are listed in Supplementary Table S2.

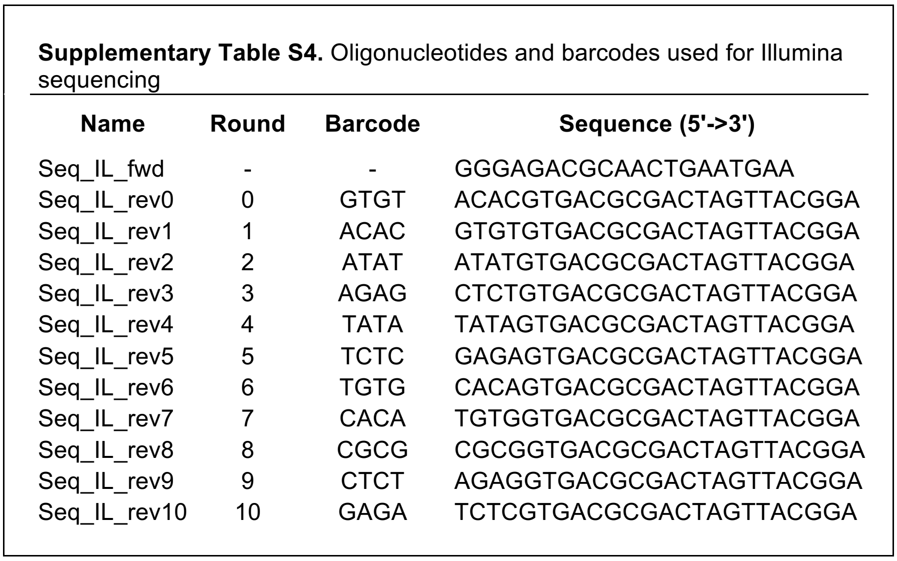
Supplementary Table S2. Oligonucleotides used for cloning

Name	Sequence (5'->3')
10A_fwd	CGCGACCGGTGGGAGACGCAACTGAATGAACATAAGTGAACGCGACTCTATCTCCCTAAACTAGGAGTCATATAGCGGCAC
10A_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATATGACTCCTAGTTTAGGGAGATAGAG
∆AUG_fwd	CGCGACCGGTGGGAGACGCAACTGAATCAACATAAGTGAACGCGACTCTATCTCCCTAAACTAGGAGTCATATAGCGGCAC
∆AUG_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATATGACTCCTAGTTTAGGGAGATAGAG
GOF_fwd	CGCGACCGGTTGGAGACGCACCTGAATCAACATACGTGAACGCGACTCTATCTCCCCAAATTAGGCGTCAGATAGCGGCACG
GOF_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATCTGACGCCTAATTTGGGGAG
G1U_fwd	CGCGACCGGTTGGAGACGCAACTGAATCAACATAAGTGAACGCGACTCTATCTCCCTAAACTAGGAGTCATATAGCGGCAC
G1U_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATATGACTCCTAGTTTAGGGAGATAGAG
A11C_fwd	CGCGACCGGTTGGAGACGCACCTGAATCAACATAAGTGAACGCGACTCTATCTCCCTAAACTAGGAGTCATATAGCGGC
A11C_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATATGACTCCTAGTTTAGGGAGATAG
A25C_fwd	CGCGACCGGTTGGAGACGCAACTGAATCAACATACGTGAACGCGACTCTATCTCCCTAAACTAGGAGTCATATAGCGGC
A25C_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATATGACTCCTAGTTTAGGGAGATAG
U47C_fwd	CGCGACCGGTTGGAGACGCAACTGAATCAACATAAGTGAACGCGACTCTATCTCCCCAAACTAGGAGTCATATAGCGGC
U47C_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATATGACTCCTAGTTTGGGGAGATAG
C51U_fwd	CGCGACCGGTTGGAGACGCAACTGAATCAACATAAGTGAACGCGACTCTATCTCCCTAAATTAGGAGTCATATAGCGGC
C51U_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATATGACTCCTAATTTAGGGAGATAG
A56C_fwd	CGCGACCGGTTGGAGACGCAACTGAATCAACATAAGTGAACGCGACTCTATCTCCCTAAACTAGGCGTCATATAGCGGC
A56C_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATATGACGCCTAGTTTAGGGAGATAG
U61G_fwd	CGCGACCGGTTGGAGACGCAACTGAATCAACATAAGTGAACGCGACTCTATCTCCCTAAACTAGGAGTCAGATAGCGGC
U61G_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATCTGACTCCTAGTTTAGGGAGATAG
A35G_fwd	CGCGACCGGTTGGAGACGCAACTGAATCAACATAAGTGAACGCGGCTCTATCTCCCTAAACTAGGAGTCATATAGCGGCAC
A35G_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATATGACTCCTAGTTTAGGGAGATAGAG
U41G_fwd	CGCGACCGGTTGGAGACGCAACTGAATCAACATAAGTGAACGCGACTCTAGCTCCCTAAACTAGGAGTCATATAGCGGCAC
U41G_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATATGACTCCTAGTTTAGGGAGCTAGAG
A50G_fwd	CGCGACCGGTTGGAGACGCAACTGAATCAACATAAGTGAACGCGACTCTATCTCCCTAAGCTAGGAGTCATATAGCGGCAC
A50G_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATATGACTCCTAGCTTAGGGAGATAGAG
U92G_fwd	CGCGACCGGTTGGAGACGCAACTGAATCAACATAAGTGAACGCGACTCTATCTCCCTAAACTAGGAGTCATATAGCGGCAC
U92G_rev	GGCCGCTAGCCATTTTGTGACGCGACTCGTTACGGATCGTGTAACTCCGTGCCGCTATATGACTCCTAGTTTAGGGAGATAGAG
A102G_fwd	CGCGACCGGTTGGAGACGCAACTGAATCAACATAAGTGAACGCGACTCTATCTCCCTAAACTAGGAGTCATATAGCGGCAC
A102G_rev	GGCCGCTAGCCATTTTGCGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATATGACTCCTAGTTTAGGGAGATAGAG
M1_fwd	CGCGACCGGTGGGAGACGCAACTGAATCAACATAAGTGAAGCCGACTCTATCTCCCTAAACTAGGAGTCATATAGCGGC
M1_frev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATATGACTCCTAGTTTAGGG
M1R_fwd	- identical to M1_fwd -
M1R_rev	GGCCGCTAGCCATTTTGTGAGCCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATATGACTCCTAGTTTAGGGAGATAGAGTCG
U37A_fwd	CGCGACCGGTTGGAGACGCACCTGAATCAACATACGTGAACGCGACACTATCTCCCCAAATTAGGCGTCAGATAGCGGC
U37A_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATCTGACGCCTAATTTGGGGAGATAGTG
G72C_fwd	CGCGACCGGTTGGAGACGCACCTGAATCAACATACGTGAACGCGACTCTATCTCCCCAAATTAGGCGTCAGATAGCGGCACCG
G72C_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTAACTCGGTGCCGCTATCTGACGCCTAATTTGGGGAGATAGAGTC
M2_fwd	CGCGACCGGTTGGAGACGCACCTGAATCAACATACGTGAACGCGACTCTATCTCCCCAAATTAGGCGTCAGATAGCGGC
M2_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTGTTTGTCCGTGCCGCTATCTGACGCCTAATTTGGGGAGATAG
M2R_fwd	CGCGACCGGTTGGAGACGCACCTGAATCAACATACGTGAACGCGACTCTATCTCCCCAAATTAGGCGTCAGATAGCGGCACGG
M2R_rev	GGCCGCTAGCCATTTTGTGACGCGACTACAAACGGATCGTGTTTGTCCGTGCCGCTATCTGACGCCTAATTTGGGGAG
M3_fwd	CGCGACCGGTTGGAGACGCACCTGAATCAACATACGTGAACGCGACTCTATCTCCCCAAATTAGGCGTCAGATAGCGGC
M3_rev	GGCCGCTAGCCATTTTGTGACGCGACTAGTTACGGATCGTCTAAGTCCGTGCCGCTATCTGACGCCTAATTTGGGGAGATAG
M3R_fwd	CGCGACCGGTTGGAGACGCACCTGAATCAACATACGTGAACGCGACTCTATCTCCCCAAATTAGGCGTCAGATAGCGGCACG
M3R_rev	GGCCGCTAGCCATTTTGTGACGCGACTACTTAGGGATCGTCTAAGTCCGTGCCGCTATCTGACGCCTAATTTGGGGAGATAG
COMP_fwd	CGCGACCGGTACCTCTGCGTGGACTTAGTTGTATGCACTTGCGCTGAGATAGAGGGGTTTAATCCGCAGTCTATCGCCGTG
COMP_rev	GGCCGCTAGCCATTTTCACTGCGCTGATCAATGCCTAGCACATTGAGGCACGGCGATAGACTGCGGATTAAACCCCTCTATCTC

Supplementary Table S3. Oligonucleotides used for cloning of doped pools for *in vivo* screening

Name	**Sequence (5'->3')**
AgeI_doped_fwd	GCATACAATCAACTCCAAGCTAGATCTACCGGT
NheI_[3.0/4.5/9.0/30.0]_doped_rev	CGAGCTAGCCATTTT**[**GTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATATGACTCCTAGTTTAGGGAGATAGAGTCGCGTTCACTTATGTTGATTCAGTTGCGTCTCCC**]**ACCGGTAGATCTAGCTTGGAGTTGATTGTATGC

For all cloning steps AgeI_doped_fwd was used for PCR. For generating different degrees of randomization, the part in brackets of NheI_ATG_Kozac_doped_rev was synthesized with mixed phosphoramidites for 3.0%, 4.5%, 9.0% and 30.0% incorporation of the other three bases.

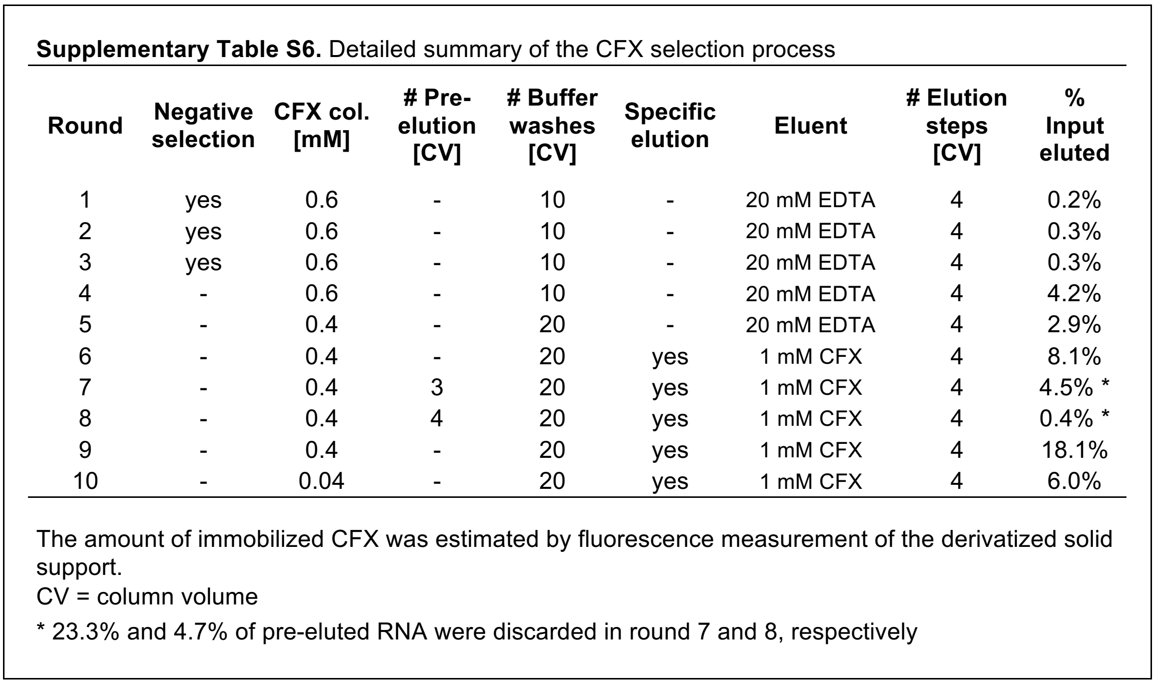


Supplementary Table S4. Oligonucleotides and barcodes used for Illumina sequencing

Name	Round	Barcode	Sequence (5'->3')
Seq_IL_fwd	-	-	GGGAGACGCAACTGAATGAA
Seq_IL_rev0	0	GTGT	ACACGTGACGCGACTAGTTACGGA
Seq_IL_rev1	1	ACAC	GTGTGTGACGCGACTAGTTACGGA
Seq_IL_rev2	2	ATAT	ATATGTGACGCGACTAGTTACGGA
Seq_IL_rev3	3	AGAG	CTCTGTGACGCGACTAGTTACGGA
Seq_IL_rev4	4	TATA	TATAGTGACGCGACTAGTTACGGA
Seq_IL_rev5	5	TCTC	GAGAGTGACGCGACTAGTTACGGA
Seq_IL_rev6	6	TGTG	CACAGTGACGCGACTAGTTACGGA
Seq_IL_rev7	7	CACA	TGTGGTGACGCGACTAGTTACGGA
Seq_IL_rev8	8	CGCG	CGCGGTGACGCGACTAGTTACGGA
Seq_IL_rev9	9	CTCT	AGAGGTGACGCGACTAGTTACGGA
Seq_IL_rev10	10	GAGA	TCTCGTGACGCGACTAGTTACGGA

Supplementary Table S5. Oligonucleotides for template generation for *in vitro* transcription

Name	Sequence (5'->3')
10A_T7_fwd	CCAAGTAATACGACTCACTATAGGGAGACGCAACTGAATGAACATAAGTGAACGCGACTCTATCTCCCTAAACTAGG
10A_T7_rev	GTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATATGACTCCTAGTTTAGGGAGATAGAGTCGCGTTC
ΔAUG_T7_fwd	CCAAGTAATACGACTCACTATAGGGAGACGCAACTGAATCAACATAAGTGAACGCGACTCTATCTCCCTAAACTAGG
ΔAUG_T7_rev	*- identical to 10A_T7_rev -*
GOF_T7_fwd	CCAAGTAATACGACTCACTATAGGGAGACGCACCTGAATCAACATACGTGAACGCGACTCTATCTCCCCAAATTAGGCGTCAG
GOF_T7_rev	GTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATCTGACGCCTAATTTGGGGAGATAGAGTCGCGTTCACG
U37A_T7_fwd	CCAAGTAATACGACTCACTATAGGGAGACGCACCTGAATCAACATACGTGAACGCGACACTATCTCCCCAAATTAGGCG
U37A_T7_rev	GTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATCTGACGCCTAATTTGGGGAGATAGTGTCGCGTTCACG
G72C_T7_fwd	CCAAGTAATACGACTCACTATAGGGAGACGCACCTGAATCAACATACGTGAACGCGACTCTATCTCCCCAAATTAGGCG
G72C_T7_rev	GTGACGCGACTAGTTACGGATCGTGTAACTCGGTGCCGCTATCTGACGCCTAATTTGGGGAGATAGAGTCGCGTTCACG
GOF_CAA4_T7_fwd	*- identical to GOF_T7_fwd -*
GOF_CAA4_T7_rev	TTGTTGTTGTTGTGACGCGACTAGTTACGGATCGTGTAACTCCGTGCCGCTATCTGACGCCTAATTTGGGGAGATAGAGTCGCGTTCACG



Supplementary Table S6. Detailed summary of the CFX selection process

Round	Negative selection	CFX col. [mM]	# Pre-elution [CV]	# Buffer washes [CV]	Specific elution	Eluent	# Elution steps [CV]	% Input eluted
1	yes	0.6	-	10	-	20 mM EDTA	4	0.2%
2	yes	0.6	-	10	-	20 mM EDTA	4	0.3%
3	yes	0.6	-	10	-	20 mM EDTA	4	0.3%
4	-	0.6	-	10	-	20 mM EDTA	4	4.2%
5	-	0.4	-	20	-	20 mM EDTA	4	2.9%
6	-	0.4	-	20	yes	1 mM CFX	4	8.1%
7	-	0.4	3	20	yes	1 mM CFX	4	4.5% *
8	-	0.4	4	20	yes	1 mM CFX	4	0.4% *
9	-	0.4	-	20	yes	1 mM CFX	4	18.1%
10	-	0.04	-	20	yes	1 mM CFX	4	6.0%

The amount of immobilized CFX was estimated by fluorescence measurement of the derivatized solid support.

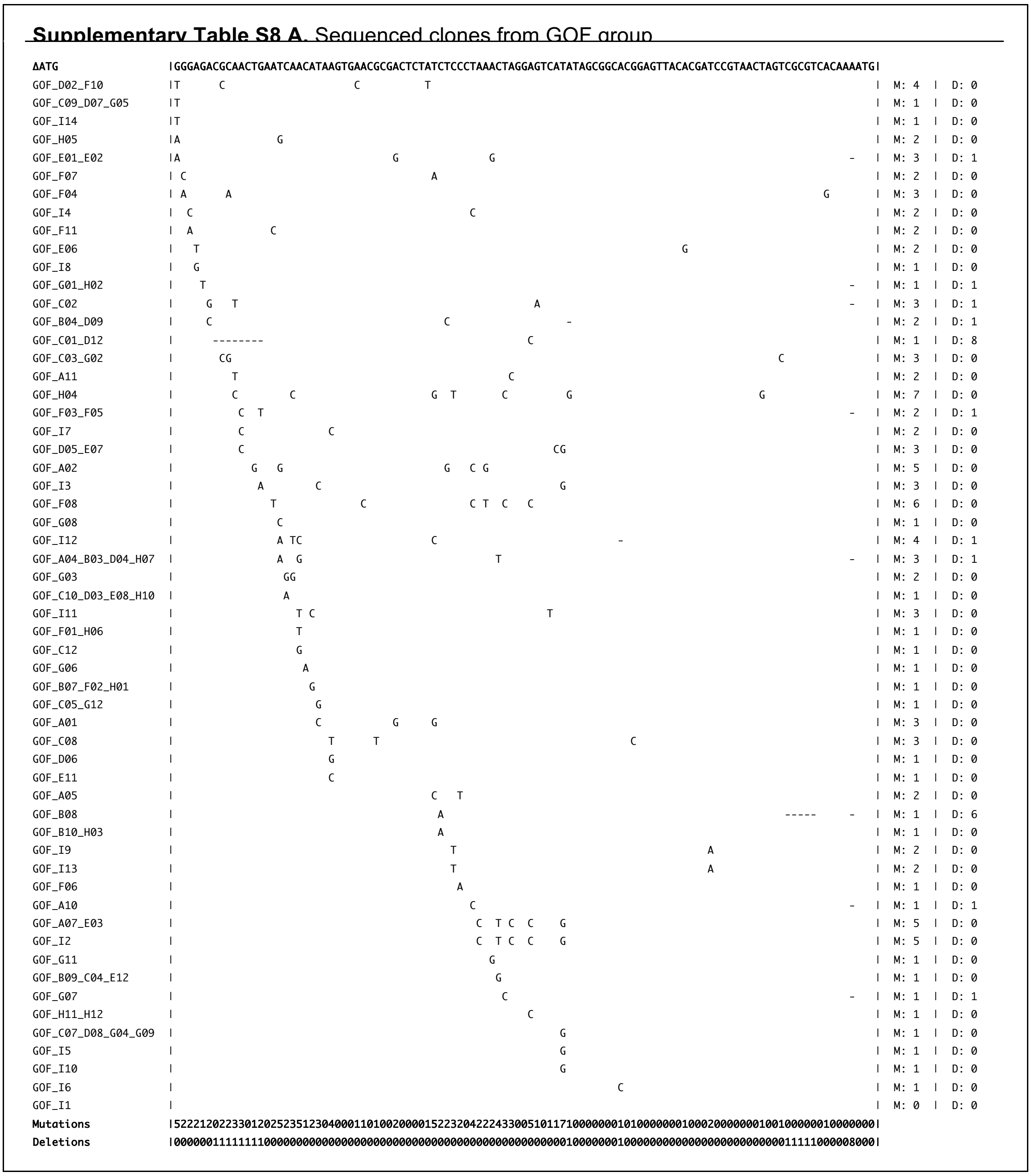
CV = column volume

* 23.3% and 4.7% of pre-eluted RNA were discarded in round 7 and 8, respectively

Supplementary Table S7. Randomized regions from clones round 10

Clone	Frequency	Sequence (5'->3')	Length*
R10K1	1	TCAGTGGCATTTCAAACACCAATTTGACGAAAAGAAGACTTAGTGAATACTAAGCGGAATTAAC	104
R10K2	3	AACCAAACAGTTCCATCAAGACCTAGGTATCTAGAAACTAGCACGTCCGGATATGTCGGTA	101
R10K3	2	ATCAGCATCCCTACAGAGGAAGTACCGCACACTATTGTGGAAAGGCCAGATTC	93
R10K4	5	GAGGTTCCCTATCATTCACAGACGCTGCTTCGGCAGTAACTAGAATGTCCGGCCACTACGTG	102
R10K6	4	AATGTCATTCAAGACTAGGTTGTGACTGCTTAGGCAGTTGTGGACGGCTAAGCCCACCAGAGG	103
R10K7	1	TTGATTTCCCGTGATGAAAAGAAGACTGCTTCGGCAGCGGAAGGAAAGTTTTCGGACCCTCCA	103
R10K9	1	TGCTGAGGACATTAGTAGCAAGTTCTCTGCTTCGGCAGGCAAATTTGGCAAGTCAGCT	98
R10K11	1	CGCAATTCATTTTCACTAGGTCGTGCTTGAAAAAGTGTTGGAGCCAGACTAATTAGCATCAGGG	104
R10K12	1	GTAGGTTCCCTATCATTCACAGACGCTGCTTCGGCAGTAACTAGAATGTCCGGCCACTACGTG	103
R10K13	1	GAGGTTCCCTATCATTCACAGACGCTGCTTCGGCGGTAACTAGAATGTCCGGCCACTACGTG	102
R10K18	1	CGTGGCCGAGCATACATCGTATCGGCCTGCTTCGACCAGGTCGGCCCTGGCG	92
R10K19	1	GACCGTCATTCATGAGTTCTTACGTGCTGCTTCGGCAGGGGGAGAATGGCTCGGACTTAAATGG	104
R10K23	1	CGAACTTCAACTAAACACTCCGATGTAATAACTAGCATCGTAGCCTGTCCCTGCGATAAAGGAG	104

Sequences found in SELEX round 10. Both, 5'- and 3'-regions are removed for clarity.
The reported stem loop (5'-CTGCTTCGGCAG-3') is underlined allowing for one mismatch/mutation.
* including constant regions.



Supplementary Table S8 A. Sequenced clones from GOF group

```
ΔATG                 |GGGAGACGCAACTGAATCAACATAAGTGAACGCGACTCTATCTCCCTAAACTAGGAGTCATATAGCGGCACGGAGTTACACGATCCGTAACTAGTCGCGTCACAAAATG|
GOF_D02_F10          |T      C                    C          T                                                                     |  M: 4  |  D: 0
GOF_C09_D07_G05      |T                                                                                                            |  M: 1  |  D: 0
GOF_I14              |T                                                                                                            |  M: 1  |  D: 0
GOF_H05              |A               G                                                                                            |  M: 2  |  D: 0
GOF_E01_E02          |A                                 G              G                                                       -   |  M: 3  |  D: 1
GOF_F07              | C                                      A                                                                    |  M: 2  |  D: 0
GOF_F04              | A      A                                                                                            G       |  M: 3  |  D: 0
GOF_I4               |  C                                           C                                                              |  M: 2  |  D: 0
GOF_F11              |  A            C                                                                                             |  M: 2  |  D: 0
GOF_E06              |   T                                                                           G                             |  M: 2  |  D: 0
GOF_I8               |   G                                                                                                         |  M: 1  |  D: 0
GOF_G01_H02          |    T                                                                                                    -   |  M: 1  |  D: 1
GOF_C02              |     G   T                                              A                                                -   |  M: 3  |  D: 1
GOF_B04_D09          |     C                                    C                  -                                               |  M: 2  |  D: 1
GOF_C01_D12          |      --------                                         C                                                     |  M: 1  |  D: 8
GOF_C03_G02          |       CG                                                                                     C              |  M: 3  |  D: 0
GOF_A11              |         T                                          C                                                        |  M: 2  |  D: 0
GOF_H04              |         C        C                     G  T       C         G                             G                 |  M: 7  |  D: 0
GOF_F03_F05          |          C  T                                                                                           -   |  M: 2  |  D: 1
GOF_I7               |          C             C                                                                                    |  M: 2  |  D: 0
GOF_D05_E07          |          C                                                CG                                                |  M: 3  |  D: 0
GOF_A02              |            G   G                         G   C G                                                            |  M: 5  |  D: 0
GOF_I3               |             A        C                                     G                                                |  M: 3  |  D: 0
GOF_F08              |               T             C                C T  C   C                                                     |  M: 6  |  D: 0
GOF_G08              |                C                                                                                            |  M: 1  |  D: 0
GOF_I12              |                A TC                    C                            -                                       |  M: 4  |  D: 1
GOF_A04_B03_D04_H07  |                A  G                              T                                                      -   |  M: 3  |  D: 1
GOF_G03              |                 GG                                                                                          |  M: 2  |  D: 0
GOF_C10_D03_E08_H10  |                 A                                                                                           |  M: 1  |  D: 0
GOF_I11              |                   T C                                    T                                                  |  M: 3  |  D: 0
GOF_F01_H06          |                   T                                                                                         |  M: 1  |  D: 0
GOF_C12              |                   G                                                                                         |  M: 1  |  D: 0
GOF_G06              |                    A                                                                                        |  M: 1  |  D: 0
GOF_B07_F02_H01      |                     G                                                                                       |  M: 1  |  D: 0
GOF_C05_G12          |                      G                                                                                      |  M: 1  |  D: 0
GOF_A01              |                      C           G     G                                                                    |  M: 3  |  D: 0
GOF_C08              |                        T      T                                       C                                     |  M: 3  |  D: 0
GOF_D06              |                        G                                                                                    |  M: 1  |  D: 0
GOF_E11              |                        C                                                                                    |  M: 1  |  D: 0
GOF_A05              |                                        C   T                                                                |  M: 2  |  D: 0
GOF_B08              |                                         A                                                     -----     -   |  M: 1  |  D: 6
GOF_B10_H03          |                                         A                                                                   |  M: 1  |  D: 0
GOF_I9               |                                           T                                       A                         |  M: 2  |  D: 0
GOF_I13              |                                           T                                       A                         |  M: 2  |  D: 0
GOF_F06              |                                            A                                                                |  M: 1  |  D: 0
GOF_A10              |                                              C                                                          -   |  M: 1  |  D: 1
GOF_A07_E03          |                                               C  T C  C    G                                                |  M: 5  |  D: 0
GOF_I2               |                                               C  T C  C    G                                                |  M: 5  |  D: 0
GOF_G11              |                                                 G                                                           |  M: 1  |  D: 0
GOF_B09_C04_E12      |                                                  G                                                          |  M: 1  |  D: 0
GOF_G07              |                                                   C                                                     -   |  M: 1  |  D: 1
GOF_H11_H12          |                                                       C                                                     |  M: 1  |  D: 0
GOF_C07_D08_G04_G09  |                                                            G                                                |  M: 1  |  D: 0
GOF_I5               |                                                            G                                                |  M: 1  |  D: 0
GOF_I10              |                                                            G                                                |  M: 1  |  D: 0
GOF_I6               |                                                                     C                                       |  M: 1  |  D: 0
GOF_I1               |                                                                                                             |  M: 0  |  D: 0
Mutations            |5222120223301202523512304000110100200001522320422243300510117100000001010000000100020000000100100000010000000|
Deletions            |0000001111111100000000000000000000000000000000000000000000000100000001000000000000000000000000011111000008000|
```

Supplementary Table S8 B. Sequenced clones from LOF group

```
ΔATG               |GGGAGACGCAACTGAATCAACATAAGTGAACGCGACTCTATCTCCCTAAACTAGGAGTCATATAGCGGCACGGAGTTACACGATCCGTAACTAGTCGCGTCACAAAATG|
LOF_C06            |TA                             C                   C      T                                                  |  M: 5  |  D: 0
LOF_D11            |  T                            C  CG          AC                                                         -   |  M: 6  |  D: 1
LOF_A02            |  C       C                                 A     A            C                             TA              |  M: 7  |  D: 0
LOF_B06_D09        |  C                        A TG  T    C                       C                 A       T             G      |  M: 10 |  D: 0
LOF_A09            |  A                        T                            A        -                                           |  M: 3  |  D: 1
LOF_C03            |  A                                          T                C               A-                   A         |  M: 5  |  D: 1
LOF_B03            |  A                                                A             G                                           |  M: 3  |  D: 0
LOF_A10            |   T                                                                          T                      G       |  M: 3  |  D: 0
LOF_H11            |   C                                                                                  C                      |  M: 2  |  D: 0
LOF_B04            |    C                                                                       A         A                      |  M: 3  |  D: 0
LOF_B08_G08        |    A        C                                                                           -   -               |  M: 2  |  D: 2
LOF_F10            |     T    G                                           AC                                                     |  M: 4  |  D: 0
LOF_F03_G06        |     T                                             A           G       A                                     |  M: 4  |  D: 0
LOF_A05            |     G                                                        G                                              |  M: 2  |  D: 0
LOF_B07            |     C                      G        A G                            T                                        |  M: 5  |  D: 0
LOF_F09            |        A                                               A                  A            G        A           |  M: 5  |  D: 0
LOF_D03            |         T                    T                                  G -T                                        |  M: 4  |  D: 1
LOF_G02            |         C   TG                       A A                               A                                    |  M: 6  |  D: 0
LOF_F07_G07        |          T              T                               C                                                   |  M: 3  |  D: 0
LOF_H10            |          C                        T          G   A                                                        --|  M: 4  |  D: 2
LOF_A06            |          C                                                                C                  G              |  M: 3  |  D: 0
LOF_C10            |           T          G            A    G                                         C                          |  M: 5  |  D: 0
LOF_E12_H09        |            A     C                                                                          -   - -         |  M: 2  |  D: 3
LOF_F02            |             T             A            A                        G                                           |  M: 4  |  D: 0
LOF_D12            |             T                                            A                  T                           -   |  M: 3  |  D: 1
LOF_D01            |             A    G       G             A                                    C                               |  M: 5  |  D: 0
LOF_F04            |              G                                                         A   C                                |  M: 3  |  D: 0
LOF_D04_H12        |              C         C  A        C                                                         C              |  M: 5  |  D: 0
LOF_D02            |              -               A               A          G         C          A                              |  M: 5  |  D: 1
LOF_C01            |                C                       C          C                                  T         C  G C-      |  M: 7  |  D: 1
LOF_F08            |                A                A      C                A                                                   |  M: 4  |  D: 0
LOF_G03            |                 G                  A                                                     T                  |  M: 3  |  D: 0
LOF_C09            |                  T                                  -                     C   C                             |  M: 3  |  D: 1
LOF_G04            |                  G                            C      C C              C T                                   |  M: 6  |  D: 0
LOF_A01_E07        |                   G                         G                  A             T           ---                |  M: 4  |  D: 3
LOF_A12            |                   C                                   G               C                     T               |  M: 4  |  D: 0
LOF_A04_H05        |                     T     C  G                                   A                                          |  M: 4  |  D: 0
LOF_H03            |                     T               G                 T                          CG     G                   |  M: 6  |  D: 0
LOF_F12            |                     G    C     T                                                                            |  M: 3  |  D: 0
LOF_G11            |                     C       T         C       T        C       C                                            |  M: 6  |  D: 0
LOF_E10            |                      G                                                    A             G                   |  M: 3  |  D: 0
LOF_H02            |                      A                            A A                  C  G            C   GA           -   |  M: 8  |  D: 1
LOF_C05            |                       T                                     C                                A              |  M: 3  |  D: 0
LOF_C07            |                        C          A                 A   G                                                   |  M: 4  |  D: 0
LOF_A03            |                          C         A       G                      A   -    A                                |  M: 5  |  D: 1
LOF_E06            |                           C          G                  A             A                                     |  M: 4  |  D: 0
LOF_G10            |                            T       A                                                                        |  M: 2  |  D: 0
LOF_E01            |                             C T                                        C              C                     |  M: 4  |  D: 0
LOF_D10_F01_G12    |                             C                    T                                        C          A --   |  M: 4  |  D: 2
LOF_D06            |                              T                                             C                                |  M: 2  |  D: 0
LOF_E05            |                               A   A                                      C             C                    |  M: 4  |  D: 0
LOF_D05            |                               A          G                                                A            --   |  M: 3  |  D: 2
LOF_A07_E03        |                               A          C                            T                                     |  M: 3  |  D: 0
LOF_C11            |                                G                      T             T             A                         |  M: 4  |  D: 0
LOF_C12            |                                 A                              C           A      C                         |  M: 4  |  D: 0
LOF_E04            |                                   A                G                                                        |  M: 2  |  D: 0
LOF_D08            |                                      G                 A            C                                       |  M: 3  |  D: 0
LOF_E09            |                                      C     G          C        C                                    --  -   |  M: 4  |  D: 3
LOF_C04            |                                      C                  C                        C  A  C                    |  M: 5  |  D: 0
LOF_F06            |                                      C                          T                                       -   |  M: 2  |  D: 1
LOF_F11            |                                        G                           T     AG       G              A          |  M: 6  |  D: 0
LOF_H04            |                                              A        G                    A  T  T                          |  M: 5  |  D: 0
LOF_E02_H06        |                                               C                                A                            |  M: 2  |  D: 0
LOF_E11            |                                                      T                                       A              |  M: 2  |  D: 0
LOF_F05            |                                                      A                                                      |  M: 1  |  D: 0
LOF_H07            |                                                         C                A C                                |  M: 3  |  D: 0
LOF_A08_E08        |                                                           G                A   A                            |  M: 3  |  D: 0
LOF_H08            |                                                                 G -       G           A                     |  M: 3  |  D: 1
LOF_B02_C02_G05    |                                                                        T                            --      |  M: 1  |  D: 2
LOF_H01            |                                                                                              C              |  M: 1  |  D: 0
Mutations          |1162240012511530214204312136245623164272702032440035124657210132451232055137824230440132521213601112022000000|
Deletions          |0000000000000010000000000000000000000000000000000000100000000000010200010000000010000000001111200010102302701 1|
```

Comparison of the sequenced clones from GOF- and LOF-group. Depiceted are only the differences compared to 10A with deleted AUG (ΔAUG). For each row and for each column, the number of mutations and deletions are listed.

Supplementary Table S9. K_D and regulatory activity of selected fluoroquinolones

Fluoroquinolone	**K_D / nM**	**Activity / x-fold**
EFX *	61.3 (1.5)	3.1 (0.1)
CFX	64.2 (1.8)	7.5 (0.3)
DFX	137.1 (10.6)	4.2 (0.7)
NFX	182.6 (22.1)	2.7 (0.2)
EX	236.6 (52.9)	2.8 (0.1)
hCFX	366.7 (67.2)	0.8 (0.0)
dCFX	829.8 (118.0)	1.7 (0.2)
PA	916.1 (195.9)	2.8 (0.2)

For every fluoroquinolone, the dissociation constant (K_D) was determined by fluorescence titration and activity *in vivo* was measured by standard GFP fluorescence assay using the CFX-riboswitch. The standard deviation (± SD) is reported in brackets for the titration experiments and regulatory activity, respectively.

* EFX reduced the growth rate of yeast approx. 10-fold [data not shown].

Supplementary references

1. Suess,B., Hanson,S., Berens,C., Fink,B., Schroeder,R. and Hillen,W. (2003) Conditional gene expression by controlling translation with tetracycline-binding aptamers. Nucleic Acids Res., 31, 1853–1858.

2. Schneider,C. and Suess,B. (2016) Identification of RNA aptamers with riboswitching properties. Methods, 97, 44–50.

7.2 SICOR: Subgraph Isomorphism Comparison of RNA Secondary Structures

The following article:

- Schmidt, M., Hamacher, K., Reinhardt, F., Lotz, T.S., Groher, F., Suess, B. and **Jager, S.** (2017) SICOR: Subgraph Isomorphism Comparison of RNA Secondary Structures, IEEE/ACM Transactions on Computational Biology and Bioinformatics (Oct., 7th, 2017 submitted)

deals with a new algorithm for the comparison of RNA SS. The algorithm is a probabilistic sub-graph isomorphism applied to compare RNA structures by mapping the nodes of one graph to another. Such a comparison is extremely important as it can help determine structural diversity in SELEX experiments. This is of particular importance since such a permutation can be found in aptamers/riboswitches. In addition, a new CG scheme was created which displays RNA structures as directed graphs. The new CG scheme has the advantage, in contrast to all previously published methods, that it also takes into account the direction (5’ to 3’) of the RNA backbone. We were able to show that SICOR shows a better performance in terms of accuracy compared to related work.

Contributions Micahel Schmidt and myself had the idea for the publication. Moreover, I created Figure 1,5 and 6 together with Michael Schmidt. The definition of the algorithm, implementation and its run-time evaluation is conducted by Michael Schmidt. Felix Reinhardt optimized the implementation of Michael Schmidt. Furthermore, I created the toy-graphs for the run-time evaluation experiment and the evaluation on a real SELEX data set. Furthermore I helped to write the paper as well as to motivate it. I am the last author in this work. Florian Groher, Thea Sabrina Lotz and Beatrix Suess helped to write and motivate the paper and provided real NGS experiments for the evaluation. Kay Hamacher helped to write and motivate the Paper and supervised the definition of SICOR.

SICOR: Subgraph Isomorphism Comparison of RNA Secondary Structures

Michael Schmidt, Kay Hamacher, Felix Reinhardt, Thea S. Lotz,
Florian Groher, Beatrix Suess and Sven Jager

January 21, 2018

Abstract

RNA aptamer selection during SELEX experiments builds on secondary structural diversity. Advanced structural comparison methods can focus this diversity.

We develop `SICOR`, which uses probabilistic subgraph isomorphisms for graph distances between RNA secondary structure graphs. `SICOR` outperforms other comparison methods and is applicable to many structural comparisons in experimental design.

1 Introduction

RNA aptamers can form a multitude of structures to promote catalytic activity or interaction with different partners such as proteins or small molecules whose affinities and selectivities can rival those of antibodies [1, 2]. Thus aptamers are of ever-increasing interest in science, as they can be used for a variety of applications, e.g. as diagnostic tools, therapeutic agents, or synthetic biosensors in various applications [3, 4].

Aptamers can be discovered with an experimental process called Systematic Evolution of Ligands by Exponential enrichment (*SELEX*) [5, 6]. This iterative method isolates aptamers with the desired properties from highly diverse nucleic acid libraries over the course of several rounds of enrichment. The analysis of SELEX data is very time-consuming and labour-intensive, since it involves the analysis of individual candidates in complex procedures. It would be a much better approach to identify the rounds with the most enriched structural motifs and analyze them afterwards.

SELEX in combination with Next Generation Sequencing (NGS) opens up entirely new possibilities for computational analysis and data mining [7], identification and characterization of aptamers. This particular knowledge can be crucial to optimize the SELEX process by identifying the SELEX iteration which enriched aptamers, motifs or the sub-structure diversity. Most prominently such structural patterns play important roles both in RNA folding and their respective biochemical function [8–10]. Thus, the need for the development of novel computational approaches that address the characteristics specific to the SELEX protocol has become highly relevant [11].

There exist several routes to analyze NGS data mainly using secondary structure information. This is often done by Minimum Free Energy (MFE) approaches using dynamic programming algorithms [12]. The resulting MFE structures are usually represented as dot-bracket (*DB*) strings (Fig. 1 c)). Based on this simple representation of RNA structure, structural differences are compared using string distance metrics (e.g. Levenshtein Distance [13]) followed by a hierarchical clustering – thus relying on conceptual very different representations. A major shortcoming of such a string-based approach is the incapability to match displaced or permuted motifs as illustrated in Figure 1 a). However, permuted motifs are often found in SELEX experiments [14–16].

To address this problem, various graph-based representaions have been proposed so far [17]. Here, RNA structures are often represented as trees [18,19] or dual graphs [20] and compared with e.g., tree alignments (RNAforester) [21–23] or general edit distances [24]. Further approaches are alignment free (e.g., GraphClust) [25] or use alignments in combination with a coarse graining scheme for RNA secondary structures (e.g., BEAR: Brand nEw Alphabet for RNA in combination with BEAM: BEAR Motif finder) [26,27].

In our approach, RNA structures can be represented in an abstract manner as directed graphs (e.g. adjacency matrices) to incorporate structural properties [28,29]. This graph-based representation makes promising methods like e.g. Frequent Subgraph Mining feasible for RNA motif (pattern) discovery in NGS experiments [30]. However, to our best knowledge, RNA-graphs have not been used for direct structure comparison yet. In this paper, we introduce *SICOR* (Subgraph Isomorphism Comparison Of Rna structure), an efficient probabilistic subgraph isomorphism for RNA structure comparison.

The remainder of this paper is structured as follows: In Sec. 2, we introduce the conceptual framework as well as our biological application scenario. We introduce basic insights on the algorithm in Sec. 3. We present a run time evaluation of our algorithm in Section 4 as well as an interpretation of applying `SICOR` to a real world example. Accordingly, we compare SICOR with state of the art RNA structure comparison algorithms. Finally, we summarize and conclude our work in Sec. 5.

2 Background

2.1 Basic Idea

The DB notation represents aptamer structures on a semi-structural level; however, it assigns a large distance value although structurally two aptamer might be closely related (e.g. having just two swapped hairpins). Computing structural similarities based on the DB notation can thus lead to overestimation of the observed structural diversity. However, quantifying the correct structural diversity in NGS data derived from SELEX experiments can help to identify the deceiving SELEX iteration(s) and thus optimize and understand the process. For example, imagine two RNA structures *A* and *B* (Fig. 1 c)). Both contain the same structural motifs, they only differ slightly in size and distribution along the respective RNA backbone. Here, the Levenshtein distance based on the DB notation clearly overestimates the structural context. In our approach, the probabilistic algorithm `SICOR` maps every nucleotide of substructure *B* based on its graph representation to reference *A* and returns only a normalized distance of 0.04.

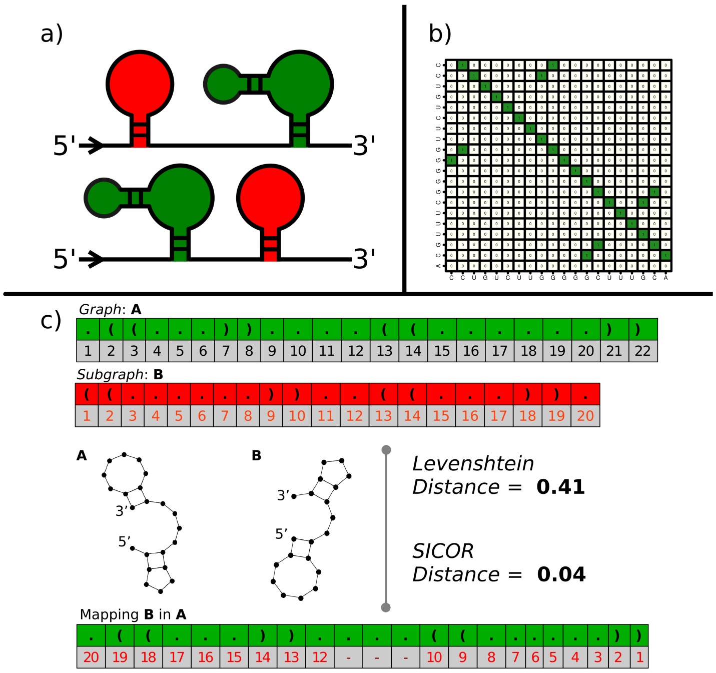


Figure 1: Illustration of our application scenario: a) Example of two RNA molecules which we aim to test for similarities. Both RNAs have the same structural motifs, suggesting similar functional properties. However, a simple string-based measure like Levenshtein distance based on DB annotation overestimates structural diversity due to a permutation and displacement of the motifs. This shortcoming can be improved by graph-based representation of the RNAs and comparison by subgraph isomorphism. b) Example of an RNA adjacency matrix. The matrix is symmetric except for the first off-diagonal, which corresponds to the backbone direction. c) Levenshtein *vs.* graph-based comparison: The normalized Levenshtein distance is 0.41 while a probabilistic subgraph isomorphism yields 0.04. Here, both distances were normalized: `SICOR` by Eq. (8) and Levenshtein by the length of the larger string. In the bottom we show the result for a subgraph-mapping of A onto B.

2.2 Data

2.2.1 Dataset from SELEX Experiments

The NGS data for the real world application scenario was derived from a RNA SELEX [1] against a small molecule ligand. The SELEX process reaches an enrichment (percentual amount of RNA exceeding a certain threshold) of ligand-binding aptamers, using a stringent experimental selection protocol. The resulting aptamers will then be applied as small molecule-triggered switches to control gene expression. We took a sample at the beginning and after applying the stringent protocol.

RNA secondary structures were predicted by Zuker's algorithm in RNAfold

---

[1] Synthetic Genetic Circuits, TU Darmstadt 2017, unpublished.

2.3.4 (RNAfold) at 300 K using thermodynamic parameters from Andronescu *et al.* [31,32].

2.2.2 Structure Comparison and Coarse Graining

Structure comparison with our method was done by conversion from DB-annotation into adjacency matrices and subsequent application of `SICOR`. In addition to benchmarks on full graphs, we extended our approach by using coarse grained representations of RNA secondary structures. Here, we converted the datasets to tree graphs using the `RNA Matrix Tool` [19,20]. "Classical" structure comparison was done using Levenshtein string distance algorithm [13] in combination with the DB annotations (as strings). Afterwards, the distance was normalized by the length of the longer string. We computed Levenshtein distance using `stringdist` library [33]. In addition, we benchmarked structure-based alignment approaches as well as tree editing and forest alignments. For the classical alignments RNA secondary structures were converted with `BEAR` [26] and aligned using `BEAM` [27] with default parameters. Tree editing was performed using the `RNAdistance` [31] programm (coarse grained, as well as full representation; using flag `-f -c`) The forest alignments were computed using the `RNAforester` programm (flag `-d`) [34].

3 Algorithm

In this Section we describe the `SICOR` algorithm, our new approach for comparison of RNA structures. We start with two RNA sequences $\boldsymbol{a}$ with length n and $\boldsymbol{b}$ with length m and $m \leq n$. The first part of our algorithm consists of predicting secondary RNA structures via the well-known minimum-free-energy approach (3.1), resulting in adjacency matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Subsequent application of an inexact subgraph isomorphism (3.2) gives a similarity score σ for the sequences and their respective secondary structures.

3.1 Conversion of Dot-Bracket Annotation to RNA Connectivity Graphs

We define an RNA connectivity *graph* as $G = (V, E)$. It can be represented as an ordered pair of *vertices* $V = \{v_1, v_2, \dots v_{|V|}\}$ and *edges* E. Here, we consider *undirected and directed graphs without loops*, i.e., $E \subseteq \{\{v, w\} : v, w \in V, v \neq w\}$. Let $\vec{s}$ be a DB vector of the length $|s|$ with $\vec{s} = \{s_1, s_2, \dots s_{|s|}\}$ where $s_i \in \{., (,)\}$ and $|s| = |V|$. The character "." denotes an unpaired nucleotide, e.g. by hydrogen bonding. Vertex i can participate in opening a base pair (BP) $s_i =$ "(" with a vertex j participating in closing the BP $s_j =$ ")". A BP corresponds to an edge between v_i and v_j. The primary structure of an RNA sequence is accounted for by assuming edges between a nucleotide and its direct neighbors along $\vec{s}$ within a distance of $i - j = 1$. The resulting *adjacency matrix* $\boldsymbol{A}$ of a graph G is a $|V| \times |V|$ symmetric matrix, defined as follows:

$$A_{ij} := \begin{cases} 1 & : i \neq j \wedge \{v_i, v_j\} \in E \\ 0 & : \text{else} \end{cases} \tag{1}$$

The direction of an RNA strand is crucial (e.g. 5' to 3') for its functionality. Thus, to further improve our representation, we included this information in our RNA representation. This is done by setting every entry $A_{i,i-1} = 0$. For simplicity, we will refer to the number of vertices $|V|$ of the two graphs to be compared as m and n respectively.

3.2 Subgraph isomorphism

Given an (n x n)-adjacency matrix $\boldsymbol{A}$ and an (m x m)-adjacency matrix $\boldsymbol{B}$ with $m \leq n$, the (sub)graph isomorphism problem can be formulated: Obtain a (m x n)-permutation matrix $\boldsymbol{P}$ such that

$$f(\boldsymbol{P}) = \left\| \boldsymbol{B} - \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^T \right\|_2^2 . \tag{2}$$

Minimizing $f(\boldsymbol{P})$ over feasible $\boldsymbol{P}$ results in a mapping from vertices in A to the ones in B. We call $\boldsymbol{P}$ a *pseudo* permutation matrix in the case of subgraph isomorphisms, meaning that $P_{ij} \in \{0,1\}$, as well as $\forall j : \sum_i P_{ij} \in \{0,1\}$ and $\sum_j P_{ij} = 1$. Eq. (2) would be zero for *exact* matching of *induced* subgraph isomorphisms. In the following we use a heuristic optimization method, leading to a *probabilistic* subgraph isomorphism for two reasons:

1. The exact subgraph isomorphism problem is NP-complete, implying that most likely no algorithm exists which finds solutions in polynomial time. In fact the combinatorial complexity is $O(\frac{n!}{(n-m)!}m^2)$, resulting from $\frac{n!}{(n-m)!}$ possible pseudo permutation matrices which can be checked in $O(m^2)$ time for isomorphism.
2. Given adjacency matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ derived for two different RNA sequences, exact matching will be impossible in most cases. This is caused by (i) edges in graph $\boldsymbol{A}$ which are not present in subgraph $\boldsymbol{B}$ and (ii) edges in $\boldsymbol{B}$ which are not present in $\boldsymbol{A}$.

For the probabilistic approach we use a similar scheme as described in [35] for graph isomorphisms. We apply a convex relaxation onto P and relax it to a (m x n)-*pseudo* bistochastic matrix $\boldsymbol{S}$. The space of pseudo bistochastic matrices Ω is the convex hull of pseudo permutation matrices and consists of elements $S_{ij} \in [0,1]$, $\sum_i S_{ij} \in [0,1]$ (*pseudo* left stochastic) and $\sum_j S_{ij} = 1$ (right stochastic). This approach leads to a continuous optimization problem

$$\min_{\boldsymbol{S}\in\Omega} f(\boldsymbol{S}) = \min_{\boldsymbol{S}\in\Omega} \left\| \boldsymbol{B} - \boldsymbol{S}\boldsymbol{A}\boldsymbol{S}^T \right\|_2^2 \tag{3}$$

with partial derivatives

$$\frac{\partial f(\boldsymbol{S})}{\partial S_{ij}} = 2 \cdot \left[\boldsymbol{S}\boldsymbol{A}\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{A}^T + \boldsymbol{S}\boldsymbol{A}^T\boldsymbol{S}^T\boldsymbol{S}\boldsymbol{A} \right. \tag{4}$$

$$\left. - \boldsymbol{B}\boldsymbol{S}\boldsymbol{A}^T - \boldsymbol{B}^T\boldsymbol{S}\boldsymbol{A} \right]_{ij}, \tag{5}$$

which we solve by local gradient-based optimization methods. To obtain a similarity score σ for the RNA sequences, $\boldsymbol{S}_{\text{opt}}$ is projected back onto the space of pseudo permutation matrices Π via orthogonal projection [35]

$$\boldsymbol{P}_{\text{opt}} = \underset{\boldsymbol{Q}\in\Pi}{\operatorname{argmax}} \operatorname{tr}\left(\boldsymbol{S}_{\text{opt}}^T \boldsymbol{Q} \right) . \tag{6}$$

Eq. (6) can be formulated as the linear assignment problem (LAP)[2]

$$\boldsymbol{P}_{\text{opt}} = \underset{\boldsymbol{Q}}{\operatorname{argmin}} \sum_{i,j} -S_{ij} Q_{ij} \tag{7}$$

with $-S_{ij}$ denoting the cost for assignment Q_{ij} of worker i to task j, fulfilling $Q_{ij} \geq 0$, $\sum_i Q_{ij} = 1$ and $\sum_j Q_{ij} = 1$. This can be solved in polynomial time using the Hungarian method[3] [36]. Similarity score is obtained by $\sigma = 1 - \delta$ with the normalized `SICOR` distance

$$\delta = \frac{\left\| \boldsymbol{B} - \boldsymbol{P}_{\text{opt}} \boldsymbol{A} \boldsymbol{P}_{\text{opt}}^T \right\|_2^2}{\min(|E|_A, m^2 - |E|_B) + \min(|E|_B, n^2 - |E|_A)}, \tag{8}$$

with $|E|_A$, $|E|_B$ denoting the number of edges and $n = |V|_A$, $m = |V|_B$ the number of vertices of the corresponding graph. The equation above represents a normalized distance as in the worst case every edge of $\boldsymbol{B}$ is mapped onto a non-edge of $\boldsymbol{A}$ and *vice versa*.

We implemented above optimization algorithm in the programming languages `julia` and `C` using `julia/C` library `NLopt` for gradient-based optimization (algorithm: augmented lagrangian method) and `julia` library `Munkres` for the Hungarian method. The implementation is integrated in a repository hosted on our webserver: `http://www.cbs.tu-darmstadt.de/SICOR.tar.gz`

Eq. (3) is non-convex and thus the results of a local minimization depends on initial values of $\boldsymbol{S}$, meaning that we need to employ a second optimizer in a nested fashion (*meta-optimization*) to infer solutions near to the global minimum. We benchmarked different Monte-Carlo based approaches such as Metropolis sampling [37] or simulated annealing [38] for the second optimizer. We observed peculiar inefficiencies for this problem due to a "spiky" potential with many local minima[4]. We found that a simple multistart algorithm (i.e. starting the local optimizer of eq. (3) with n different values of $\boldsymbol{S}$) yielded better results.

Note that initial inputs $\boldsymbol{S}$ for the subgraph isomorphism in eq. (3) must be random pseudo bistochastic matrices. We achieve this by first constructing a random matrix with i.i.d. elements drawn from a uniform distribution in the interval $[0, 1]$. Pseudo bistochasticity is accomplished by an iterative procedure, consisting of projection $\xi_1(\boldsymbol{S})$ onto the space of right stochastic matrices (by normalizing row sums to $\forall i : \sum_j S_{ij} = 1$), subsequently followed by projection $\xi_2(\xi_1(\boldsymbol{S}))$ onto pseudo left stochastic matrices (by normalizing column sums to $\forall j : \sum_i S_{ij} \leq 1$). Iterative application of ξ_1 and ξ_2 is guaranteed to converge to some point in the intersection of right- and pseudo left stochastic matrices as they both form closed convex sets [39,40]. We show in Algorithm 1 our method in pseudo-code form.

---

[2] To be formally correct, in order to map eq. 6 onto LAP, $\boldsymbol{S}_{\text{opt}}$ has to be filled up with zeros ($\hat{=}$ "virtual" workers with no cost for each arbitrary task; this is necessary due to the constraint $\sum_j Q_{ij} = 1$) to form a quadratic (n x n)-matrix $\boldsymbol{S}'_{\text{opt}}$, also resulting in a quadratic matrix $\boldsymbol{Q}'$. However, an optimal assignment of real workers ($\hat{=}$ minimizing the total cost of eq. (7)) is also optimal in the presence of virtual workers, as the latter have equal cost for every task and thus do not induce further constraints onto the problem.

[3] There is guaranteed to exist a solution $\boldsymbol{Q}$ with integer values due to total unimodularity of the constraint matrix.

[4] Our scheme consisted of taking the locally optimized value $\boldsymbol{S}_k$ of the k-th run and modulating it with noise for the $(k+1)$-th run.

```
Data: RNA sequences $\boldsymbol{a}$ and $\boldsymbol{b}$ with length($\boldsymbol{a}$) $\geq$ length($\boldsymbol{b}$)

Function SubGrIso($\boldsymbol{A}$,$\boldsymbol{B}$,$\boldsymbol{S}$)
    minimize eq. (3) with gradient (5)
    subject to constraints $\forall j : \sum_i S_{ij} \leq 1$ and $\forall i : \sum_j S_{ij} = 1$
return $\boldsymbol{S}_{\text{opt}}$

begin
    compute adjacency matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ e.g. with Zuker algorithm
    initialize Res=Array(k)                    /* do $k$ multistarts */
    for $i$ in 1 to $k$ do
        construct random pseudo bistochastic matrix $\boldsymbol{S}$
        $\boldsymbol{S}_{\text{opt}}$=SubGrIso($\boldsymbol{A}$,$\boldsymbol{B}$,$\boldsymbol{S}$)
        Res[$i$]=Hungarian($\boldsymbol{S}_{\text{opt}}$)  /* Project $\boldsymbol{S}_{\text{opt}}$ acc. to Eq. (7) */
        if $f(\boldsymbol{S}) == 0$                /* cmp. Eq. (3) */
         then
            break          /* Break if perfect match is found */
        end
    end
end
```

Algorithm 1: Summary of all *SICOR*-steps for comparing two RNA sequences.

Although we use above algorithm for planar graphs (due to restrictions of omitting pseudo-knots in the Zuker algorithm), it should be noted that it is valid for all kind of graphs. This implies, one could replace Zuker with a more complex structure prediction algorithm (which might have pseudo-knots and thus leads to non-planarity) or even some database-search/homology-modeling approach. Still, the above implementation would be applicable. In the following section, we benchmark this on different graphs.

4 Evaluation

In this section, we evaluate the run-time performance of `SICOR`. First, two benchmarks were performed on synthetic data, namely Erdős–Rényi graphs, where we investigate the influence of graph and subgraph sizes and sparsities on the run time performance. The last benchmark was carried out on realistic toy models of RNA graphs to investigate the general applicability of our algorithm. For the real world application, we analyzed a SELEX dataset and compared our approach to state of the art analytics using Levenshtein distance on DB annotation.

4.1 Benchmark on Erdős–Rényi graphs

The first benchmark was performed on random (n x n)-Erdős–Rényi graphs $\boldsymbol{A}$ with sparsity[5] $s = 0.5$, leading to $\sum_{ij} A_{ij} \approx \frac{n\cdot(n-1)}{2}$ as we set the diagonal to

---

[5] Sparsity s is defined as the percentage of nonzero elements in the strictly upper triangular matrix.

zero. Erdős–Rényi graphs were generated with `julia` package `Erdos` for different number of vertices $n \in \{25, 50, 75, 100\}$. We randomly chose induced (m x m)-subgraphs $\boldsymbol{B}$ out of $\boldsymbol{A}$, meaning that exact matching is possible. Multistart scheme as described in Alg. 1 with a maximum trial number of $k = 10^5$ was used for accuracy- and run time evaluation. We expected an increasing runtime for growing subgraph sizes due to combinatorial complexity.

We observed the mismatch error according to Eq. (3) to vanish with a few exceptions (2 % of all runs). Results of runtime comparison[6] for different n and m are summarized in Figure 2. Runtime consists of the average duration of a `SubGrIso` run multiplied by the number of multistarts. Overall results can be summarized into three different regimes:

1. For small subgraph sizes there is a high number of degenerated global minima as many exact matchings are possible. Due to the multistart scheme, `SICOR` finds a global minimum in short runtimes.

2. As subgraph size increases, the number of (degenerated) global optima decreases. However, there is now a variety of steep local minima which are close in depth to the global minimum. Landing of the gradient-descent algorithm in such a potential well (but *not* in the local minimum) leads to high gradients and big jumps on the potential landscape. We suggest this to be the reason for longer convergence times.

3. For subgraphs larger than a threshold t there exists one global minimum and relative sizes of local minima decrease. We expect this to flatten the potential landscape, leading to faster convergence times. Interestingly, runtimes do not grow for subgraph sizes $m \gg t$, which makes `SICOR` suitable for comparison of large RNA subgraphs.

In the second benchmark on Erdős–Rényi graphs, we evaluated the influence of sparsity on the runtime performance. We expect high runtimes for sparse and dense graphs as there exist many local minima which are close in depth to the global minimum, leading to a high number of required multistarts. Runtimes should decrease in the intermediate region as a higher number of constraints transforms the potential landscape into a regime similar to the third one mentioned above. Results[7] are summarized in Figure 3. We do not observe a significant influence of s on runtime except for a small increase for dense graphs. Note that in our RNA application we have sparse graphs with $s \geq 0.95$.

4.2 Benchmark on realistic RNA toy graphs

In this section we benchmark `SICOR` on RNA toy graphs for which a perfect matching exists. Toy graphs are generated in a two step procedure, consisting of creating a random DB string $\vec{s}$ and subsequent transformation into graph-based representation as described in Section 3. To create the random DB string

---

[6] Benchmark system was a Debian-operating server with two *Intel(R) Xeon(R) CPU E5-2687W v2 @ 3.40GHz* and disabled hyperthreading. Every `SICOR`-evaluation was performed on exactly one physical core.

[7] Benchmark system was a Debian-operating server with two *Intel(R) Xeon(R) CPU X5660 @ 2.80GHz* and disabled hyperthreading. Every `SICOR`-evaluation was performed on exactly one physical core.

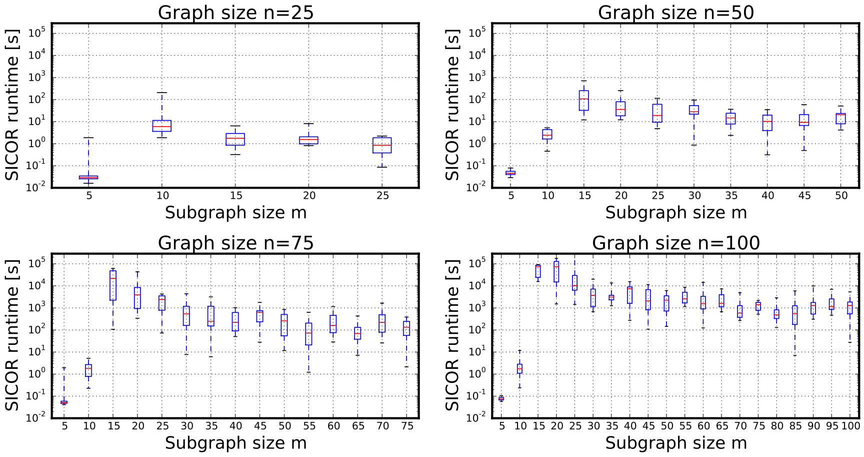


Figure 2: Runtime comparison for different combinations of graph size n and subgraph size m. Each box-and-whisker consists of 10 different samples and whiskers were chosen to contain the $[0, 100]$ percentile.

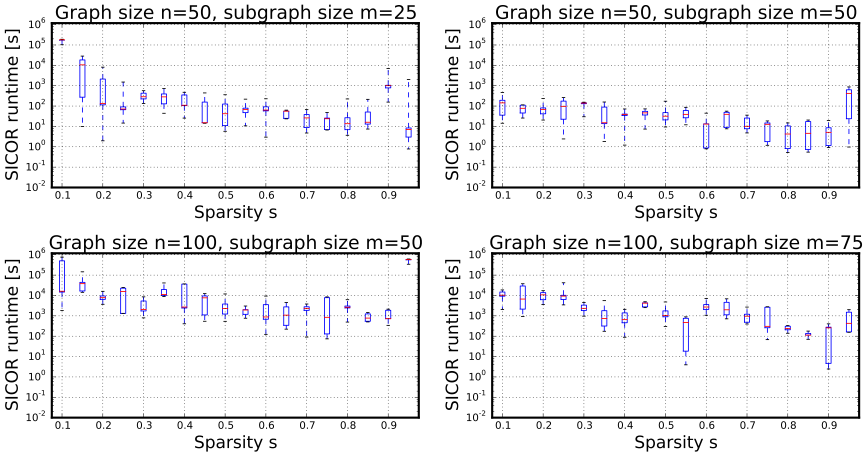


Figure 3: Runtime comparison for different sparsities s. Each box-and-whisker consists of 5 different samples and whiskers were chosen to contain the $[0, 100]$ percentile.

we use a Markov chain $\vec{p}_{i+1} = \boldsymbol{T}\vec{p}_i$ with transition matrix

$$\boldsymbol{T} = \begin{pmatrix} \frac{1}{2} & \frac{1}{3} & 0 \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{2} \\ 0 & \frac{1}{3} & \frac{1}{2} \end{pmatrix} \tag{9}$$

starting in the state $\vec{p}_0 = (0, 1, 0)^T$. Afterwards string entries $s_i \in \{-1, 0, 1\}$ are drawn according to the corresponding probability vector $\vec{p}_i$. Here, -1 corresponds to an opening bracket in DB (`(`), +1 a closing bracket in DB (`)`) and 0 a dot (`.`).

Only $\vec{s}$ with an equal number of open and closed parantheses are chosen for conversion into RNA toy graphs. We implemented this generator with the `markovchain` library in `R` [41, 42]. The resulting graphs are still random while preserving the structural elements of RNA. We generated graphs of size $n \in \{50, 100, 150, 200\}$ and drew randomly connected subgraphs of size $m \in \{5, 10, ..., 30\}$ from it[8]. Apart from toy model generation, benchmark setup was the same as in previous section[6]. It should be noted that we used undirected graphs for the purpose of simplicity.

We expect absolute runtimes to be higher than in previous section due to the higher complexity of fully connected subgraphs. We observe a slightly increasing error of the optimal `SICOR`-match for large subgraphs (up to 4 mismatched edges for $m = 30$). Runtime benchmark is shown in Figure 4. Despite higher absolute runtimes, behavior is similar to Erdős–Rényi graphs as we have short runtimes for small subgraph sizes which saturate for large sizes. However, we do not observe an increase in runtime for medium sizes. We assume this comes due to a "friendlier" potential landscape of planar RNA graphs as compared to Erdős–Rényi graphs.

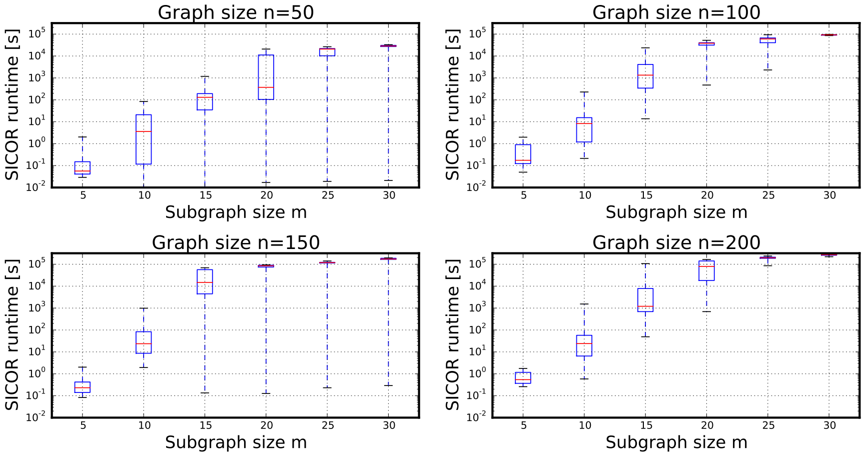


Figure 4: Runtime comparison for different combinations of graph size n and subgraph size m. Each box-and-whisker consists of 10 different samples and whiskers were chosen to contain the $[0, 100]$ percentile.

4.3 Application to Synthetic Biology: SELEX & Aptamers

As a real world example, we used the two datasets described in Section 2.2.1. The datasets consist of a sample from the beginning of a SELEX process and a sample after selection took place. For this setup, we want to benchmark the previously introduced scoring of `SICOR` in Section 3 versus different "state of the art" RNA sequence comparison methods. Here, we consider four different kinds of approaches: Tree editing on full and coarse grained trees (implemented in Vienna RNA e.g., `RNAdistance` [31]), classical alignments using a structural alphabet for RNA secondary structures (e.g., `BEAR` [26]), Levenshtein distance

---

[8] This scheme represents a simplification of real RNA structures as they include slightly different motifs and their different permutation.

computed on DB strings and finally structure comparison by forest alignments using `RNAForester` [34]. In addition, we perfomed SICOR also on the coarse grained tree graphs [19, 20]. For each method, we compared the structure with the highest abundance (reads per million, rpm) with all remaining structures of the set. In order to find an optimal matching within `SICOR`, we used 10^4 multistarts for each comparison and chose the minimal distance afterwards. The first SELEX iteration contains 327 sequences. For this round, we expect a high diversity in terms of structural patterns. The last SELEX iteration contains 166 sequences. Here, we expect a high similarity between the sequences. This assumption can be explained due to an experimental treatment with consistent artificially imposed selection pressure. Average distance values are shown in Tab. 1.

Table 1: Average distances for the benchmarked methods and their corresponding standard deviation.

	First SELEX round	Last SELEX round
BEAR alignment	0.44 ± 0.13	0.31 ± 0.18
Edit tree	0.53 ± 0.08	0.41 ± 0.22
Edit tree (coarse-grained)	0.26 ± 0.06	0.21 ± 0.12
Forest alignment	215.47 ± 56.62	173.66 ± 111.96
Levenshtein	0.45 ± 0.08	0.35 ± 0.2
`SICOR`	0.24 ± 0.03	0.08 ± 0.06
`SICOR` (tree graph)	0.07 ± 0.08	0.05 ± 0.07

As expected, all methods show lower distances for the last round than the first round, verifying their applicability for RNA sequence comparison. However, to identify the aptamer-enriched SELEX iteration in an experimental application, a distinct signal in form of a high difference between the two SELEX rounds is needed. To measure the performance of the different methods, we used the relative mean deviation

$$\nu = \frac{\bar{\delta}_{\text{first}} - \bar{\delta}_{\text{last}}}{\bar{\delta}_{\text{first}} + \bar{\delta}_{\text{last}}}, \tag{10}$$

where $\bar{\delta}_{\text{first}}/\bar{\delta}_{\text{last}}$ denotes the corresponding average distance of the first/last SELEX round. Accordingly, ν qunatifies differences in structural diversity or in other words, secondary structure motif enchrichment. Figure 5 shows relative deviations for the different methods.

As can be seen, `SICOR` applied on raw DB data performs best with a relative deviation of about $\nu = 0.5$, followed by `SICOR` applied on coarse-grained tree graphs and BEAR alignments (both $\nu = 0.17$). Moreover, `SICOR` outperforms advanced approaches like forest alignments. Even coarse-grained tree representation in combination with `SICOR` performs better than tree editing on a full tree graph (where every vertex is one nucleotide).

This makes `SICOR` well suited for comparison of RNA secondary structures and identification of enriched SELEX rounds. Moreover, `SICOR` can be used to identify the corresponding structural motif in the considered round. The result gives also rise to future developments, as one might find a more adapted/advanced coarse-graining scheme to have the same accuracy as raw

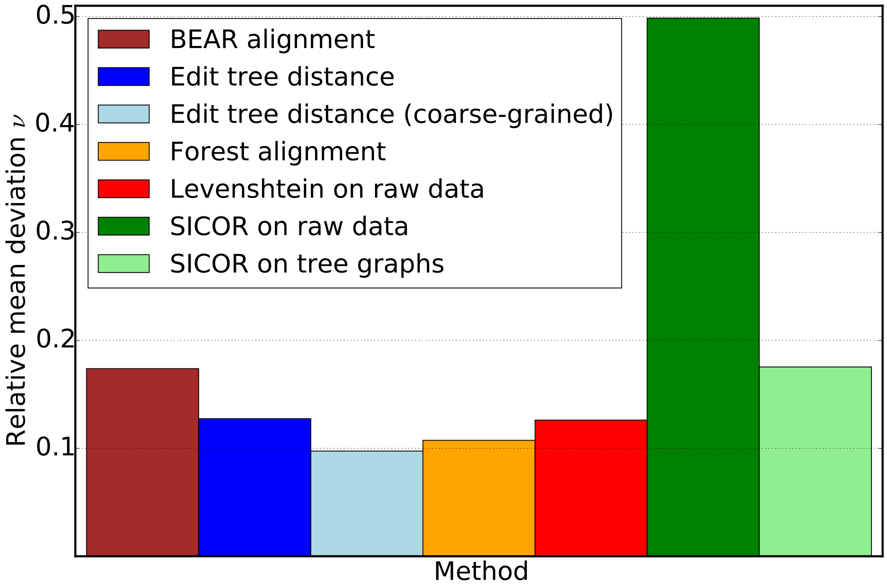


Figure 5: Evaluation on real world datasets: Relative mean deviations for our benchmarked methods (higher is better). For calculation of ν-values, data was weighted by the corresponding rpm values.

DB data, however significantly decreasing computation time. Interestingly, all remaining methods perform equally well in a region of about $\nu = 0.1$ to 0.13. Figure 6 shows a detailed histogram of `SICOR` and Levenshtein distances, both applied on raw DB data. Both methods show relatively high distances in the first SELEX iteration, with `SICOR`'s values lowering in the final round. In contrast, the Levenshtein scheme is clearly overestimating structural diversities and having a large statistical dispersion. This again shows the advantage of a graph-based approach for structure comparison to a purely string-based metric.

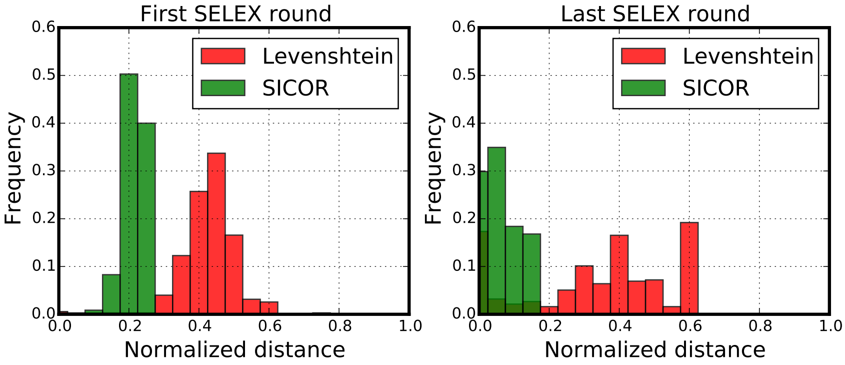


Figure 6: Histogram of normalized distances of the SELEX runs. Data is weighted by the corresponding rpm values.

5 Summary, Conclusion, & Future Work

In this work we demonstrate that subgraph isomorphisms are well suited for structural diversity quantification of directed RNA graphs. The main advantage of this graph-based approach (as opposed to string-based metrics) lies in its ability to compare structural motifs *independently* of their appearance on the string. This enables e.g. the mapping of two similar motifs, which have a relative permutation in the string-based representation (Fig. 1 a)). In this way, suitable structural motifs for synthetic biology applications (e.g., Riboswichtes, Ribozymes) can be identified much more quickly.

Furthermore, our proposed algorithm `SICOR` is a *probabilistic* subgraph isomorphism, allowing it to find good matches between similar, but different, RNA graphs.

Within the scope of our evaluation `SICOR` clearly outperforms Levenshtein distance in terms of accuracy, as the latter can assign too large structural diversity values in (among other) the real world example.

Moreover, `SICOR` outperforms advanced comparison and coarse graining schemes (Fig. 5) and showed a better accuracy on coarse grained tree graphs (with coarse-graining according to [19]) in contrast to classical tree editing (according to [31]). However, tree editing comes also with several limitations like "scattering" effects (e.g., missmatching due to a misleading objective function of total edit costs.) [43]. The particular coarse graining scheme of [19] represents secondary structure elements as vertices and, in related work, it could be shown to be benficial for the 3D structure prediction of aptamers [44]. This renders `SICOR` more applicable to SELEX datasets, despite better run time performance of other metrics (e.g., Levenshtein's algorithm).

More precisely, using `SICOR` could be beneficial for the analysis of aptamer structures from NGS data obtained by SELEX rounds. Knowing the correct structural diversity for a SELEX experiment can help to optimize the aptamer selection process and help to identify important structural motifs.

`SICOR` is ideal for the comparison of RNA structures, due to their inherent modularity. On the basis of this, new unknown patterns/motifs can also be identified. Hence, we envision that `SICOR` will be used for *de novo* RNA design of, e.g., switchable, binding or catalytical motifs. Our method is easily transferable to other areas of molecular and synthetic biology where a comparison of topological structures is required, for example comparison of metabolic pathways or protein structures. In addition to that we chose to use a *generic* subgraph isomorphism, meaning that it is feasible for all kind of graphs. For future applications, in principle one even could replace Zuker with another structure prediction algorithm or experimental structures with pseudoknots. Since `SICOR` perfomed good on coarse grained tree graphs, another future development could be the usage of more adapted coarse-graining schemes like forgi [45] or more complex networks like dual graphs [19]. This will lead to significant reduction of run time. We conclude, that our proposed algorithm `SICOR` fulfills sensible demands in terms of run time and accuracy and is thus a powerful tool for the analysis of NGS data. Due to its transferability it will considerably improve a broad range of molecular/synthetic biology applications.

6 Acknowledgements

This work was supported by the Landesoffensive für wissenschaftliche Exzellenz (LOEWE; initiative to increase research excellence in the state of Hessen, Germany) as part of the LOEWE Schwerpunkt CompuGene and iNAPO, as well as the DFG (GRK 1657 project 1A, CRC 902 A2).

References

[1] S. Hanson, G. Bauer, B. Fink, and B. Suess, "Molecular analysis of a synthetic tetracycline-binding riboswitch," *RNA*, pp. 2549 – 2556, 2005.

[2] B. M. Warfield and P. C. Anderson, "Molecular simulations and Markov state modeling reveal the structural diversity and dynamics of a theophylline-binding RNA aptamer in its unbound state," *PLOS ONE*, 2017.

[3] E. J. Cho, J.-W. Lee, and A. D. Ellington, "Applications of aptamers as sensors." *Annual review of analytical chemistry (Palo Alto, Calif.)*, vol. 2, pp. 241–64, 2009.

[4] D. E. Cameron, C. J. Bashor, and J. J. Collins, "A brief history of synthetic biology." *Nature reviews. Microbiology*, vol. 12, no. 5, pp. 381–90, 2014.

[5] a. D. Ellington and J. W. Szostak, "In vitro selection of RNA molecules that bind specific ligands." *Nature*, vol. 346, no. 6287, pp. 818–22, 1990.

[6] C. Tuerk and L. Gold, "Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage t4 dna polymerase," *Science*, vol. 249, no. 4968, pp. 505–510, 1990.

[7] N. Nguyen Quang, G. Perret, and F. Ducongé, "Applications of High-Throughput Sequencing for In Vitro Selection and Characterization of Aptamers," *Pharmaceuticals*, vol. 9, no. 4, p. 76, 2016.

[8] T. Schlick, "Mathematical and Biological Scientists Assess the State of the Art in RNA Science At an Ima Workshop, RNA in Biology, Bioengineering, and Biotechnology," *International Journal for Multiscale Computational Engineering*, vol. 8, no. 4, pp. 369–378, 2010.

[9] M. Parisien and F. Major, "The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data," *Nature*, vol. 452, no. 7183, pp. 51–55, 2008.

[10] C. Laing and T. Schlick, "Computational approaches to RNA structure prediction, analysis, and design," *Curr. Opin. Struct. Biol.*, vol. 21, no. 3, pp. 306–318, 2011.

[11] P. Dao, J. Hoinka, M. Takahashi, J. Zhou, M. Ho, Y. Wang, F. Costa, J. J. Rossi, R. Backofen, J. Burnett, and T. M. Przytycka, "Aptatrace elucidates rna sequence-structure motifs from selection trends in ht-selex experiments," *Cell Systems*, vol. 3, no. 1, pp. 62 – 70, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2405471216302204

[12] M. Zuker and P. Stiegler, "This paper presents a new computer method for folding an RNA molecule Nucleic Acids Research," *Nucleic Acids Research*, vol. 9, no. 11981, pp. 133–148, 1980.

[13] V. I. Levenshtein, "On the minimal redundancy of binary error-correcting codes," *Information and Control*, vol. 28, no. 4, pp. 268–291, 1975.

[14] Rani P. G. Cruz, Johanna B. Withers and Y. Li, "Dinucleotide Junction Cleavage Versatility of 8-17 Deoxyribozyme Rani," *Chemistry & Biology*, vol. 11, pp. 57–67, 2004.

[15] M. Legiewicz, C. Lozupone, R. Knight, and M. Yarus, "Size, constant sequences, and optimal selection." *RNA (New York, N.Y.)*, vol. 11, no. 11, pp. 1701–9, 2005. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16177137{%}5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1370856

[16] M. Legiewicz and M. Yarus, "A more complex isoleucine aptamer with a cognate triplet," *Journal of Biological Chemistry*, vol. 280, no. 20, pp. 19 815–19 822, 2005.

[17] P. Tijerina and R. Russell, *Biophysics of RNA Folding*, R. Russell, Ed. Springer, 2013, no. January 2013. [Online]. Available: http://www.springerlink.com/index/10.1007/978-1-4614-4954-6

[18] N. Kim, K. N. Fuhr, and T. Schlick, *Graph Applications to RNA Structure and Function.* New York, NY: Springer New York, 2013, pp. 23–51. [Online]. Available: http://dx.doi.org/10.1007/978-1-4614-4954-6_3

[19] N. Kim, L. Petingi, and T. Schlick, "Network theory tools for rna modeling," *WSEAS transactions on mathematics*, vol. 9, pp. 941–955, 09 2013.

[20] D. Fera, N. Kim, N. Shiffeldrim, J. Zorn, U. Laserson, H. H. Gan, and T. Schlick, "RAG: RNA-As-Graphs web resource." *BMC bioinformatics*, vol. 5, p. 88, 2004.

[21] M. Höchsmann, T. Töller, R. Giegerich, and S. Kurtz, "Local similarity in RNA secondary structures," in *2nd IEEE Computer Society Bioinformatics Conference, CSB 2003, Stanford, CA, USA, August 11-14, 2003*, 2003, pp. 159–168. [Online]. Available: https://doi.org/10.1109/CSB.2003.1227315

[22] G. Blin, A. Denise, S. Dulucq, C. Herrbach, and H. Touzet, "Alignments of rna structures," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, no. 2, pp. 309–322, April 2010.

[23] S. Schirmer and R. Giegerich, *Forest Alignment with Affine Gaps and Anchors.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 104–117. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-21458-5{_}11

[24] T. Jiang, G. Lin, B. Ma, and K. Zhang, "A general edit distance between RNA structures." *Journal of computational biology : a journal of computational molecular cell biology*, vol. 9, no. 2, pp. 371–388, 2002.

[25] S. Heyne, F. Costa, D. Rose, and R. Backofen, “Graphclust: Alignment-free structural clustering of local RNA secondary structures,” *Bioinformatics*, vol. 28, no. 12, pp. 224–232, 2012.

[26] E. Mattei, G. Ausiello, F. Ferrè, and M. Helmer-Citterich, “A novel approach to represent and compare RNA secondary structures,” *Nucleic Acids Research*, vol. 42, no. 10, pp. 6146–6157, 2014.

[27] M. Pietrosanto, E. Mattei, M. Helmer-Citterich, and F. Ferrè, “A novel method for the identification of conserved structural patterns in RNA: From small scale to high-throughput applications,” *Nucleic Acids Research*, vol. 44, no. 18, pp. 8600–8609, 2016.

[28] S. Jager, B. Schiller, T. Strufe, and K. Hamacher, “StreAM-T_g : Algorithms for analyzing coarse grained rna dynamics based on markov models of connectivity-graphs,” *Springer LNCS*, vol. 9838, 2016.

[29] K. Hamacher, J. Trylska, and J. A. McCammon, “Dependency map of proteins in the small ribosomal subunit,” *PLoS Comput Biol*, vol. 2, no. 2, pp. 1–8, 2006.

[30] M. Gawronski, Alex R.and Turcotte, “Ribofsm: Frequent subgraph mining for the discovery of RNA structures and interactions,” *BMC Bioinformatics*, vol. 15, no. 13, p. S2, 2014.

[31] R. Lorenz, S. H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, “Viennarna package 2.0,” *Algorithms for Molecular Biology*, vol. 6, no. 1, p. 26, 2011.

[32] M. Andronescu, A. Condon, H. H. Hoos, D. H. Mathews, and K. P. Murphy, “Efficient parameter estimation for RNA secondary structure prediction,” *Bioinformatics*, vol. 23, no. 13, pp. 19–28, 2007.

[33] M. van der Loo, “{stringdist}: an {R} Package for Approximate String Matching,” *The R Journal*, vol. 6, no. 1, pp. 111–122, 2014.

[34] M. Höchsmann, “The tree alignment model: algorithms, implementations and applications for the analysis of RNA secondary structures,” *Naturwissenschaften*, 2005.

[35] Y. Aflalo, A. Bronstein, and R. Kimmel, “On convex relaxation of graph isomorphism,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, pp. 2942–7, Mar 2015.

[36] H. W. Kuhn, “The hungarian method for the assignment problem,” *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[37] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, “Equation of state calculations by fast computing machines,” *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[38] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, “Optimization by simulated annealing,” *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[39] H. H. Bauschke and J. M. Borwein, “On projection algorithms for solving convex feasibility problems,” *SIAM Review*, vol. 38, no. 3, pp. 367–426, 1996.

[40] V. Cappellini, H.-J. Sommers, W. Bruzda, and K. Życzkowski, “Random bistochastic matrices,” *Journal of Physics A: Mathematical and Theoretical*, vol. 42, no. 36, p. 365209, 2009.

[41] G. A. Spedicato, *markovchain: discrete time Markov chains made easy*, 2015, R package version 0.4.3.

[42] R. D. C. Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008.

[43] J. Allali and M.-F. Sagot, “Novel Tree Edit Operations for RNA Secondary Structure Comparison,” *Algorithms in Bioinformatics, 4th International Workshop, WABI 2004, Proceedings*, vol. 3240, pp. 412–425, 2004.

[44] N. Kim, M. Zahran, and T. Schlick, “Computational prediction of riboswitch tertiary structures including pseudoknots by ragtop: A hierarchical graph sampling approach.” *Methods in enzymology*, vol. 553C, pp. 115–135, 03 2015.

[45] P. Kerpedjiev, C. Höner, Z. Siederdissen, and I. L. Hofacker, “Predicting RNA 3D structure using a coarse-grain helix-centered model,” *Rna*, vol. 21, no. 6, pp. 1110–1121, 2015.

7.3 Motif Based Analysis of NGS Data

The following section illustrates the usage of dynamic graphs and motif counts to model NGS data from SELEX experiments. In the previous chapters, the structural diverstity of SELEX rounds was determined by comparing the different RNA structures for each SELEX rounds. On the one hand, the LD (*170*) in combination with Hofacker's DB representation and on the other hand the structures were used as graphs compared by SICOR. By means of these methods, the structural diversity of a particular SELEX round can be determined and accordingly compared with other rounds. Thus, the search space for follow-up experiments can be drastically reduced. Yet, the methods only help to evaluate the structural context of each round. A much more interesting question is actually whether defined substructures in the underlying rounds are being enriched.A much more interesting question is actually whether special substructures in the underlying rounds are being enriched. These substructures are modules or parts of modules created by the evolutionary process of SELEX. There is a great deal of interest in this modules, because these are the evolved functional part of the aptamers and can be used for engineering, re-engineering or bio-conjugation. To address all mentioned problems we present *NGStream*. Here, NGStream allows for bridging a methodological gap between motif-based analysis of MD simulation and graph-based analysis of HTS-SELEX.

Methods

The results of NGS experiments quickly end up in a large number of sequences. Especially when RNA-seq (i.e. NGS) for SELEX (cf. Section 4.2.2) rounds is applied, a significant number of sequences per rounds can be recorded (up to millions of different sequences). Thanks to Zuker's algorithm, we can predict a secondary structure from these sequences alone (cf. Section 4.4.4). This feature makes it possible to identify structural diversity and enrichment within SELEX. At the same time, it also offers the possibility to count motifs with the help of StreaM$_k$ in the RNA graphs. The NGS data originate from Florian Groher and have been analyzed in the manuscript in Section 7.1 before. From this data set, the first aptamers were created against *ciprofloxaicin* (CFX) and it consists of NGS data from 12 SELEX rounds (details can be found in Section 7.1). In order to create the respective dynamic graphs, we convert the DB data set (cf. Section 7.1) to Hofacker's full-tree representation. We choose this representation because it offers a good trade-off between full and CG representation (cf. Section 4.4.4). This makes it possible to map bigger structural-motifs because by now we can only count up to 7-vertex motifs. Since tree-motifs only depict a small subset of the motifs, the search space is also reduced here. Here, k-vertex motifs can also depict structural-motifs/configurations (e.g., stacks, loops, hairpins, and bulges, etc.) in the RNA, depending on which representation one choose.

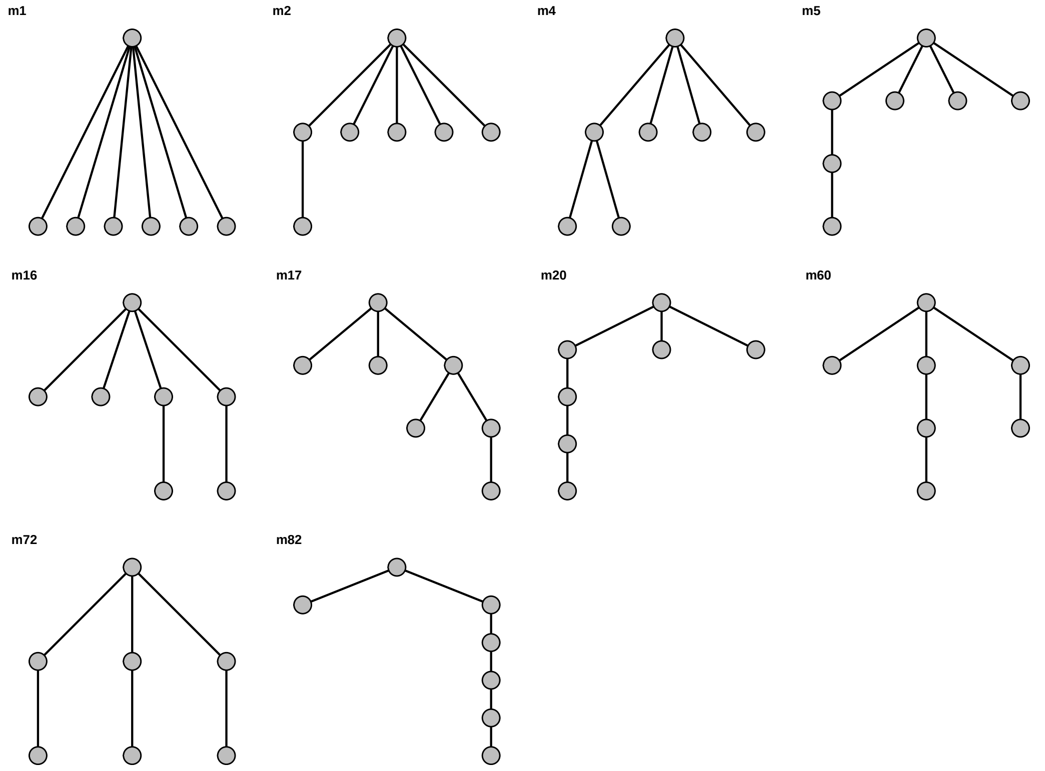


Figure 7.1: All 7-vertex tree motifs

Another advantage of using Zuker's algorithm is that the distance parameter d for creating the graphs is completely omitted (cf. Section 4.4.3). A short statistic of the 12 dynamic graphs can be found in Table. However, a classical analysis of secondary structure strings (comparison via LD) (*170*) was applied to this data. From this experiment some aptamers like *R10K6* and *10A* were created, both exhibit special rare motifs.[1].

---

[1] A special thanks goes to Florian Groher who helped with the evaluation and interpretation of the data

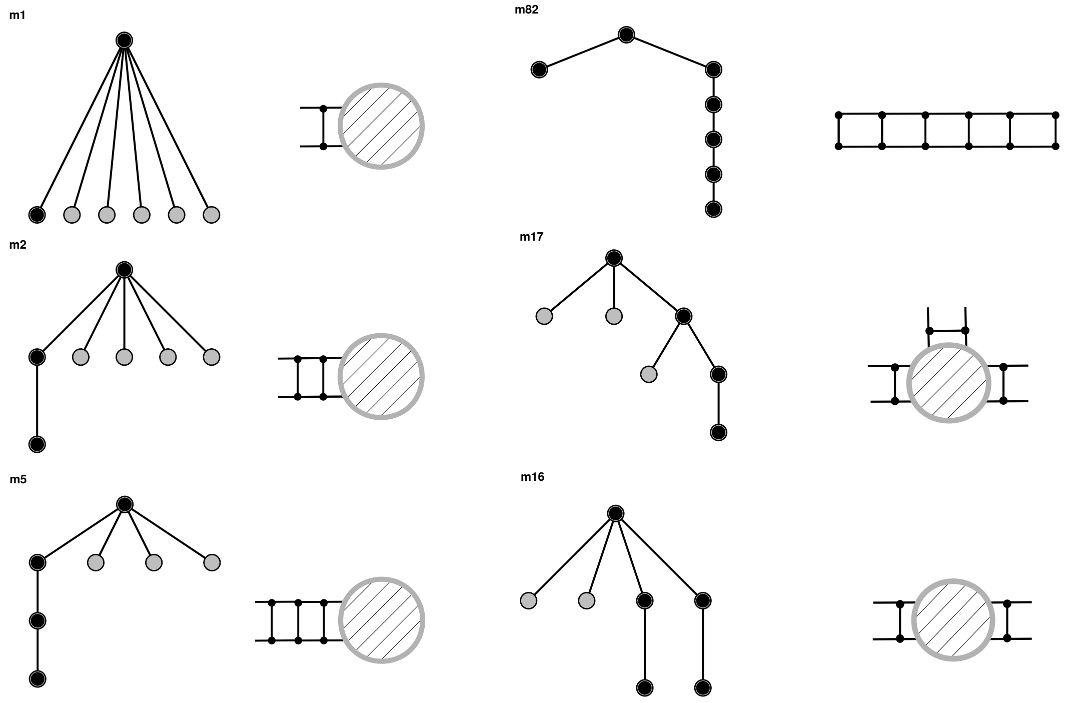


Figure 7.2: 7-vertex motifs on the right and on the left corresponding RNA structural-motifs

Results

In the following, the mean motif counts are investigated for $k \in \{3, 4, 5, 6, 7\}$ in each of the 12 SELEX rounds. We expect, e.g., the mean of the counted structural motifs $\overline{F_{\mathcal{M}_k}(m)}$, to rise for special enrichment events. Especially for motifs of the size $k \in \{3, 4, 5\}$, it is not possible to record any interesting structural features, as an entire unit, because of their size. Figure 7.3 shows the mean motif counts of motifs of the size $k \in \{3, 4, 5\}$.

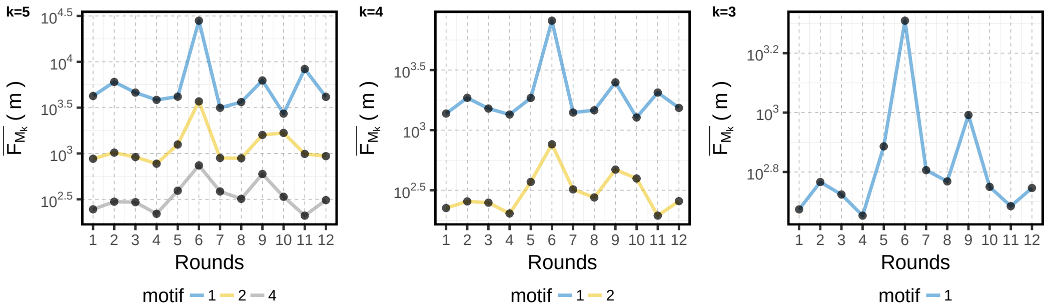


Figure 7.3: Analysis of CFX-SELEX using average motif counts for each SELEX round. Here we consider the size $k \in \{3, 4, 5\}$

For motif $k = 3$ there is only one tree motif for $k = 4$ two and for $k = 3$ five tree motifs. In each case, a peak appears at round six (cf. Figure 7.3). The strong peak at round six occurs on almost every motif up to m_{72}, this motif is one of the motifs found in multi-loop structures. However, to explain this peak one have to look into the experimental protocol. For this step, the process was relieved of background binders (washing step). These aptamers are mostly unstructured with a high free energy.

Moreover, they have an advantage in terms of amplification via PCR after each round, compared to the structured RNA molecules. After the washing step, we now have a strongly structured population which drops again after the next amplification at round seven. Here, we could detect that the selections pressure shifts again from the side of high affinity sequences (aptamers, motifs) to sequences which are just better amplifying via PCR.

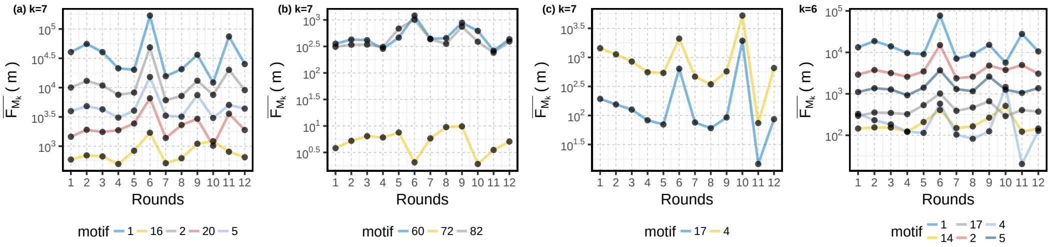


Figure 7.4: Analysis of CFX-SELEX using average motif counts for each SELEX round. Here we consider the size $k \in \{6, 7\}$

Figure 7.4 shows the mean motif counts for $k \in \{6, 7\}$. What is noticeable here is, that the course for some motif counts is offset. Namely, for the counts of motif: m_1, m_2, m_5 and m_{20}. These motifs represent hairpins with an additional stacking pair (cf. Figure 7.3; all illustrated motifs, cf. Figure 10.1)). It's an almost ascending order when motif four is omitted. This observation makes perfect sense, because this motif is very simple to find. Accordingly, the extension of the motif (e.g., by a stacked pair) is only a small change and ensures that the course of the rounds is the same. In addition to this findings, two motifs could also be found which are enriched by the factor ten. This could be observed in round ten. Figure 7.4 shows motif m_4 and m_{17} with their corresponding structural motif.

Exactly these motifs can be found in R10k6 secondary structure [2]. This aptamer was found in round ten and afterward engineered.

---

[2] see Section 7.1

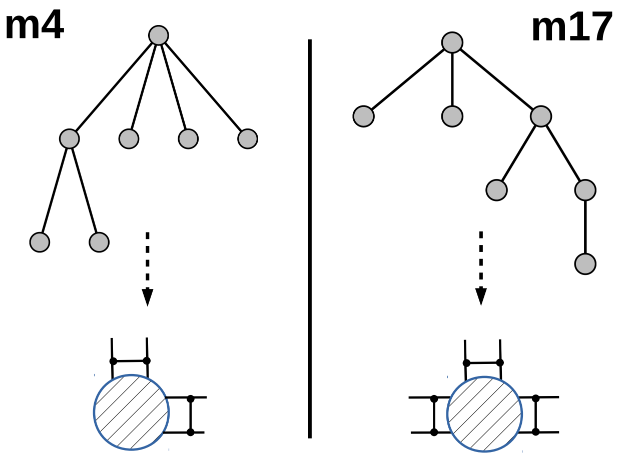


Figure 7.5: 7-vertex motif class m_4 and m_{17} and the corresponding structural element.

At this moment it could be shown that NGStream can identify the rounds in which aptamers appear or rare structural motifs enrich. However the most impressive thing here is, this happens without comparing a single sequence. Especially the dynamic graph format, which showed a good performance for MD simulations (consisting of millions of snapshots), is perfectly suited for RNA graphs. However, setting up the dynamic graph offers a small optimization problem, because the structural progression is not continuous here in contrast to MD. Nevertheless, the performance of a stream-based approach depends heavily on the batch (change-set, cf. Section 8) between the frames in the respective graphs. Furthermore, we also have the possibility to choose each of the presented representations for RNA structure (cf. Section 4.4.4), in order to model the dynamic graph. Especially in terms of different representations, this motif-based semantic would give a complementary description of the CFX-SELEX set.

8 Contributions

8.1 Puplications

- Groher, F., Bofill-Bosch, C., Schneider, C., Braun, J., **Jager, S.**, Geißler, K., Hamacher, K., Suess, B., Riboswitching with ciprofloxacin – Development and characterization of a novel RNA regulator. Nucleic Acids Research NAR, gkx1319, https://doi.org/10.1093/nar/gkx1319

- Dombrowsky, M.J., **Jager, S.***, Schiller, B., Mayer, B.E., Stammler, S., Hamacher, K., StreaMD: Advanced analysis of Molecular Dynamics using R, Journal of Computational Chemistry JCC (Oct., 20th, 2017 submitted and in revision)

- Schmidt, M., Hamacher, K., Reinhardt, F., Lotz, T.S., Groher, F., Suess, B. and **Jager, S.**, SICOR: Subgraph Isomorphism Comparison of RNA Secondary Structures, IEEE/ACM Transactions on Computational Biology and Bioinformatics (Oct., 7th, 2017 submitted)

- **Jager, S.***, Schiller, B., Babel, P., Blumenroth, M., Strufe, T., & Hamacher, K. (2017). StreAM-Tg: algorithms for analyzing coarse grained RNA dynamics based on Markov models of connectivity-graphs. Algorithms for Molecular Biology, 12(1), 15.

- Gross, C.*, Hamacher, K., Schmitz, K., & **Jager, S.** (2017). Cleavage Product Accumulation Decreases the Activity of Cutinase during PET Hydrolysis. Journal of Chemical Information and Modeling, 57(2),243-255.

- **Jager, S.***, Schiller B., Strufe T., Hamacher K. (2016) StreAM-Tg: Algorithms for Analyzing Coarse Grained RNA Dynamics Based on Markov Models of Connectivity-Graphs. In: Frith M., Storm Pedersen C. (eds)

Algorithms in Bioinformatics. WABI 2016. Lecture Notes in Computer Science, vol 9838. Springer

- Buss, O.*,**Jager, S.***, Dold, S.-M., Zimmermann, S., Hamacher, K., Schmitz, K., & Rudat, J. (2016). Statistical Evaluation of HTS Assays for Enzymatic Hydrolysis of β-Keto Esters. PloS One, 11(1), e0146104.

- Schiller B.*, **Jager, S.**, Hamacher K., Strufe T. (2015) StreaM - A Stream-Based Algorithm for Counting Motifs in Dynamic Graphs. In: Dediu AH., Hernandez-Quiroz F., Martin-Vide C., Rosenblueth D. (eds) Algorithms for Computational Biology. AlCoB 2015. Lecture Notes in Computer Science, vol 9199. Springer

- **Jager, S.**, Buss, O. (2018) Neue in silico-Methoden für die Etablierung einer Grünen Chemie, BIOspektrum-Ausgabe 01/18, Springer

8.2 Conference Poster

- Helmer D.*, Rink I. ,Dalton J.A.R.,Brahm K, **Jager, S.**, Joest, M., Wadhwani, P., Brenner-Weiss, G., Rapp, E.B., Giraldo, J., Hamacher, K., Schmitz, K. Peptides and peptide mimetics to inhibit the interaction of CXCL8 with its receptors, European Chemokine and Cell Migration Conference (ECMC, 2016)

- Buss, O.*, **Jager, S.***, Syldatk C., Rabe, K., and Rudat, J., β-Amino Acid Synthesis by an Engineered ω-Aminotransferase, International Conference on Molecular Interaction Engineering (MIE, 2016)

- **Jager S.***, Reinhardt F., Schmidt, M., Hamacher, K., In collaboration with: Lehr, F., Koeppl, H., Groher, A., Suess, B. RNA Aptamer Optimization through Secondary Structure Prediction, CompuGene Symposium: Computer-aided Engineering of Synthetic Genetic Circuits (CESGC, 2017)

- Groher A.*, **Jager, S.***, Schneider, C., Hamacher, K., Suess, B. A Large Scale Approach towards Riboswitch Design, CompuGene Symposium: Computer-aided Engineering of Synthetic Genetic Circuits (CESGC, 2017)

- Hamacher, K., Gross, C., Schmitz, K., **Jager, S.**, Cleavage Product Accumulation Decreases the Activity of Cutinase during PET Hydrolysis, GDCh-Wissenschaftsforum Chemie 2017 (GDCh 150 anniversary, 2017)
- Schlichting, N., **Jager S.**, Reinhardt F., Huxhorn, T., Schmidt M., Koeppl, H., Kabisch, J., Hamacher, K., Computer-aided Prediction of DNA Assembly Reactions and Experimental Workflows, 69. Mosbacher Kolloquium Synthetic Biology (GBM Symposium, 2018)

8.3 Invited Talks

- **Jager, S.*** and Hamacher, K., From Molecular Co-Evolution to Biophysical Annotation –Computational Methods for Synthetic Biology and Beyond, 1st RMU Bioinformaics Symposium (RMU, Jul., 27th, 2016), Frankfurt
- **Jager, S.***, Schiller, B., Strufe, T. and Hamacher K. : StreAM-T_g: Algorithms for Analyzing Coarse Grained RNA Dynamics based on Markov Models of Connectivity-Graphs, 16th Workshop on Algorithms in Bioinformatics (WABI, Aug., 24th, 2016), Arhaus
- Schiller, B.,**Jager, S.**, Hamacher, K., Strufe, T.: StreaM - a Stream-based Algorithm for Counting Motifs in Dynamic Graphs 2nd International Conference on Algorithms for Computational Biology, (AlCoB, Aug., 4-5, 2015), Mexico
- **Jager, S.*** and Rohden, F., Biologie trifft Informatik, Meta Rhein Main Chaos Days (MRMCD, Sep., 9th, 2013), Darmstadt

8.4 Supervised Work

- Development and optimization of dynamic programming approaches in the field of RNA bioinformatics, Thomas Huxhorn, TU Darmstadt, Jan. 2018, Bachelor Thesis
- Investigation of the influences of disulfide bonds on the activity of *Fusarium solani* Cutinase using computational and experimental methods, Christine Groß, TU Darmstadt, Oct. 2014, Master Thesis

- In silico Design of Murine Terminal Deoxyribonucleotidyl-Transferase Variants to Increase the Tolerance for 3'-O-Modified Nucleoside Triphosphates, Sebastian Palluk, TU Darmstadt, Oct. 2015, TU Darmstadt, Master Thesis
- Setup of an MD-Simulation to Evaluate the Catalytic Activity of Mice Terminal Desoxyribonucleotidyl-Transferase Variants, Sebastian Palluk, TU Darmstadt, Apr. 2015, Internship - M.FPR Biomolecular Engineering

9 Summary and Discussion

This thesis described several computational approaches to support the design of functional biomolecules in synthetic biology. Among them were protein (namely enzymes) as well as RNA (in the form of aptamers).

Enzyme design aims to improve bio-catalyst properties to optimize stability in the presence of organic solvents, temperature, and turn-over rates as well as the acceptance of high substrate and product concentrations. Here, design and engineering tools are greatly needed, as evolution has not yet had sufficient time to develop efficient enzymes for the hydrolysis of man-made materials like PET. Here, MD offers the possibility to analyze the interaction of the enzyme *Fusarium Solanie Cutinase* (FsC) and its reaction-product *Ethan-1,2-diol* (EG) on a molecular level.

In Section 5.1, the interaction of EG and the FsC was investigated via MD simulations. Solvent-induced interactions were to be found in the conformational dynamics of the enzyme. Starting from this, we analyzed not only the protein itself but also the enzyme's environment (solvent; EG) as well as hydration shells (water). As a result, a set of amino acids was identified on the surface, which interacts very strongly with EG. In addition, a design rule was derived: mutants with increased flexibility near the active site seem to compensate for the solvent-mediated reduction in activity. This result could form the basis of an advanced HTS approach, which yields an efficient bio-catalyst for the conversion of PET or polyesters. The generated knowledge limits the solution space (amino acid positions) for randomization. Accordingly, the next step would be to screen the randomized cutinase variants with a suitable assay, combined with a test substrate for activity. The assay should focus on a relevant substrate (PET).

However, finding the right assay is an often underestimated factor for the success of any HTS approach. Section 5.2 introduced a procedure for the selection of chemical assays in HTS. We were able to base this process on statistical tests which quantify the distances of negative and positive controls by endpoint measurements.

The result of this procedure is to identify that assay, which provides a clear separation of positive and negative signals. Such a pre-screening makes it possible to minimize false positive and negative rates automatically. Lipases were found by using this procedure, which converted the desired substrate (β-Keto ester) efficiently. It was also possible to illustrate that the pNB-Est13 (experimental contribution) is suitable for the conversion of β-Keto esters and thus could potentially extend the spectrum for applications because esterases can efficiently convert hydrophilic substrates.

Simulation methods like MD of Section 5.1 provides information on the motion of molecule and thus supports identification of new variants with desired properties. In addition new graph-based analysis methods have also been developed for the analysis of MD simulations. In particular, we investigated whether motifs are suitable for describing molecular structures and interactions. Motifs (see definition in Section 4.4.1) are patterns of connectivity occurring in complex networks at numbers that are significantly higher than those in randomized networks (*171–173*).

Hence, the counts of motifs in these networks help us understand their complex organization and were used in this thesis to describe dynamic properties of systems in computational biology, to compare spatial structures, to design and engineer novel functional sites, and to predict the structure and function of uncharacterized proteins (*123*, *124*).

The approach proposed in Section 6.1 relies on the transformation of molecular structures into distance-dependent dynamic graphs. Molecular proximities were mapped into an abstract "motif-space." Based on this idea, the StreaM algorithm was introduced and applied to MD simulations of the pNB-Est13 complex. StreaM turned out to be an efficient algorithm for 4-vertex motifs. It was found that certain 4-vertex motif counts increased in stable secondary structural elements, and others were equally likely over the protein in general.

This finding shows that secondary structures, like α-helices, correspond to a unique topology in their graph representation. From thereon, the algorithm was extended by Schiller *et al.* (*128*) to lokk into motifs of up to 7 vertices. The algorithm was named StreaM$_k$ and achieves speedups up to 19 043 fold on synthetic dynamic graphs and up to 2882.20 fold dynamic graphs taken from MD simulations compared to preexisting work. This performance increase enables the analysis of MD simulations with high granularity.

In Section 6.2.2, StreaM$_k$ was applied to two simulations of a helical peptide at different temperatures. In this basal example, it was identified that 7-vertex motif counts are well suited for quantifying essential structure-based dynamics (e. g. folding or unfolding). For instance, those 7-vertex motifs were identified, which occur more often during unfolding or folding. This renders a motif-based semantic as very sensitive in the classification of dynamical processes. In addition, we were able to show that traditional analysis methods such as the RMSD, underestimate the ensemble of conformations created during a simulation.

Furthermore, we could demonstrate that the application of these techniques is not limited to protein simulations. Rather, it has been shown that RNA in particular, thanks to its modular structure, is ideally suited for these motif-based semantics. Since the MD community has mainly focused on proteins, MD simulations of RNA are an area in which there is a lack of analysis methods. For this reason, the StreAM-T_g algorithm, which calculates MSM from motif transitions, was developed as part of this work.

The combination of MD and the StreAM-T_g (cmp. Section 6.3 algorithm not only revealed the nature of the conformational change but also identified the participating nucleotides. The knowledge gathered here can help identify design possibilities and propose hypotheses and new experiments.

As mentioned in the previous Section 5.1, the characterization of the environment (context) can be an important part of the analysis of a simulation; this method was extended one last time. In Section 6.2, we analyzed simulations of SPC/E water and approximated the number of interactions (cf. Equation 6.1) between water molecules based on 3-vertex motif counts.

This is a well suited graph-based representation hence a water molecule also consists of three atoms and therefore each water molecule represents one of two possible 3-vertex motifs. The other 3-vertex motif can now be used to measure water interactions. This methodology aims to describe the hydrogen bonding network of SPC/E water in the liquid phase of 27 MD simulations (cf. Table 10.3).

Here, we obtained statistically significant correlations between these motif-based results and experimental entropy values (cmp. Section 6.2). Second, we studied the organization and disorder as well as thermodynamics of water in pore-shaped confinements in Section 6.2.2.

Here, again we identified a benefit of our motif-based approach over commonly investigated distance-based properties. In conclusion, it could be demonstrated that the derivation of (time-dependent) graph properties reveals insights into the dynamical structure of the modeled (bio)molecular systems and the relations between its components on a more expressive semantic level (k-vertex motifs) than 3D coordinates.

However, in the context of these findings, the question arises as to whether the simple analysis of spatial 3D coordinates of atoms is semantically advanced enough to describe molecular complexity at all. Nonetheless, 3D coordinate-based metrics such as RMSD are often used for modeling MSMs or HMMs from simulations or in advanced sampling methods like *meta-dynamics* or *umbrella sampling*. It would be fascinating to use our unique kind of graph-based semantic in the future for *de novo* structure prediction to improve or accelerate sampling methods within the field of MD or MC.

The proposed motif-based concepts always require a structure for modeling. Unfortunately, from today's point of view, it is not possible to compute the correct topology of arbitrary proteins (with high accuracy) directly from the sequences on-the-fly.

There exists pure-sequence-based approaches such as *Mutual Information* based alignment analysis or *Direct Coupling Analysis*, which can predict the topology of a proteins only from available sequence data. Each of these approaches quantify a possible (co)evolution in an underlying sequence set from the same protein.

Nevertheless, these methods are error-prone and require many sequences. Moreover, in a recent publication by Schmidt and Hamacher (*44*), it could be shown that two-point interactions alone are not sufficient enough to derive contact maps for proteins because these models improve significantly when three-body interactions are used for the computation. This insight indicates that the structural prediction of proteins from the sequence is more complex than expected and (co)evolution has to be considered between at least three amino acids. Consequently, these methods are not suitable for use on NGS data. Nevertheless, in the light of the growing attractiveness and accessibility of NGS for scientists (e.g., Nanopore Project, *MinION* (*174*)), it is becoming more and more important to the scientific community to develop efficient and expressive methods in order to gather knowledge from NGS experiments.

The unique property of RNA is the ability to predict (secondary) structures directly from the sequence (using Zukers algorithm, cf. Section 4.2.3). With the help of this structural information, RNA secondary structure graphs can be generated that serve as the basis for an advanced graph based representation and pattern search. This possibility facilitates the analysis and understanding of e.g. the SELEX process (cmp. Sec. 4.2.2) on a structural level. During this iterative HTS process, aptamers were isolated with the desired properties from highly diverse synthetic nucleic acid libraries over the course of several rounds of enrichment. This process is also often referred to as a "black box", which contains a lot of experimental pitfalls.

Consequently, reducing the number of rounds to screen is a necessity because screening large libraries can be quite laborious and wasteful in terms of materials and time. Nevertheless, if this process is monitored with the help of NGS, this gives a unique insight into the variety of sequences that occur during the different SELEX rounds. Moreover, it is possible to calculate the MFE structure, which has many advantages: for instance, if there are only a few changes in the primary sequence, thanks to the non-linearity of the MFE prediction, the impact onto the structure could be much larger. In the same way a methodical combination of NGS and SELEX was used in the Chapter 7.1 in order to discover the first Riboswitch for CFX, a well-known fluoroquinolone antibiotic which can be used besides bacteria (i.e. prokaryotic cells) in mammalian cells.

Here, for each sequence in each SELEX round the secondary structure was determined using Zuker's algorithm. The structures were presented as DB strings and compared with the help of LD for each SELEX round. Despite the optimized software and distributed parallel computing of LD, the comparison of the entire data needed some three weeks on 280 CPU cores.

However, the LD distributions analysis was able to identify rounds in which a strong, structural enrichment occurred. Moreover, the novel CFX-Riboswitch (10A) emerged from one of these rounds. A major shortcoming of a string-based approach is its incapability to match displaced or perturbed structural motifs. Based on this representation of RNA structure, topological differences are compared using string distance metrics. However, by applying these simple metrics, the distance between two structures is usually overestimated. To overcome this shortcoming we introduced *SubgraphIsomorphism Comparison of RNA structure* (SICOR), an efficient probabilistic subgraph isomorphism for RNA structure comparison (cf. Section 7.2). *SICOR* outperforms string and also tree based metrics regarding accuracy. This renders `SICOR` more applicable to SELEX data sets. Nonetheless, compared to the string-based methods (e.g., LD with a run-time complexity of $\mathcal{O}(nm)$), SICOR is still too slow (run-time) taking the amount of data generated by NGS into account.

It is also possible to perform a mapping from one structure to another. This information can further be used to help to identify previously enriched structural motifs in the SELEX iterations on the nucleotide level. In order to find enriched (sub)structures, the motif counting concepts were applied to the NGS data of CFX-SELEX[1]. In order to do this, the StreaM$_k$ protocol was redesigned to handle NGS data. The result of this is called NGStream and can be found in Section 7.3. NGStream delivers a different result/output than SICOR. While the SICOR score or LD maintains the similarity of RNA structures (besides the mapping information), NGStream counts structural elements in the form of k-vertex motifs. Thus NGStream allows the structural composition to be viewed on multiple scales of SELEX iteration. Both pieces of information are complementary and can help to find the SELEX rounds, and from where novel aptamers or motifs could be derived.

The combination of the introduced methods could be integrated into a single process capable of identifying structural motifs and the nucleotides from which they are composed. This process could help to identify new switchable motifs from RNA pools. The first step in this process would be to predict secondary structure from the NGS data derived from SELEX experiments. Afterwards, the MFE structures (i.e. string representation) are then modeled into directed or undirected graphs in the next step.

Furthermore, many different CG schemes from Section 4.4 can be augmented. Accordingly, the next step involves converting this graph set into a dynamic graph. In this step, the possibility of optimizing the dynamic graphs regarding its sequential arrangement arises. This is because the run-time of the algorithms on dynamic graphs depends strongly on the batch size (e.g., Section 6.1). Thus, it is possible that due to a perfect arrangement of the graphs, prior to the conversion into dynamic graphs, the run-time of StreaM$_k$ can be further minimized. Afterwards, motifs can be counted for different k-vertex sizes. Depending on which CG scheme is chosen, it is possible to count defined structural elements, in the form of motifs, directly for every

---

[1] F. Groher, Suess Lab

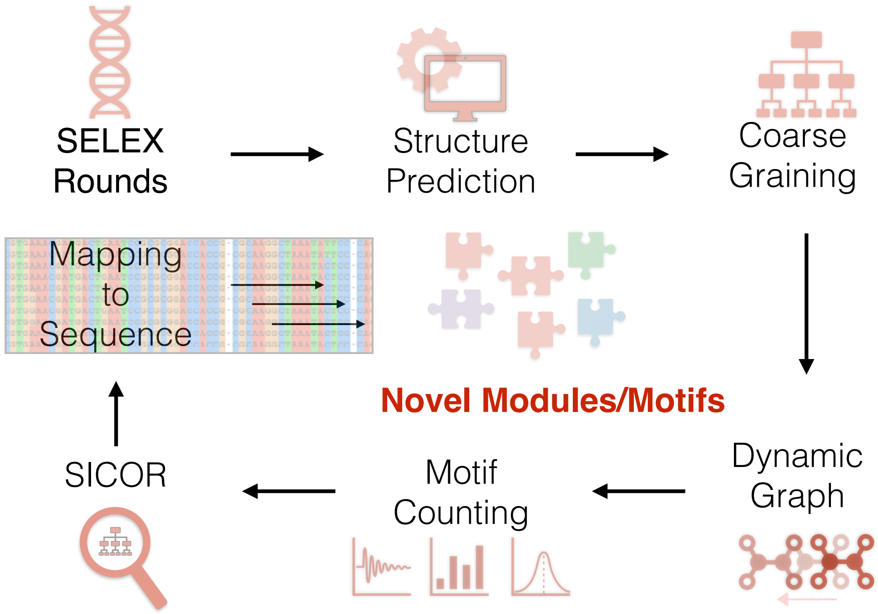


Figure 9.1: Schematic representation of a workflow using Zukers algorithm, coarse-graining, motif counting and the SICOR algorithm to improve *in vitro* selection, SELEX

SELEX rounds. After the successful identification of the functional and structural motif, SICOR can now search for this substructure in the corresponding round. As a result, the sequences that encode the corresponding structure in the SELEX round would be identified. The last step would be to perform experimental studies relating the motif sequence/structure to function to place these structural features into a biological context.

With this work, we established a motif-based semantic suitable to describe processes in which the functional structures of biomolecules need to be assessed. Moreover, these semantics are also suitable for protein, RNA structures and water. It could also be shown that this representation can also be beneficial analyzing HTS (e.g., SELEX, Section 7.3 and Section 7.1) experiments. In addition, we could show in Section 7.1, that combining statistics and structure prediction alone can also improve HTS by limiting the search space for subsequent experiments and screening. Furthermore, it could also be shown that for enzymes (lipases, esterases) classic statistical methods without structure prediction can lead to optimization in terms of assay selection. Especially the combination of structure prediction, simulation and screening can help to solve future problems, especially when it comes to the engineering of molecules which are supposed to have an artificial origin (aptamers, riboswitches) or a synthetic substance (β-Keto ester).

All motif-based methods developed in the context of this dissertation have been integrated into a software package for the programming language `R` (cf. Section 6.4). The package is called `streaMD` and includes a high-level API developed with a focus on enabling fast experimentation and prototyping using dynamic graphs, motif-counts, motif-based entropy and motif-based MSM's in combination with molecular structures as well as MD trajectories.

10 Appendix

10.1 Material to Gromacs GPU Benchmarks

The test system consisted of one HCN1 pores domain 327 DPPC lipids, 1743 TIPS3p, 1743 TIPS3p, 65 K and 245 clones. This results in a total system size of 103329 and particles. Speical thanks to Daniel Bauer and Philipp Babel.

	System	**Performance** [ns/day]
1	4 Threads, i5-4670 @ 3.40 GHz + GTX 1080	14.23
2	8 Threads i7-6700K@4.00 GHz + GTX 1080	20.50
3	32 threads E5-2690@2.90 GHz	10.10
4	8 Threads i7-7700K@4.20 GHz + GTX 1080	23.11
5	16 Threads Ryzen 7 1700X@3.95 GHz + GTX 1080	30.70
6	16 Threads Ryzen 7 1700X@3.90 GHz + Tesla c2070	8.00

Table 10.1: Statistics regarding CPU-GPU benchmarks

10.2 Material to Graph-based Analysis of MD Simulations

k	**2**	**3**	**4**	**5**	**6**	**7**
$\|\mathcal{A}_k\|$	2	8	64	1,024	32,768	2,097,152
$\|\mathcal{A}_k^{con}\|$	1	4	38	827	26,704	1,866,256
$\|\mathcal{M}_k\|$	1	2	6	21	112	853

Table 10.2: Statistics on motifs, connected components and adjacency matrices

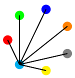
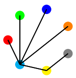
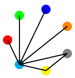
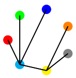
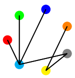
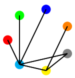
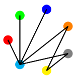
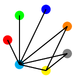
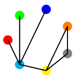
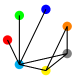
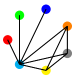
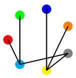
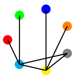
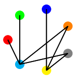
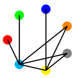
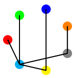
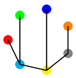
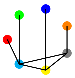
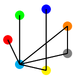
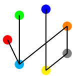
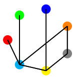
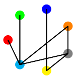
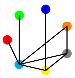
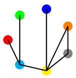
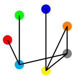
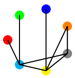
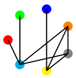
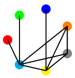
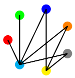
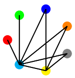
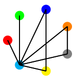
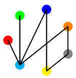
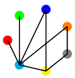
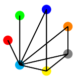
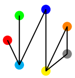
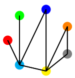
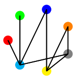
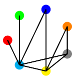
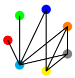
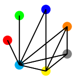
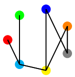
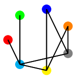
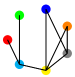
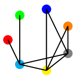
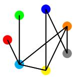
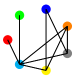
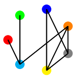
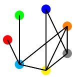
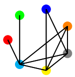
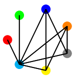
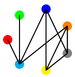
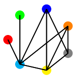
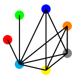
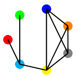
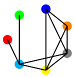
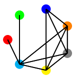
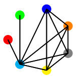
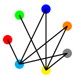
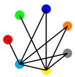
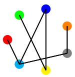
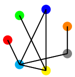
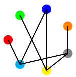
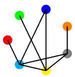
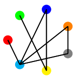
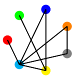
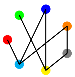
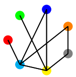
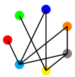
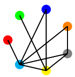
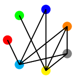
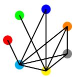
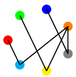
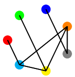
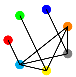
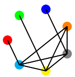
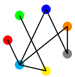
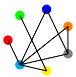
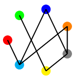
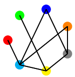
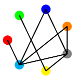
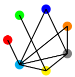
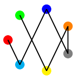
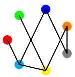
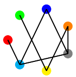
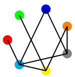
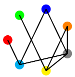
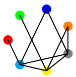
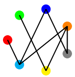
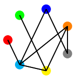
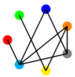
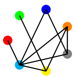
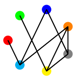
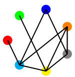
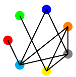
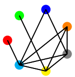
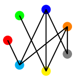
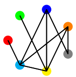
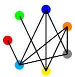
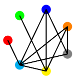
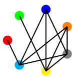
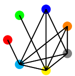
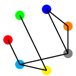
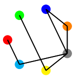
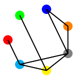
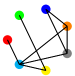
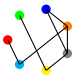
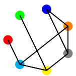
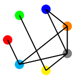
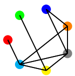
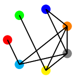
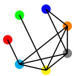
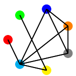
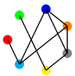
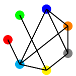
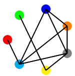
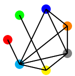
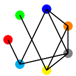
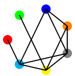
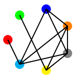
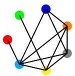
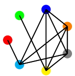
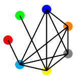
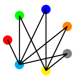
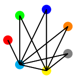
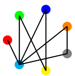
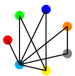
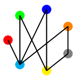
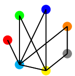
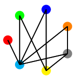
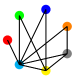
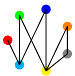
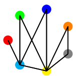
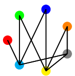
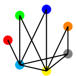
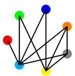
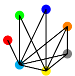
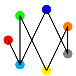
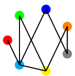
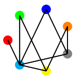
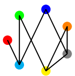
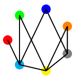
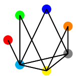
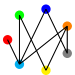
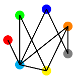
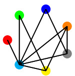
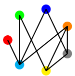
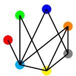
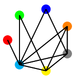
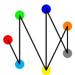
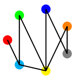
Simulation id	**Temperature** $[K]$	**Density** $[kg/m^3]$	**Entropy** $[J/mol\,K]$
1	273.16	0.9998	61.21
2	285.15	0.9993	65.40
3	298.15	0.9968	68.78
4	313.15	0.9925	72.60
5	333.15	0.9833	77.35
6	353.15	0.9718	82.31
7	373.15	0.9579	86.88
8	423.15	0.9166	96.42
9	473.15	0.8651	105.41
10	523.15	0.7994	113.14
11	573.15	0.7128	121.54
12	613.15	0.6105	128.49
13	633.15	0.5283	132.69
14	647.29	0.3170	141.63
15	633.15	0.1440	150.70
16	613.15	0.0926	154.85
17	573.15	0.0461	160.90
18	523.15	0.0199	167.33
19	473.15	0.0079	174.95
20	423.15	0.0025	184.50
21	373.15	0.000598	195.77
22	353.15	0.000294	202.35
23	333.15	0.00013	207.57
24	313.15	0.0000512	213.86
25	298.15	0.0000231	217.63
26	285.15	0.0000107	222.85
27	273.16	0.00000485	227.89

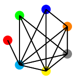
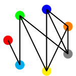
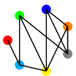
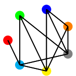
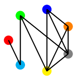
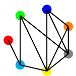
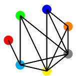
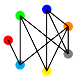
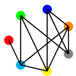
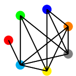
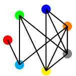
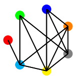
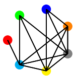
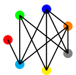
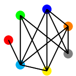
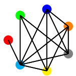
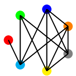
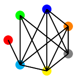
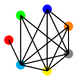
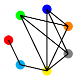
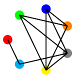
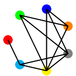
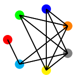
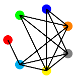
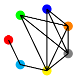
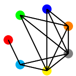
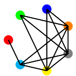
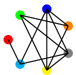
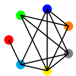
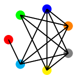
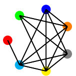
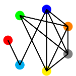
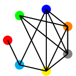
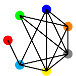
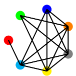
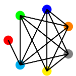
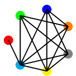
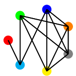
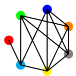
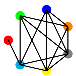
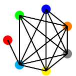
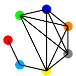
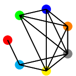
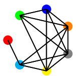
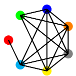
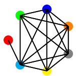
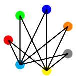
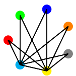
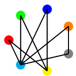
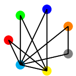
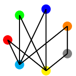
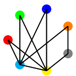
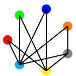
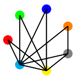
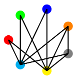
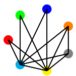
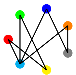
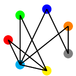
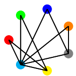
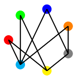
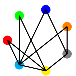
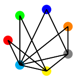
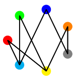
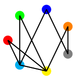
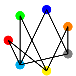
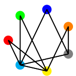
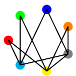
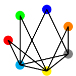
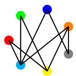
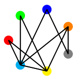
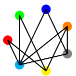
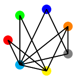
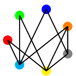
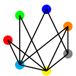
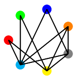
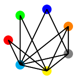
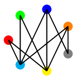
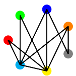
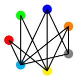
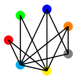
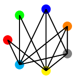
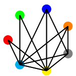
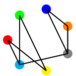
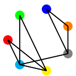
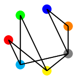
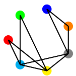
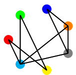
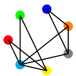
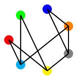
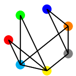
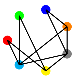
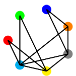
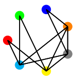
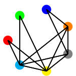
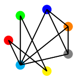
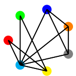
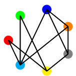
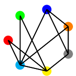
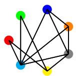
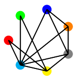
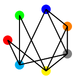
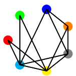
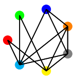
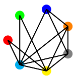
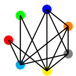
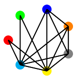
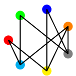
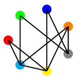
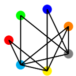
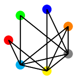
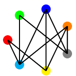
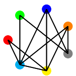
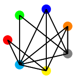
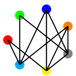
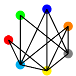
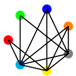
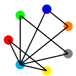
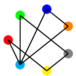
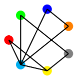
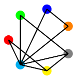
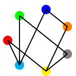
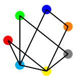
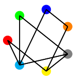
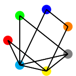
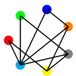
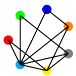
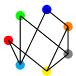
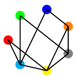
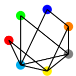
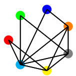
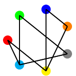
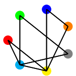
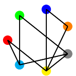
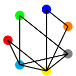
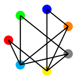
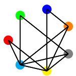
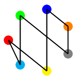
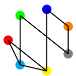
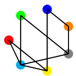
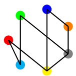
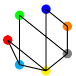
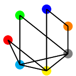
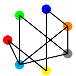
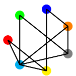
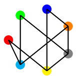
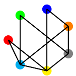
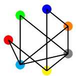
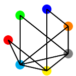
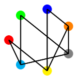
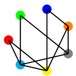
Table 10.3: Thermodynamic Parameters for SPC/E water

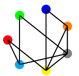
Temperature	$[r', r]$	$\overline{\lvert V\rvert}$	$\overline{\lvert E\rvert}$	$\overline{\lvert B\rvert}$
200 K	$[0.0, 0.3]$	32	14	1.57
	$[0.3, 0.5]$	151	99	15.90
	$[0.5, 0.7]$	366	247	47.35
	$[0.7, 0.9]$	571	475	71.48
	$[0.9, 1.1]$	644	616	83.15
	$[1.1, 1.3]$	786	833	112.81
	$[1.3, 1.5]$	869	1,011	119.66
	$[1.5, 1.7]$	889	1,089	129.28
	$[1.7, 1.9]$	907	1,148	126.31
250 K	$[0.0, 0.3]$	38	10	0.72
	$[0.3, 0.5]$	300	85	15.77
	$[0.5, 0.7]$	601	234	58.75
	$[0.7, 0.9]$	743	442	89.83
	$[0.9, 1.1]$	790	582	111.14
	$[1.1, 1.3]$	879	795	150.08
	$[1.3, 1.5]$	893	998	183.00
	$[1.5, 1.7]$	901	1,089	182.14
	$[1.7, 1.9]$	909	1,152	178.73
270 K	$[0.0, 0.3]$	44	11	1.17
	$[0.3, 0.5]$	267	110	15.69
	$[0.5, 0.7]$	535	300	49.87
	$[0.7, 0.9]$	711	522	83.99
	$[0.9, 1.1]$	785	645	101.53
	$[1.1, 1.3]$	847	848	130.90
	$[1.3, 1.5]$	876	1,021	157.90
	$[1.5, 1.7]$	888	1,110	163.54
	$[1.7, 1.9]$	908	1,136	165.50
350 K	$[0.0, 0.3]$	50	9	1.30
	$[0.3, 0.5]$	282	59	11.23
	$[0.5, 0.7]$	615	249	42.21
	$[0.7, 0.9]$	768	512	70.69
	$[0.9, 1.1]$	826	608	92.59
	$[1.1, 1.3]$	873	837	131.78
	$[1.3, 1.5]$	889	975	148.85
	$[1.5, 1.7]$	891	1,070	156.85
	$[1.7, 1.9]$	911	1,095	157.01

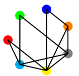
Table 10.4: Graph statistics of water in mineral confinement

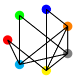
10.3 k-vertex Motif Visualization

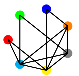

M-7-1

M-7-2

M-7-3

M-7-4

M-7-5

M-7-6

M-7-7

M-7-8

M-7-9

M-7-10

M-7-11

M-7-12

M-7-13

M-7-14

M-7-15

M-7-16

M-7-17

M-7-18

M-7-19

M-7-20

M-7-21

M-7-22

M-7-23

M-7-24

M-7-25

M-7-26

M-7-27

M-7-28

M-7-29

M-7-30

M-7-31

M-7-32

M-7-33

M-7-34

M-7-35

M-7-36

M-7-37

M-7-38

M-7-39

M-7-40

M-7-41

M-7-42

M-7-43

M-7-44

M-7-45

M-7-46

M-7-47

M-7-48

M-7-49

M-7-50

M-7-51

M-7-52

M-7-53

M-7-54

M-7-55

M-7-56

M-7-57

M-7-58

M-7-59

M-7-60

M-7-61

M-7-62

M-7-63

M-7-64

M-7-65

M-7-66

M-7-67

M-7-68

M-7-69

M-7-70

M-7-71

M-7-72

M-7-73

M-7-74

M-7-75

M-7-76

M-7-77

M-7-78

M-7-79

M-7-80

M-7-81

M-7-82

M-7-83

M-7-84

M-7-85

M-7-86

M-7-87

M-7-88

M-7-89

M-7-90

M-7-91

M-7-92

M-7-93

M-7-94

M-7-95

M-7-96

M-7-97

M-7-98

M-7-99

M-7-100

M-7-101

M-7-102

M-7-103

M-7-104

M-7-105

M-7-106

M-7-107

M-7-108

M-7-109

M-7-110

M-7-111

M-7-112

M-7-113

M-7-114

M-7-115

M-7-116

M-7-117

M-7-118

M-7-119

M-7-120

M-7-121

M-7-122

M-7-123

M-7-124

M-7-125

M-7-126

M-7-127

M-7-128

M-7-129

M-7-130

M-7-131

M-7-132

M-7-133

M-7-134

M-7-135

M-7-136

M-7-137

M-7-138

M-7-139

M-7-140

M-7-141

M-7-142

M-7-143

M-7-144

M-7-145

M-7-146

M-7-147

M-7-148

M-7-149

M-7-150

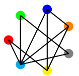

M-7-301

M-7-302

M-7-303

M-7-304

M-7-305

M-7-306

M-7-307

M-7-308

M-7-309

M-7-310

M-7-311

M-7-312

M-7-313

M-7-314

M-7-315

M-7-316

M-7-317

M-7-318

M-7-319

M-7-320

M-7-321

M-7-322

M-7-323

M-7-324

M-7-325

M-7-326

M-7-327

M-7-328

M-7-329

M-7-330

M-7-331

M-7-332

M-7-333

M-7-334

M-7-335

M-7-336

M-7-337

M-7-338

M-7-339

M-7-340

M-7-341

M-7-342

M-7-343

M-7-344

M-7-345

M-7-346

M-7-347

M-7-348

M-7-349

M-7-350

M-7-351

M-7-352

M-7-353

M-7-354

M-7-355

M-7-356

M-7-357

M-7-358

M-7-359

M-7-360

M-7-361

M-7-362

M-7-363

M-7-364

M-7-365

M-7-366

M-7-367

M-7-368

M-7-369

M-7-370

M-7-371

M-7-372

M-7-373

M-7-374

M-7-375

M-7-376

M-7-377

M-7-378

M-7-379

M-7-380

M-7-381

M-7-382

M-7-383

M-7-384

M-7-385

M-7-386

M-7-387

M-7-388

M-7-389

M-7-390

M-7-391

M-7-392

M-7-393

M-7-394

M-7-395

M-7-396

M-7-397

M-7-398

M-7-399

M-7-400

M-7-401

M-7-402

M-7-403

M-7-404

M-7-405

M-7-406

M-7-407

M-7-408

M-7-409

M-7-410

M-7-411

M-7-412

M-7-413

M-7-414

M-7-415

M-7-416

M-7-417

M-7-418

M-7-419

M-7-420

M-7-421

M-7-422

M-7-423

M-7-424

M-7-425

M-7-426

M-7-427

M-7-428

M-7-429

M-7-430

M-7-431

M-7-432

M-7-433

M-7-434

M-7-435

M-7-436

M-7-437

M-7-438

M-7-439

M-7-440

M-7-441

M-7-442

M-7-443

M-7-444

M-7-445

M-7-446

M-7-447

M-7-448

M-7-449

M-7-450

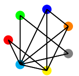

M-7-451

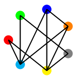

M-7-452

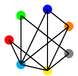

M-7-453

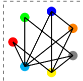

M-7-454

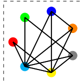

M-7-455

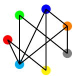

M-7-456

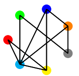

M-7-457

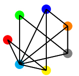

M-7-458

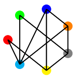

M-7-459

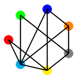

M-7-460

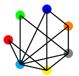

M-7-461

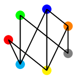

M-7-462

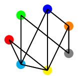

M-7-463

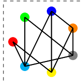

M-7-464

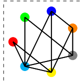

M-7-465

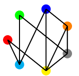

M-7-466

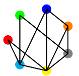

M-7-467

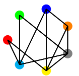

M-7-468

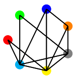

M-7-469

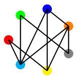

M-7-470

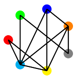

M-7-471

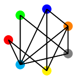

M-7-472

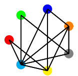

M-7-473

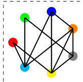

M-7-474

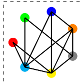

M-7-475

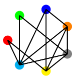

M-7-476

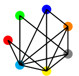

M-7-477

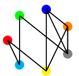

M-7-478

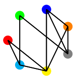

M-7-479

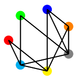

M-7-480

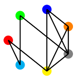

M-7-481

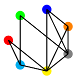

M-7-482

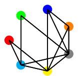

M-7-483

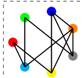

M-7-484

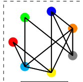

M-7-485

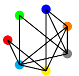

M-7-486

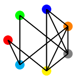

M-7-487

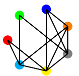

M-7-488

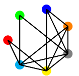

M-7-489

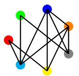

M-7-490

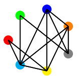

M-7-491

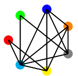

M-7-492

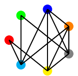

M-7-493

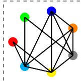

M-7-494

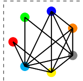

M-7-495

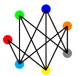

M-7-496

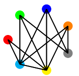

M-7-497

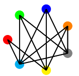

M-7-498

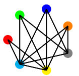

M-7-499

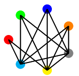

M-7-500

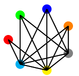

M-7-501

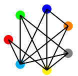

M-7-502

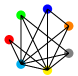

M-7-503

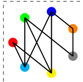

M-7-504

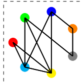

M-7-505

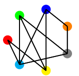

M-7-506

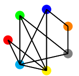

M-7-507

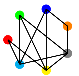

M-7-508

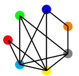

M-7-509

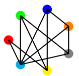

M-7-510

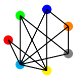

M-7-511

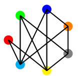

M-7-512

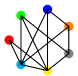

M-7-513

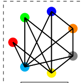

M-7-514

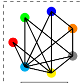

M-7-515

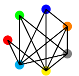

M-7-516

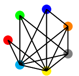

M-7-517

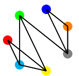

M-7-518

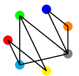

M-7-519

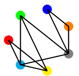

M-7-520

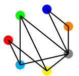

M-7-521

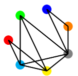

M-7-522

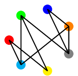

M-7-523

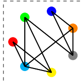

M-7-524

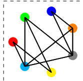

M-7-525

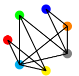

M-7-526

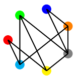

M-7-527

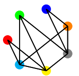

M-7-528

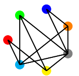

M-7-529

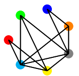

M-7-530

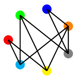

M-7-531

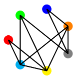

M-7-532

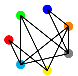

M-7-533

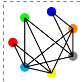

M-7-534

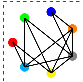

M-7-535

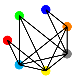

M-7-536

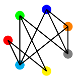

M-7-537

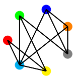

M-7-538

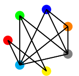

M-7-539

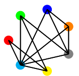

M-7-540

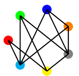

M-7-541

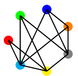

M-7-542

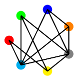

M-7-543

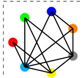

M-7-544

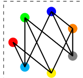

M-7-545

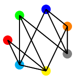

M-7-546

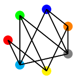

M-7-547

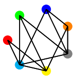

M-7-548

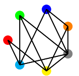

M-7-549

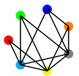

M-7-550

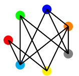

M-7-551

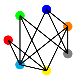

M-7-552

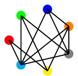

M-7-553

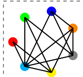

M-7-554

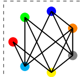

M-7-555

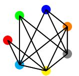

M-7-556

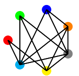

M-7-557

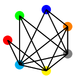

M-7-558

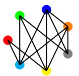

M-7-559

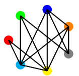

M-7-560

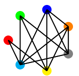

M-7-561

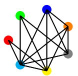

M-7-562

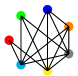

M-7-563

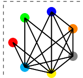

M-7-564

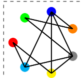

M-7-565

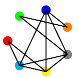

M-7-566

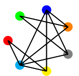

M-7-567

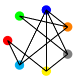

M-7-568

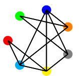

M-7-569

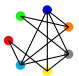

M-7-570

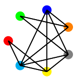

M-7-571

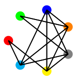

M-7-572

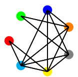

M-7-573

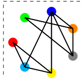

M-7-574

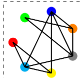

M-7-575

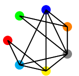

M-7-576

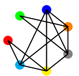

M-7-577

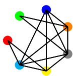

M-7-578

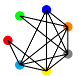

M-7-579

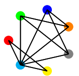

M-7-580

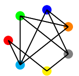

M-7-581

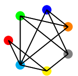

M-7-582

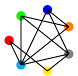

M-7-583

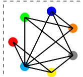

M-7-584

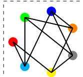

M-7-585

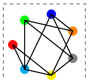

M-7-586

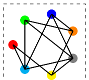

M-7-587

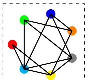

M-7-588

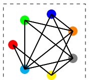

M-7-589

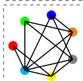

M-7-590

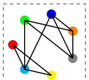

M-7-591

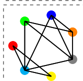

M-7-592

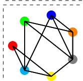

M-7-593

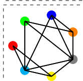

M-7-594

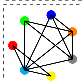

M-7-595

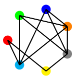
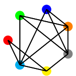
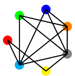
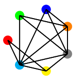
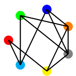
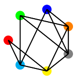
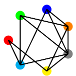
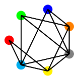
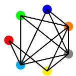
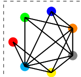
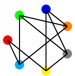
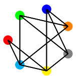
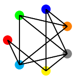
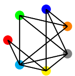
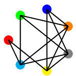
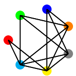
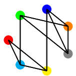
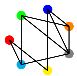
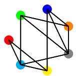
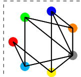
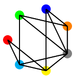
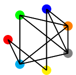
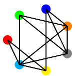
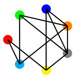
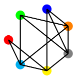
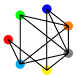
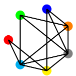
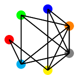
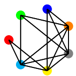
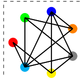
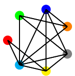
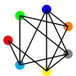
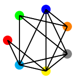
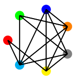
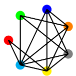
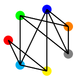
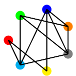
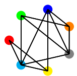
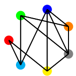
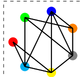
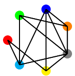
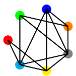
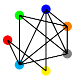
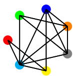
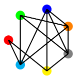
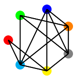
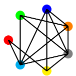
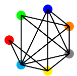
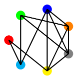
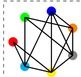
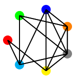
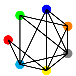
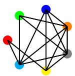
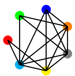
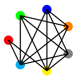
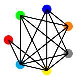
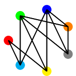
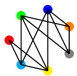
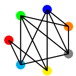
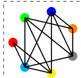
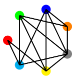
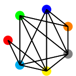
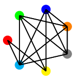
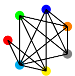
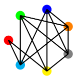
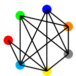
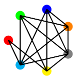
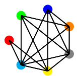
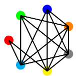
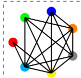
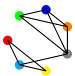
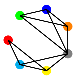
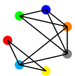
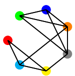
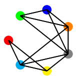
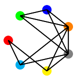
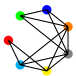
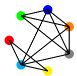
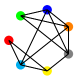
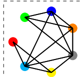
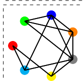
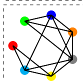
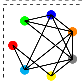
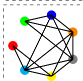
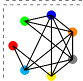
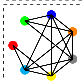
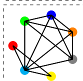
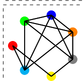
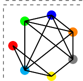
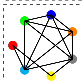
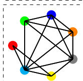
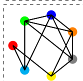
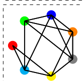
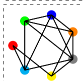
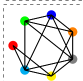
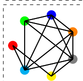
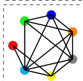
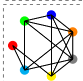
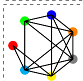
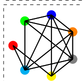
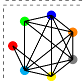
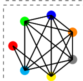
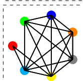
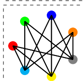
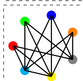
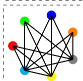
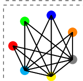
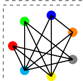
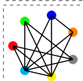
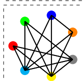
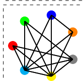
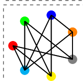
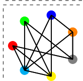
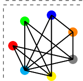
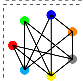
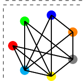
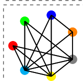
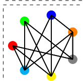
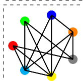
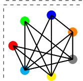
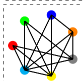
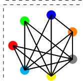
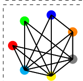
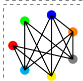
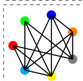
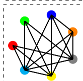
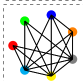
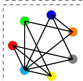
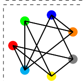
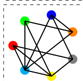
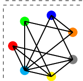
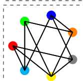
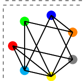
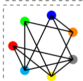
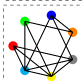
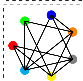
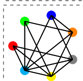
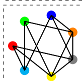
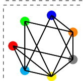
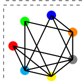
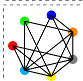
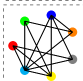
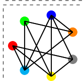
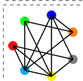
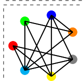
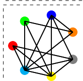
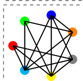
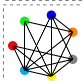
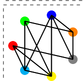
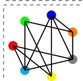

M-7-596


M-7-597


M-7-598


M-7-599


M-7-600


M-7-601

M-7-602

M-7-603

M-7-604

M-7-605

M-7-606

M-7-607

M-7-608

M-7-609

M-7-610

M-7-611

M-7-612

M-7-613

M-7-614

M-7-615

M-7-616

M-7-617

M-7-618

M-7-619

M-7-620

M-7-621

M-7-622

M-7-623

M-7-624

M-7-625

M-7-626

M-7-627

M-7-628

M-7-629

M-7-630

M-7-631

M-7-632

M-7-633

M-7-634

M-7-635

M-7-636

M-7-637

M-7-638

M-7-639

M-7-640

M-7-641

M-7-642

M-7-643

M-7-644

M-7-645

M-7-646

M-7-647

M-7-648

M-7-649

M-7-650

M-7-651

M-7-652

M-7-653

M-7-654

M-7-655

M-7-656

M-7-657

M-7-658

M-7-659

M-7-660

M-7-661

M-7-662

M-7-663

M-7-664

M-7-665

M-7-666

M-7-667

M-7-668

M-7-669

M-7-670

M-7-671

M-7-672

M-7-673

M-7-674

M-7-675

M-7-676

M-7-677

M-7-678

M-7-679

M-7-680

M-7-681

M-7-682

M-7-683

M-7-684

M-7-685

M-7-686

M-7-687

M-7-688

M-7-689

M-7-690

M-7-691

M-7-692

M-7-693

M-7-694

M-7-695

M-7-696

M-7-697

M-7-698

M-7-699

M-7-700

M-7-701

M-7-702

M-7-703

M-7-704

M-7-705

M-7-706

M-7-707

M-7-708

M-7-709

M-7-710

M-7-711

M-7-712

M-7-713

M-7-714

M-7-715

M-7-716

M-7-717

M-7-718

M-7-719

M-7-720

M-7-721

M-7-722

M-7-723

M-7-724

M-7-725

M-7-726

M-7-727

M-7-728

M-7-729

M-7-730

M-7-731

M-7-732

M-7-733

M-7-734

M-7-735

M-7-736

M-7-737

M-7-738

M-7-739

M-7-740

M-7-741

M-7-742

M-7-743

M-7-744

M-7-745

M-7-746

M-7-747

M-7-748

M-7-749

M-7-750

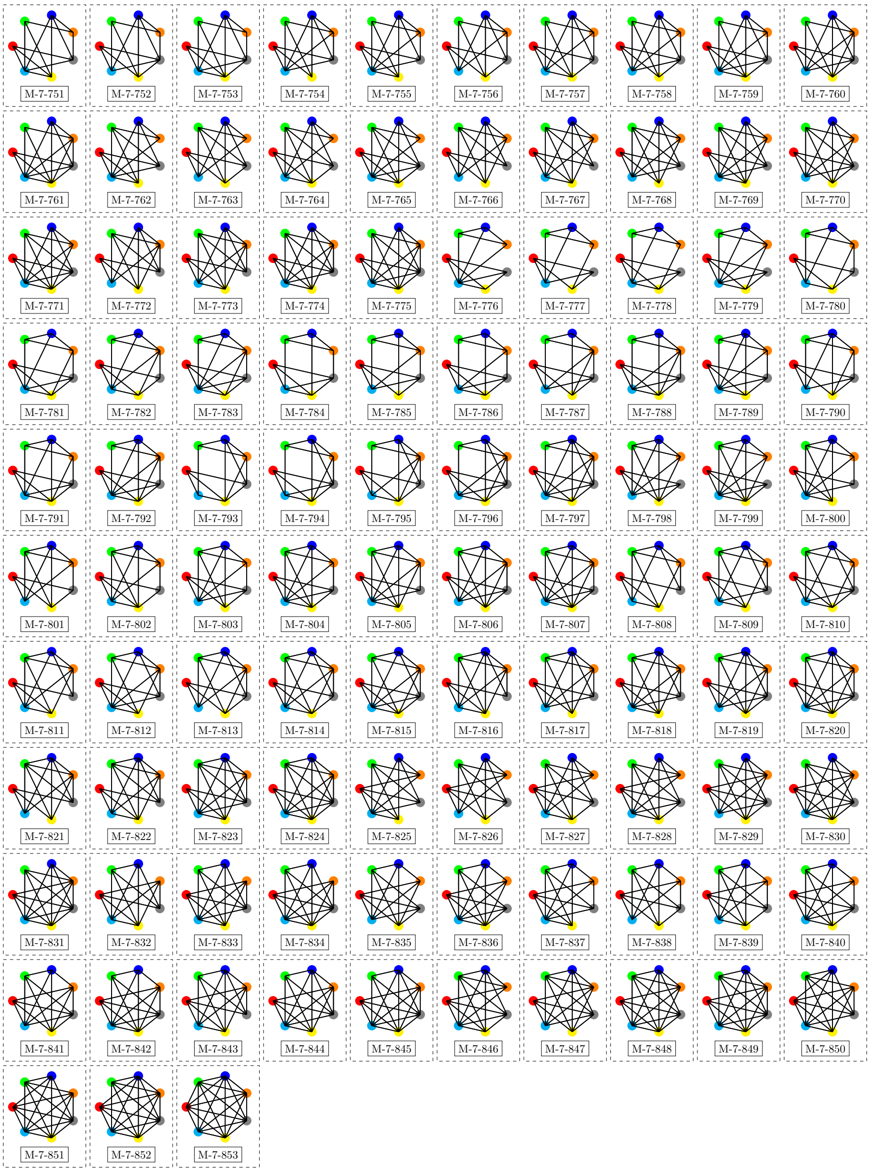


Figure 10.1: Illustration of 7-vertex motifs

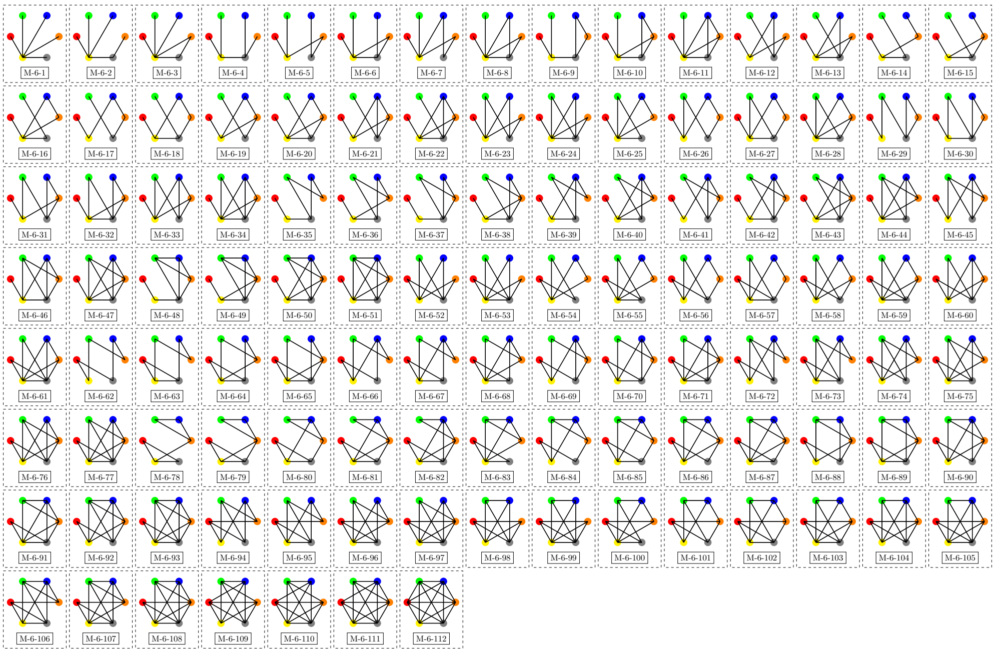


Figure 10.2: Illustration of 6-vertex motifs

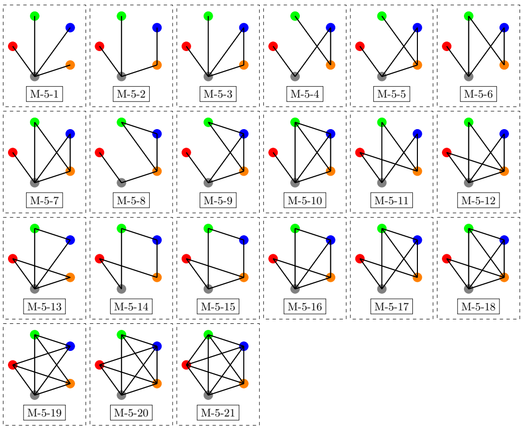


Figure 10.3: Illustration of 5-vertex motifs

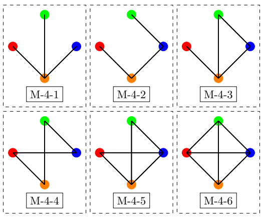


Figure 10.4: Illustration of 4-vertex motifs

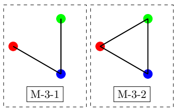


Figure 10.5: Illustration of 3-vertex motifs

10.4 Abbreviations

BP Base Pair (A,T,C,G,U)

MD Molecular Dynamics

MC Monte Carlo

DNA Deoxyribonucleic acid

RNA Ribonucleic acid

PCR Polymerase chain reaction

ASCII American Standard Code for Information Interchange

MFE Minimum free energy

SIMD Single Instruction, Multiple Data

HPC high-performance computing

NGS Next Generation Sequencing

SELEX Systematic Evolution of Ligands by EXponential Enrichment

MLM Machine Learning Models

MSM Markov State Models

SPC Simple Point Charge

HTS High Throuput Screenings

PET Polyethlene Terphtalate

LNCS Linear Constraint Solver

FSC Fusarium Solanie Cutinase

CG Coarse Graining

RMSD Root Mean Square Deviation

HMM Hidden Markov Models

pNB-Est13	para Nitro Butyrate Esterase-13
LD	Levensthein Distance
KS	Kolmogorov Smirnoff
API	Application Programming Interface

List of Tables

10.1 Statistics regarding CPU-GPU benchmarks III
10.2 Statistics on motifs, connected components and adjacency matrices III
10.3 Thermodynamic Parameters for SPC/E water IV
10.4 Graph statistics of water in mineral confinement VI

List of Figures

3.1 GPU Benchmarks 16

4.1 Proposed tertiary structure of the pNB-Est13 23
4.2 Chemical structures of the four nucleotides of DNA/RNA 25
4.3 Chemical structures of different interaction schemes. 25
4.4 Schematic representation of the in vitro selection, SELEX 26
4.5 Decomposition of RNA secondary structure 29
4.6 Different time-scales for simulations and the according experimental methods 32
4.7 $\mathcal{A}_3$ - all adjacency matrices of 3-vertex graphs 38
4.8 Representation of proteins as a dynamic graph 39
4.9 Different types of RNA structures representations 41

6.1 Model of water in mineral confinement with pore dimensions . . 140
6.2 Spatial measures and visualizations of SPC/E water in a box . . 143
6.3 Comparison of thermodynamic entropy and motif-based interactions 145
6.4 Convergence of the number of interactions for different ρ 146
6.5 Motif-based properties depending on pore wall distance r and temperature 148
6.6 Motif-based results for simulations of water in mineral confinement 149
6.7 RMSD and graph-based representation of an helical molecule at two different temperatures 151
6.9 7-vertex motif counts as a function of time for two 100ns simulation of *1hu5* 154
6.10 Visualization of the simulated complex over time 155
6.11 Distance-depended properties of the protein complex over time. 156
6.12 Selected 7-vertex motifs ($\mathcal{M}_7$) 156
6.13 Counts of 3-vertex motif $m_1 \in \mathcal{M}_3$ and 7-vertex motifs $m_{21}, m_{101}, m_{416} \in \mathcal{M}_7$ over time for Complex 157

7.1 All 7-vertex tree motifs 247
7.2 7-vertex motifs on the right and on the left corresponding RNA structural-motifs 248
7.3 Analysis of CFX-SELEX using average motif counts for each SELEX round. Here we consider the size $k \in \{3, 4, 5\}$ 248
7.4 Analysis of CFX-SELEX using average motif counts for each SELEX round. Here we consider the size $k \in \{6, 7\}$ 249
7.5 7-vertex motif class m_4 and m_{17} and the corresponding structural element. 250

9.1 Schematic representation of a workflow using Zukers algorithm, coarse-graining, motif counting and the SICOR algorithm 261

10.1 Illustration of 7-vertex motifs XII
10.2 Illustration of 6-vertex motifs XIII
10.3 Illustration of 5-vertex motifs XIII
10.4 Illustration of 4-vertex motifs XIII
10.5 Illustration of 3-vertex motifs XIV

References for Introduction, Discussion and Summary

(*1*) N. Kim, L. Petingi, and T. Schlick. “Network Theory Tools for RNA Modeling.” In: *WSEAS Transactions on Mathematics.* Vol. 9. Sept. 2013, pp. 941–955.

(*2*) F. H. Arnold. “Design by Directed Evolution.” In: *Accounts of Chemical Research* 31.3 (1998), pp. 125–131. DOI: `10.1021/ar960017f`.

(*3*) M. K. Tiwari et al. “Computational approaches for rational design of proteins with novel functionalities.” In: *Computational and Structural Biotechnology Journal* 2.3 (2012), e201204002. ISSN: 2001-0370. DOI: `10.5936/csbj.201209002`.

(*4*) R. Vianello, C. Domene, and J. Mavri. “The Use of Multiscale Molecular Simulations in Understanding a Relationship between the Structure and Function of Biological Systems of the Brain: The Application to Monoamine Oxidase Enzymes.” In: *Frontiers in Neuroscience* 10 (2016), p. 327. ISSN: 1662-453X. DOI: `10.3389/fnins.2016.00327`.

(*5*) I. Callebaut et al. “Molecular modelling and molecular dynamics of CFTR.” In: *Cellular and Molecular Life Sciences* 74.1 (2017), pp. 3–22. DOI: `10.1007/s00018-016-2385-9`.

(*6*) S. Jager et al. “StreAM-Tg : algorithms for analyzing coarse grained RNA dynamics based on Markov models of connectivity-graphs.” In: *Algorithms for Molecular Biology* 12.1 (2017), p. 15. DOI: `10.1186/s13015-017-0105-0`.

(*7*) C. Groß et al. “Cleavage Product Accumulation Decreases the Activity of Cutinase during PET Hydrolysis.” In: *Journal of Chemical Information and Modeling* 57.2 (2017). PMID: 28128951, pp. 243–255. DOI: `10.1021/acs.jcim.6b00556`.

(8) C. Scholz et al. “DOCKTITE—A Highly Versatile Step-by-Step Workflow for Covalent Docking and Virtual Screening in the Molecular Operating Environment.” In: *Journal of Chemical Information and Modeling* 55.2 (2015). PMID: 25541749, pp. 398–406. DOI: `10.1021/ci500681r`.

(9) K. E. Deigan et al. “Accurate SHAPE-directed RNA structure determination.” In: *Proceedings of the National Academy of Sciences of the United States of America* 106 (2009), pp. 97–102.

(10) J. Wang et al. “Molecular Dynamics Simulation Directed Rational Design of Inhibitors Targeting Drug-Resistant Mutants of Influenza A Virus M2.” In: *Journal of the American Chemical Society* 133.32 (2011), pp. 12834–12841. DOI: `10.1021/ja204969m`.

(11) D. He et al. “Molecular dynamics directed rational design and fluorescence binding assay of phosphopeptide ligands for PLK polo-box domain.” In: *Molecular Simulation* 43.3 (2017), pp. 176–182. DOI: `10.1080/08927022.2016.1244605`. eprint: `http://dx.doi.org/10.1080/08927022.2016.1244605`. URL: `http://dx.doi.org/10.1080/08927022.2016.1244605`.

(12) C. N. Bedbrook et al. “Machine learning to design integral membrane channelrhodopsins for efficient eukaryotic expression and plasma membrane localization.” In: *PLOS Computational Biology* 13.10 (Oct. 2017), pp. 1–21. DOI: `10.1371/journal.pcbi.1005786`.

(13) J. M. Carothers et al. “Model-driven engineering of RNA devices to quantitatively program gene expression.” In: *Science* 334.6063 (2011), pp. 1716–9.

(14) M. Senne, B. Trendelkamp-schroer, and F. Noe. “EMMA: A Software Package for Markov Model Building and Analysis.” In: *Journal of Chemical Theory and Computation* (2012).

(15) D. Osorio, P. Rondón-Villarreal, and R. Torres. “Peptides: A Package for Data Mining of Antimicrobial Peptides.” In: *The R Journal* 7.1 (2015), pp. 4–14. URL: `https://journal.r-project.org/archive/2015/RJ-2015-001/index.html`.

(*16*) D. C. Koboldt et al. “The Next-Generation Sequencing Revolution and Its Impact on Genomics.” In: *Cell* 155.1 (2013), pp. 27–38. DOI: `doi.org/10.1016/j.cell.2013.09.006`.

(*17*) S. Giguère et al. “Machine Learning Assisted Design of Highly Active Peptides for Drug Discovery.” In: *PLOS Computational Biology* 11.4 (Apr. 2015), pp. 1–21. DOI: `10.1371/journal.pcbi.1004074`. URL: `https://doi.org/10.1371/journal.pcbi.1004074`.

(*18*) P. A. Dalby. “Strategy and success for the directed evolution of enzymes.” In: *Current Opinion in Structural Biology* 21.4 (2011). Engineering and design / Membranes, pp. 473–480. ISSN: 0959-440X. DOI: `10.1016/j.sbi.2011.05.003`.

(*19*) J. Inglese et al. “High-throughput screening assays for the identification of chemical probes.” In: *Nature Chemical Biology* 3.8 (2007), pp. 466–479. ISSN: 1552-4450. DOI: `10.1038/nchembio.2007.17`.

(*20*) S. S. Carroll et al. “Drug Screening: Assay Development Issues.” In: *Molecular Cancer Therapeutics*. Ed. by G. C. Prendergast. John Wiley & Sons, Inc., 2004, pp. 119–140. ISBN: 9780471656166. DOI: `10.1002/047165616X.ch7`.

(*21*) J.-L. Reymond, V. S. Fluxa, and N. Maillard. “Enzyme assays.” In: *The Royal Society of Chemistry: Chemical Communications* (1 2008), pp. 10–46. DOI: `10.1039/B813732C`.

(*22*) M. Dörr et al. “Fully automatized high-throughput enzyme library screening using a robotic platform.” In: *Biotechnology and Bioengineering* 113.7 (2016), pp. 1421–1432. ISSN: 1097-0290. DOI: `10.1002/bit.25925`.

(*23*) N. Nguyen Quang, G. Perret, and F. Ducongé. “Applications of High-Throughput Sequencing for In Vitro Selection and Characterization of Aptamers.” In: *Pharmaceuticals* 9.4 (2016), p. 76. ISSN: 1424-8247. DOI: `10.3390/ph9040076`.

(*24*) C. Schneider and B. Suess. “Identification of RNA aptamers with riboswitching properties.” In: *Methods* 97.Supplement C (2016). Nucleic Acid Aptamers, pp. 44–50. DOI: `doi.org/10.1016/j.ymeth.2015.12.001`.

(25) L. Jiang et al. “Saccharide–RNA recognition in a complex formed between neomycin B and an RNA aptamer.” In: *Structure* 7.7 (1999), 817–S7. ISSN: 0969-2126. DOI: `10.1016/S0969-2126(99)80105-1`.

(26) M. Legiewicz and M. Yarus. “A more complex isoleucine aptamer with a cognate triplet.” In: *Journal of Biological Chemistry* 280.20 (2005), pp. 19815–19822. DOI: `10.1074/jbc.M502329200`.

(27) J. E. Weigand et al. “Screening for engineered neomycin riboswitches that control translation initiation.” In: *RNA (New York, N.Y.)* 14.1 (2008), pp. 89–97. DOI: `10.1261/rna.772408`.

(28) M. Legiewicz et al. “Size, constant sequences, and optimal selection.” In: *RNA (New York, N.Y.)* 11.11 (2005), pp. 1701–9. DOI: `10.1261/rna.2161305`.

(29) F. Groher and B. Suess. “In vitro selection of antibiotic-binding aptamers.” In: *Methods* 106.Supplement C (2016). In vitro selection and evolution, pp. 42–50. ISSN: 1046-2023. DOI: `doi.org/10.1016/j.ymeth.2016.05.008`.

(30) C. Berens, F. Groher, and B. Suess. “RNA aptamers as genetic control devices: The potential of riboswitches as synthetic elements for regulating gene expression.” In: *Biotechnology Journal* 10.2 (2015), pp. 246–257. DOI: `10.1002/biot.201300498`.

(31) L. You and F. Arnold. “Directed evolution of subtilisin E in Bacillus subtilis to enhance total activity in aqueous dimethylformamide.” In: *Protein Engineering* 9.1 (1996), pp. 77–83. DOI: `10.1093/protein/9.1.77`. eprint: `http://peds.oxfordjournals.org/content/9/1/77.full.pdf+html`. URL: `http://peds.oxfordjournals.org/content/9/1/77.abstract`.

(32) M. T. Reetz. *High-throughput Screening Systems for Assaying the Enantioselectivity of Enzymes*. Wiley-VCH Verlag GmbH & Co. KGaA, 2005, pp. 41–76. ISBN: 9783527607846. DOI: `10.1002/3527607846.ch2`. URL: `http://dx.doi.org/10.1002/3527607846.ch2`.

(33) C. Schmidt-Dannert and F. H. Arnold. “Directed evolution of industrial enzymes.” In: *Trends in Biotechnology* 17.4 (1999), pp. 135–136. ISSN: 0167-7799. DOI: `http://dx.doi.org/10.1016/S0167-7799(98)`

01283-9. URL: `http://www.sciencedirect.com/science/article/pii/S0167779998012839`.

(*34*) U. T. Bornscheuer. “High-Throughput-Screening Systems for Hydrolases.” In: *Engineering in life sciences* 4.6 (2004), pp. 539–542.

(*35*) M. C. Childers and V. Daggett. “Insights from molecular dynamics simulations for computational protein design.” In: *Molecular System Design and Engineering* 2 (1 2017), pp. 9–33. DOI: `10.1039/C6ME00083E`.

(*36*) K. A. Dill et al. “The Protein Folding Problem.” In: *Annual Review of Biophysics* 37.1 (2008). PMID: 18573083, pp. 289–316. DOI: `10.1146/annurev.biophys.37.092707.153558`.

(*37*) K. G. (Krishna Mohan Poluri. *Protein Engineering Techniques: Gateways to Synthetic Protein Universe*. 1st ed. SpringerBriefs in Applied Sciences and Technology. 2017.

(*38*) M. Hedvat et al. “Selected Approaches for Rational Drug Design and High-Throughput Screening to Identify Anti-Cancer Molecules.” In: *Anticancer Agents Med Chemistry* 12.9 (2012), pp. 1143–1155. DOI: `10.1161/CIRCULATIONAHA.111.087940.The`. arXiv: `NIHMS150003`.

(*39*) D. Bensinger et al. “Elastase-like Activity Is Dominant to Chymotrypsin-like Activity in 20S Proteasome’s beta5 Catalytic Subunit.” In: *ACS Chemical Biology* 11.7 (2016). PMID: 27111844, pp. 1800–1804. DOI: `10.1021/acschembio.6b00023`.

(*40*) S. Fahimi et al. “Developing a visco-hyperelastic material model for 3D finite deformation of elastomers.” In: *Finite Elements in Analysis and Design* 140 (2018), pp. 1–10. ISSN: 0168-874X. DOI: `10.1016/j.finel.2017.10.009`.

(*41*) H. M. Berman et al. “The Protein Data Bank.” In: *Nucleic Acids Research* 28.1 (2000), pp. 235–242. DOI: `10.1093/nar/28.1.235`.

(*42*) H. Fan and A. E. Mark. “Refinement of homology-based protein structures by molecular dynamics simulation techniques.” In: *Protein Science* 13.1 (2004), pp. 211–220. ISSN: 1469-896X. DOI: `10.1110/ps.03381404`.

(*43*) M. R. Lee et al. “Molecular dynamics in the endgame of protein structure prediction.” In: *Journal of Molecular Biology* 313.2 (2001), pp. 417–430. ISSN: 0022-2836. DOI: `10.1006/jmbi.2001.5032`.

(*44*) M. Schmidt and K. Hamacher. “Three-body interactions improve contact prediction within direct-coupling analysis.” In: *Physical Review E* 96 (5 Nov. 2017), p. 10. DOI: `10.1103/PhysRevE.96.052405`.

(*45*) Y. Zhang, A. Kolinski, and J. Skolnick. “TOUCHSTONE II: A New Approach to Ab Initio Protein Structure Prediction.” In: *Biophysical Journal* 85.2 (2017), pp. 1145–1164. ISSN: 0006-3495. DOI: `10.1016/S0006-3495(03)74551-2`.

(*46*) T. Strunk et al. “Structural Model of the Gas Vesicle Protein GvpA and Analysis of GvpA Mutants *in vivo*.” In: *Molecular Microbiology* 81.1 (2011), pp. 56–68.

(*47*) J. Skolnick. “In quest of an empirical potential for protein structure prediction.” In: *Current Opinion in Structural Biology* 16.2 (2006). Theory and simulation/Macromolecular assemblages, pp. 166–171. ISSN: 0959-440X. DOI: `10.1016/j.sbi.2006.02.004`.

(*48*) G. A. Tribello et al. “PLUMED 2: New feathers for an old bird.” In: *Computer Physics Communications* 185.2 (2014), pp. 604–613. ISSN: 0010-4655. DOI: `10.1016/j.cpc.2013.09.018`.

(*49*) N. Kim, K. N. Fuhr, and T. Schlick. “Graph Applications to RNA Structure and Function.” In: *Biophysics of RNA Folding*. Ed. by R. Russell. New York, NY, 2013, pp. 23–51. ISBN: 978-1-4614-4954-6. DOI: `10.1007/978-1-4614-4954-6_3`.

(*50*) Q. R. V. Ferry, R. Lyutova, and T. A. Fulga. “Rational design of inducible CRISPR guide RNAs for de novo assembly of transcriptional programs.” In: *Nature Communications* 8 (2017), p. 14633. ISSN: 2041-1723. DOI: `10.1038/ncomms14633`.

(*51*) M. Zuker and P. Stiegler. “Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.” In: *Nucleic Acids Research* 9.1 (1981), pp. 133–148.

(*52*) R. Lorenz et al. “ViennaRNA Package 2.0.” In: *Algorithms for Molecular Biolog* 6:26 (2011). URL: `http://www.tbi.univie.ac.at/RNA`.

(*53*) A. Wittmann and B. Suess. “Engineered riboswitches: Expanding researchers’ toolbox with synthetic RNA regulators.” In: *FEBS Letters* 586.15 (2012). Synthetic Biology, pp. 2076–2083. ISSN: 0014-5793. DOI: `doi.org/10.1016/j.febslet.2012.02.038`.

(*54*) D. E. Shaw et al. “Millisecond-scale Molecular Dynamics Simulations on Anton.” In: *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*. SC ’09 39. Portland, Oregon: ACM, 2009, p. 11. ISBN: 978-1-60558-744-8. DOI: `10.1145/1654059.1654099`. URL: `http://doi.acm.org/10.1145/1654059.1654099`.

(*55*) D. E. Shaw et al. “Anton 2: Raising the Bar for Performance and Programmability in a Special-purpose Molecular Dynamics Supercomputer.” In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. SC ’14. New Orleans, Louisana: IEEE Press, 2014, p. 13. ISBN: 978-1-4799-5500-8. DOI: `10.1109/SC.2014.9`.

(*56*) K. Hamacher. “Information Theoretical Measures to Analyze Trajectories in Rational Molecular Design.” In: *Journal of Computaional Chemistry* 28 (2007). DOI: `10.1002/jcc.20759`.

(*57*) C. Kutzner et al. “Best bang for your buck: GPU nodes for GROMACS biomolecular simulations.” In: *Journal of Computational Chemistry* 36.26 (2015), pp. 1990–2008. DOI: `10.1002/jcc.24030`.

(*58*) D. E. Shaw. “Anton, a special-purpose machine for molecular dynamics simulation.” In: *ACM* 51.7 (2008). DOI: `10.1145/1364782`.

(*59*) E. Krieger and G. Vriend. “New ways to boost molecular dynamics simulations.” In: *Journal of Computational Chemistry* 36.13 (2015), pp. 996–1007. ISSN: 1096-987X. DOI: `10.1002/jcc.23899`.

(*60*) S. Páll et al. “Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS.” In: *Solving Software Challenges for Exascale: International Conference on Exascale Applications and Software, EASC 2014, Stockholm, Sweden, April 2-3, 2014, Revised Selected Papers*. Ed. by S. Markidis and E. Laure. Cham, 2015, pp. 3–27. ISBN: 978-3-319-15976-8. DOI: `10.1007/978-3-319-15976-8_1`.

(*61*) H. Khandelia and Y. N. Kaznessis. “Molecular dynamics investigation of the influence of anionic and zwitterionic interfaces on antimicrobial peptides’ structure: Implications for peptide toxicity and activity.” In: *Peptides* 27.6 (2006), pp. 1192–1200. ISSN: 0196-9781. DOI: `10.1016/j.peptides.2005.10.022`.

(*62*) K. Moritsugu and J. C. Smith. “REACH Coarse-Grained Biomolecular Simulation: Transferability between Different Protein Structural Classes.” In: *Biophysical Journal* 95.4 (2008), pp. 1639–1648. ISSN: 0006-3495. DOI: `https://doi.org/10.1529/biophysj.108.131714`.

(*63*) A. R. Atilgan et al. “Anisotropy of fluctuation dynamics of proteins with an elastic network model.” In: *Biophysical Journal* 80.1 (2001), pp. 505–515. DOI: `10.1016/S0006-3495(01)76033-X`.

(*64*) K. Hamacher. “Coarse-grained molecular models for high-throughput and multi-scale functional investigations.” In: *Chemistry Central Journal* 2.1 (Mar. 2008), S14. DOI: `10.1186/1752-153X-2-S1-S14`.

(*65*) F. Hoffgaard, P. Weil, and K. Hamacher. “BioPhysConnectoR: Connecting Sequence Information and Biophysical Models.” In: *BMC Bioinformatics* 11 (2010), p. 199. DOI: `10.1186/1471-2105-11-199`.

(*66*) M. Deng and G. E. Karniadakis. “Coarse-Grained Modeling of Protein Unfolding Dynamics.” In: *Multiscale Modeling & Simulation* 12.1 (2014), pp. 109–118. DOI: `10.1137/130921519`.

(*67*) K. Hamacher and J. A. McCammon. “Computing the amino acid specificity of fluctuations in biomolecular systems.” In: *Journal of Chemical Theory and Computation* 2.3 (2006), pp. 873–878. ISSN: 15499618. DOI: `10.1021/ct050247s`.

(*68*) S. J. Marrink et al. “The MARTINI force field: coarse grained model for biomolecular simulations.” In: *Journal of Physical Chemistry B* 111.27 (2007), pp. 7812–7824. DOI: `10.1021/jp071097f`.

(*69*) A. Górecki et al. “RedMD—Reduced molecular dynamics package.” In: *Journal of Computational Chemistry* 30.14 (2009), pp. 2364–2373. ISSN: 1096-987X. DOI: `10.1002/jcc.21223`. URL: `http://dx.doi.org/10.1002/jcc.21223`.

(*70*) Y. Dehouck and A. S. Mikhailov. “Effective Harmonic Potentials: Insights into the Internal Cooperativity and Sequence-Specificity of Protein Dynamics.” In: *PLOS Computational Biology* 9.8 (Aug. 2013), pp. 1–11. DOI: `10.1371/journal.pcbi.1003209`.

(*71*) J. M. Berg, J. L. Tymoczko, and L. Stryer. “Molekulare Motoren.” In: *Stryer Biochemie*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 1025–1048. ISBN: 978-3-8274-2989-6. DOI: `10.1007/978-3-8274-2989-6_35`.

(*72*) R. A. Sheldon. “Fundamentals of green chemistry: efficiency in reaction design.” In: *Chemical Society Reviews* 41.4 (2012), pp. 1437–1451. ISSN: 0306-0012. DOI: `10.1039/C1CS15219J`.

(*73*) J.-M. Claverie. “Life from an RNA World: The Ancestor Within by Michael Yarus.” In: *The Quarterly Review of Biology* 87.1 (2012), pp. 65–66. DOI: `10.1086/663920`.

(*74*) K.-M. Song, S. Lee, and C. Ban. “Aptamers and Their Biological Applications.” In: *Sensors* 12.1 (2012), pp. 612–631. ISSN: 1424-8220. DOI: `10.3390/s120100612`.

(*75*) D. E. Cameron, C. J. Bashor, and J. J. Collins. “A brief history of synthetic biology.” In: *Nature reviews. Microbiology* 12.5 (2014), pp. 381–90.

(*76*) A. D. Mcnaught and A. Wilkinson. *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book")*. WileyBlackwell; 2nd Revised edition edition, 1997. ISBN: 978-0865426849.

(*77*) Schrödinger, LLC. “The PyMOL Molecular Graphics System, Version 1.3r1.” 2010.

(*78*) S. Subramaniam et al. “Resolution advances in cryo-EM enable application to drug discovery.” In: *Current Opinion in Structural Biology* 41 (2016). Multi-protein assemblies in signaling • Catalysis and regulation, pp. 194–202. ISSN: 0959-440X. DOI: `10.1016/j.sbi.2016.07.009`. URL: `http://www.sciencedirect.com/science/article/pii/S0959440X16300847`.

(*79*) K. A. Dill and H. S. Chan. “From Levinthal to pathways to funnels.” In: *Nature Structure Molecular Biology* 4.1 (Jan. 1997), pp. 10–19. DOI: `10.1038/nsb0197-10`.

(*80*) D. J. Rigden. *From Protein Structure to Function with Bioinformatics.* Springer, Dordrecht, 2009. ISBN: 978-1-4020-9057-8. DOI: `10.1007/978-1-4020-9058-5`.

(*81*) K. Hoogsteen. "The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine." In: *Acta Crystallographica* 16.9 (Sept. 1963), pp. 907–916. DOI: `10.1107/S0365110X63002437`.

(*82*) F. H. C. Crick et al. "Codon-anticodon pairing: The wobble hypothesis." In: *Journal of Molecular Biology* (1966), pp. 548–555.

(*83*) S. Hanson et al. "Molecular analysis of a synthetic tetracycline-binding riboswitch." In: *RNA* (2005), pp. 2549–2556.

(*84*) D. Ellington and J. W. Szostak. "In vitro selection of RNA molecules that bind specific ligands." In: *Nature* 346.6287 (1990), pp. 818–22. DOI: `10.1038/346818a0`. arXiv: `0801.3609`.

(*85*) A. Wittmann and B. Suess. "Engineered riboswitches: Expanding researchers' toolbox with synthetic RNA regulators." In: *FEBS Letters* 586.15 (2012), pp. 2076–2083. ISSN: 1873-3468. DOI: `10.1016/j.febslet.2012.02.038`.

(*86*) C. Berens, A. Thain, and R. Schroeder. "A tetracycline-binding RNA aptamer." In: *Bioorganic and Medicinal Chemistry* 9.10 (2001), pp. 2549–2556.

(*87*) C. Tuerk and L. Gold. "Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase." In: *Science* 249.4968 (1990), pp. 505–510. ISSN: 0036-8075. DOI: `10.1126/science.2200121`.

(*88*) Z. Zhuo et al. "Recent Advances in SELEX Technology and Aptamer Applications in Biomedicine." In: *International Journal of Molecular Sciences* 18.2142 (2017). ISSN: 1422-0067. DOI: `10.3390/ijms18102142`. URL: `http://www.mdpi.com/1422-0067/18/10/2142`.

(*89*) C. Laing and T. Schlick. "Computational approaches to RNA structure prediction, analysis, and design." In: *Current Opinion in Structural Biology* 21.3 (2011), pp. 306–318.

(*90*) T. Liu et al. “RNA Secondary Structure Prediction Using Extreme Learning Machine with Clustering Under-Sampling Technique.” In: *Proceedings of ELM-2015 Volume 2: Theory, Algorithms and Applications (II)*. Ed. by J. Cao et al. 2016, pp. 317–324. ISBN: 978-3-319-28373-9. DOI: `10.1007/978-3-319-28373-9_27`.

(*91*) Q. Su, J. Jiang, and Y. Fu. “A Hardware Implementation of Nussinov RNA Folding Algorithm.” In: *Computer Engineering and Technology: 16th National Conference, NCCET 2012, Shanghai, China, August 17-19, 2012, Revised Selected Papers*. Ed. by W. Xu et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 84–91. ISBN: 978-3-642-35898-2. DOI: `10.1007/978-3-642-35898-2_10`.

(*92*) B. A. Shapiro et al. “The massively parallel genetic algorithm for RNA folding: MIMD implementation and population variation.” In: *Bioinformatics* 17.2 (2001), pp. 137–148. DOI: `10.1093/bioinformatics/17.2.137`.

(*93*) J. McCaskill. “The equilibrium partition function and base pair binding probabilities for RNA secondary structure.” In: *Biopolymers* 29.6-7 (1990), pp. 1105–1119.

(*94*) D. J. Evers. “RNA folding via algebraic dynamic programming.” PhD thesis. Bielefeld University, 2003.

(*95*) S. Bernhart et al. “RNAalifold: Improved consensus structure prediction for RNA alignments.” In: *BMC Bioinformatics* 9.1 (2008), p. 474.

(*96*) I. Hofacker et al. “Fast folding and comparison of RNA secondary structures.” In: *Monatshefte für Chemie/Chemical Monthly* 125.2 (1994), pp. 167–188.

(*97*) P. Kerpedjiev et al. “Predicting RNA 3D structure using a coarse-grain helix-centered model.” In: *Rna* 21.6 (2015), pp. 1110–1121. DOI: `10.1261/rna.047522.114`.

(*98*) R. B. Lyngsø and C. N. S. Pedersen. “RNA Pseudoknot Prediction in Energy-Based Models.” In: *Journal of Computational Biology* 7.3-4 (Aug. 2000), pp. 409–427. DOI: `10.1089/106652700750050862`.

(*99*) E. Rivas and S. R. Eddy. "A dynamic programming algorithm for RNA structure prediction including pseudoknots11Edited by I. Tinoco." In: *Journal of Molecular Biology* 285.5 (1999), pp. 2053–2068. ISSN: 0022-2836. DOI: `doi.org/10.1006/jmbi.1998.2436`.

(*100*) D. Mathews et al. "Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure." In: *Journal of Molecular Biology* 288.5 (1999), pp. 911–940.

(*101*) D. Mathews et al. "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure." In: *Proceedings of the National Academy of Sciences of the United States of America* 101.19 (2004), p. 7287.

(*102*) D. Turner and D. Mathews. "NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure." In: *Nucleic Acids Research* 38.suppl 1 (2010), pp. D280–D282.

(*103*) R. Dimitrov and M. Zuker. "Prediction of hybridization and melting for double-stranded nucleic acids." In: *Biophysical Journal* 87.1 (2004), pp. 215–226.

(*104*) B. R. Brooks et al. "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations." In: *Journal of Computational Chemistry* 4.2 (1983), pp. 187–217. ISSN: 1096-987X.

(*105*) M. J. Boniecki et al. "SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction." In: *Nucleic Acids Research* 44.7 (2016), e63. DOI: `10.1093/nar/gkv1479`.

(*106*) M. Parisien and F. Major. "The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data." In: *Nature* 452.7183 (2008), pp. 51–55.

(*107*) A. Russo et al. "In Silico Generation of Peptides by Replica Exchange Monte Carlo: Docking-Based Optimization of Maltose-Binding-Protein Ligands." In: *PLOS ONE* 10.8 (Aug. 2015), pp. 1–16. DOI: `10.1371/journal.pone.0133571`. URL: `https://doi.org/10.1371/journal.pone.0133571`.

(*108*) O. Zimmermann and U. H. Hansmann. “Understanding protein folding: Small proteins in silico.” In: *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* 1784.1 (2008). Inhibitors of Protein Kinases (5th International Conference, IPK-2007) and Workshop Session on Molecular Design and Simulation Methods (Warsaw, Poland, June 23-27, 2007), pp. 252–258. ISSN: 1570-9639. DOI: `http://dx.doi.org/10.1016/j.bbapap.2007.10.010`.

(*109*) B. J. Alder and T. E. Wainwright. “Studies in Molecular Dynamics.” In: *Journal Chemical Physics* 30 (1959), pp. 459–466.

(*110*) M. P. Allen. “Introduction to molecular dynamics simulation.” In: *Computational Soft Matter: From Synthetic Polymers to Proteins, Lecture Notes, Norbert Attig, Kurt Binder, Helmut Grubmuller, Kurt Kremer (Eds.), John von Neumann Institute for Computing, Julich* (2004).

(*111*) G. Ortega, M. Pons, and O. Millet. “Chapter Six - Protein Functional Dynamics in Multiple Timescales as Studied by NMR Spectroscopy.” In: *Dynamics of Proteins and Nucleic Acids.* Ed. by T. Karabencheva-Christova. Vol. 92. Advances in Protein Chemistry and Structural Biology. 2013, pp. 219–251. DOI: `10.1016/B978-0-12-411636-8.00006-7`.

(*112*) J.-P. Hansen and I. R. McDonald. *Theory of Simple Liquids (Third Edition).* third. Elsevier, 2006.

(*113*) M. Rovere. “Lecture notes on Monte Carlo and Molecular Dynamics Simulations.” In: *School of Neutron Scattering "F. P. Ricci", Santa Margherita di Pula, 22 Sep.-3 Oct.* (2008).

(*114*) D. van der Spoel et al. *Gromacs User Manual version 4.5.6.* 2010.

(*115*) H. Berendsen, D. van der Spoel, and R. van Drunen. “GROMACS: A message-passing parallel molecular dynamics implementation.” In: *Computer Physics Communications* 91.1 (1995), pp. 43–56. ISSN: 0010-4655. DOI: `http://dx.doi.org/10.1016/0010-4655(95)00042-E`.

(*116*) E. Lindahl, B. Hess, and D. van der Spoel. “GROMACS 3.0: a package for molecular simulation and trajectory analysis.” In: *Molecular Modeling Annual* 7.8 (2001), pp. 306–317. ISSN: 0948-5023. DOI: `10.1007/s008940100045`.

(*117*) D. Van Der Spoel et al. “GROMACS: Fast, flexible, and free.” In: *Journal of Computational Chemistry* 26.16 (2005), pp. 1701–1718. ISSN: 1096-987X. DOI: `10.1002/jcc.20291`.

(*118*) B. Hess, H. Bekker, and H. J. C. Berendsen. “LINCS: A linear constraint solver for molecular simulations.” In: *Journal of Computational Chemistry* 18.12 (1997), pp. 1463–1472.

(*119*) W. G. Hoover. “Canonical dynamics: Equilibrium phase-space distributions.” In: *Physical Review A* 31.3 (1985), p. 1695.

(*120*) M. Parrinello and A. Rahman. “Polymorphic transitions in single crystals: A new molecular dynamics method.” In: *Journal of Applied Physics* 52 (1981), p. 7182. DOI: `10.1063/1.328693`.

(*121*) S. Nosé and M. L. Klein. “Constant pressure molecular dynamics for molecular systems.” In: *Molecular Physics* 50.5 (1983), pp. 1055–1076. DOI: `10.1080/00268978300102851`.

(*122*) M. Parrinello and A. Rahman. “Crystal Structure and Pair Potentials: A Molecular-Dynamics Study.” In: *Physical Review Letters* 45 (14 Oct. 1980), pp. 1196–1199. DOI: `10.1103/PhysRevLett.45.1196`.

(*123*) R. Milo et al. “Network Motifs: Simple Building Blocks of Complex Networks.” In: *Science* 298.5594 (2002), pp. 824–827. DOI: `10.1126/science.298.5594.824`.

(*124*) A. R. Benson, D. F. Gleich, and J. Leskovec. “Higher-order organization of complex networks.” In: *Science* 353.6295 (2016), pp. 163–166. DOI: `10.1126/science.aad9029`.

(*125*) S. Panni and S. E. Rombo. “Searching for repetitions in biological networks: methods, resources and tools.” In: *Briefings in Bioinformatics* 16.1 (2015), pp. 118–136.

(*126*) B. Schiller et al. “StreaM - A Stream-Based Algorithm for Counting Motifs in Dynamic Graphs.” In: *Algorithms for Computational Biology - Second International Conference, AlCoB 2015, Mexico City, Mexico, August 4-5, 2015, Proceedings.* Ed. by A. H. Dediu et al. Vol. 9199. Lecture Notes in Computer Science. Springer, 2015, pp. 53–67. ISBN: 978-3-319-21232-6. DOI: `10.1007/978-3-319-21233-3{\_}5`.

(*127*) B. A. Shapiro et al. “Bridging the gap in RNA structure prediction.” In: *Current Opinion in Structural Biology* 17.2 (2007), pp. 157–165.

(*128*) B. Schiller. “Graph-based Analysis of Dynamic Systems.” PhD thesis. TU Dresden, 2016.

(*129*) B. Schiller, J. Castrillon, and T. Strufe. “Efficient data structures for dynamic graph analysis.” In: *Proceedings of the 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. Ed. by L. O’Conner. SITIS 2015. Bangkok, Thailand: IEEE Computer Society, Nov. 2015, pp. 497–504. DOI: `10.1109/SITIS.2015.94`.

(*130*) M. A. Alvarez and C. Yan. “A new protein graph model for function prediction.” In: *Computational biology and chemistry* 37 (2012), pp. 6–10.

(*131*) N. C. Benson and V. Daggett. “A chemical group graph representation for efficient high-throughput analysis of atomistic protein simulations.” In: *Journal of Bioinformatics and Computational Biology* 10.04 (2012).

(*132*) S. Vishveshwara, K. Brinda, and N. Kannan. “Protein structure: insights from graph theory.” In: *Journal of Theoretical and Computational Chemistry* 1.01 (2002), pp. 187–211.

(*133*) Y. Yan, S. Zhang, and F.-X. Wu. “Applications of graph theory in protein structure identification.” In: *Proteome Science* 9.Suppl 1 (2011), S17.

(*134*) X. Daura et al. “Reversible peptide folding in solution by molecular dynamics simulation.” In: *Journal of Molecular Biology* 280.5 (1998), pp. 925–932.

(*135*) E. A. Coutsias, C. Seok, and K. A. Dill. “Using quaternions to calculate RMSD.” In: *Journal of Computational chemistry* 25.15 (2004), pp. 1849–1857.

(*136*) Z. Bikadi and E. Hazai. “Application of the PM6 semi-empirical method to modeling proteins enhances docking accuracy of AutoDock.” In: *Journal of Cheminformatics* 1.1 (2009), p. 1.

(*137*) T.-H. Chiang, D. Hsu, and J.-C. Latombe. “Markov dynamic models for long-timescale protein motion.” In: *Bioinformatics* 26.12 (2010), pp. i269–i277.

(*138*) B. Knapp et al. “Is an intuitive convergence definition of molecular dynamics simulations solely based on the root mean square deviation possible?” In: *Journal of Computational Biology : a journal of computational molecular cell biology* 18.8 (2011), pp. 997–1005. ISSN: 1557-8666. DOI: `10.1089/cmb.2010.0237`.

(*139*) P. Tijerina and R. Russell. *Biophysics of RNA Folding.* January 2013. 2013, pp. 205–230. ISBN: 978-1-4614-4953-9. DOI: `10.1007/978-1-4614-4954-6`.

(*140*) P. Dao et al. “AptaTRACE Elucidates RNA Sequence-Structure Motifs from Selection Trends in HT-SELEX Experiments.” In: *Cell Systems* 3.1 (2016), pp. 62–70. ISSN: 2405-4712. DOI: `doi.org/10.1016/j.cels.2016.07.003`.

(*141*) M. A. Ditzler et al. “High-Throughput sequence analysis reveals structural diversity and improved potency among RNA inhibitors of HIV reverse transcriptase.” In: *Nucleic Acids Research* 41.3 (2013), pp. 1873–1884. DOI: `10.1093/nar/gks1190`.

(*142*) A. R. Gawronski and M. Turcotte. “RiboFSM: Frequent subgraph mining for the discovery of RNA structures and interactions.” In: *BMC Bioinformatics* 15.13 (2014), S2. ISSN: 1471-2105. DOI: `10.1186/1471-2105-15-S13-S2`.

(*143*) B. Shapiro. “An algorithm for comparing multiple RNA secondary structures.” In: *Computer Applications in the Biosciences: CABIOS* 4.3 (1988), pp. 387–393.

(*144*) B. Shapiro and K. Zhang. “Comparing multiple RNA secondary structures using tree comparisons.” In: *Computer Applications in the Biosciences: CABIOS* 6.4 (1990), pp. 309–318.

(*145*) R. Lorenz et al. “ViennaRNA Package 2.0.” In: *Algorithms for Molecular Biology* 6.1 (Nov. 2011), p. 26. ISSN: 1748-7188. DOI: `10.1186/1748-7188-6-26`.

(*146*) I. Hofacker, M. Fekete, and P. Stadler. “Secondary structure prediction for aligned RNA sequences.” In: *Journal of Molecular Biology* 319.5 (2002), pp. 1059–1066.

(*147*) W. Fontana et al. “RNA folding and combinatory landscapes.” In: *Physical Review E* 47.3 (1993), p. 2083.

(*148*) W. Fontana et al. “Statistics of RNA secondary structures.” In: *Biopolymers* 33.9 (1993), pp. 1389–1404.

(*149*) S. G. Kobourov. “Spring Embedders and Force Directed Graph Drawing Algorithms.” In: *arXiv preprint arXiv:1201.3011* (2012), pp. 1–23. arXiv: `1201.3011`. URL: `http://arxiv.org/abs/1201.3011`.

(*150*) S. Schirmer and R. Giegerich. “Forest Alignment with Affine Gaps and Anchors.” In: *Combinatorial Pattern Matching: 22nd Annual Symposium, CPM 2011, Palermo, Italy, June 27-29, 2011. Proceedings.* Ed. by R. Giancarlo and G. Manzini. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 104–117. ISBN: 978-3-642-21458-5. DOI: `10.1007/978-3-642-21458-5_11`.

(*151*) T. Schlick. “Mathematical and Biological Scientists Assess the State of the Art in RNA Science At an Ima Workshop, RNA in Biology, Bioengineering, and Biotechnology.” In: *International Journal for Multiscale Computational Engineering* 8.4 (2010), pp. 369–378.

(*152*) M. Ikeguchi et al. “Protein Structural Change Upon Ligand Binding: Linear Response Theory.” In: *Physical Review Letters* 94.7 (2005).

(*153*) N. Xiao. *ggsci: Scientific Journal and Sci-Fi Themed Color Palettes for 'ggplot2'.* R package version 2.8. 2017. URL: `https://CRAN.R-project.org/package=ggsci`.

(*154*) M. J. Dombrowsky et al. “StreaMD: Efficient analysis of Molecular Dynamics using R as an Integration Tool.” revision to JCC. 2017.

(*155*) H. Wickham. *ggplot2: Elegant Graphics for Data Analysis.* 2009. ISBN: 978-0-387-98140-6. URL: `http://ggplot2.org`.

(*156*) C. O. Wilke. *cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'.* R package version 0.7.0. 2016. URL: `https://CRAN.R-project.org/package=cowplot`.

(*157*) H. J. C. Berendsen et al. “Molecular dynamics with coupling to an external bath.” In: *The Journal of Chemical Physics* 81.8 (1984), pp. 3684–3690. DOI: `10.1063/1.448118`.

(*158*) S. Pronk et al. “GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit.” In: *Bioinformatics* 29.7 (2013), pp. 845–854. DOI: `10.1093/bioinformatics/btt055`. eprint: `http://bioinformatics.oxfordjournals.org/content/29/7/845.full.pdf+html`. URL: `http://bioinformatics.oxfordjournals.org/content/29/7/845.abstract`.

(*159*) C. Allolio et al. “Cover Picture: Ab Initio H2O in Realistic Hydrophilic Confinement (ChemPhysChem 18/2014).” In: *ChemPhysChem* 15.18 (2014), pp. 3881–3881. ISSN: 1439-7641. DOI: `10.1002/cphc.201490089`.

(*160*) M. F. Harrach et al. “Effect of the hydroaffinity and topology of pore walls on the structure and dynamics of confined water.” In: *The Journal of Chemical Physics* 142.3, 034703 (2015). DOI: `http://dx.doi.org/10.1063/1.4905557`. URL: `http://scitation.aip.org/content/aip/journal/jcp/142/3/10.1063/1.4905557`.

(*161*) D. Helmer et al. “Rational design of a peptide capture agent for CXCL8 based on a model of the CXCL8:CXCR1 complex.” In: *Royal Society Advances* 5 (33 2015), pp. 25657–25668. DOI: `10.1039/C4RA13749C`.

(*162*) B. Webb and A. Sali. “Comparative Protein Structure Modeling Using MODELLER.” In: *Current Protocols in Bioinformatics.* John Wiley and Sons, Inc., 2002. Chap. 5.6.1-5.6.32. ISBN: 9780471250951. DOI: `10.1002/0471250953.bi0506s47`. URL: `http://dx.doi.org/10.1002/0471250953.bi0506s47`.

(*163*) K. B. Koziara et al. “Testing and validation of the Automated Topology Builder (ATB) version 2.0: prediction of hydration free enthalpies.” In: *Journal of Computer-Aided Molecular Design* 28.3 (2014), pp. 221–233. ISSN: 1573-4951. DOI: `10.1007/s10822-014-9713-7`. URL: `http://dx.doi.org/10.1007/s10822-014-9713-7`.

(*164*) H. J. C. Berendsen et al. “Molecular dynamics with coupling to an external bath.” In: *The Journal of Chemical Physics* 81.8 (1984), pp. 3684–3690. DOI: `http://dx.doi.org/10.1063/1.448118`.

(*165*) B. Jana et al. “Entropy of Water in the Hydration Layer of Major and Minor Grooves of DNA.” In: *The Journal of Physical Chemistry B* 110.39 (2006). PMID: 17004828, pp. 19611–19618. DOI: `10.1021/`

`jp061588k`. eprint: `http://dx.doi.org/10.1021/jp061588k`. URL: `http://dx.doi.org/10.1021/jp061588k`.

(*166*) S. Cheluvaraja and H. Meirovitch. “Calculation of the Entropy and Free Energy from Monte Carlo Simulations of a Peptide Stretched by an External Force.” In: *The Journal of Physical Chemistry B* 109.46 (2005). PMID: 16853854, pp. 21963–21970. DOI: `10.1021/jp052969l`. eprint: `http://dx.doi.org/10.1021/jp052969l`. URL: `http://dx.doi.org/10.1021/jp052969l`.

(*167*) R. Baron, P. H. Hünenberger, and J. A. McCammon. “Absolute Single-Molecule Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction Terms and Convergence Properties.” In: *Journal of Chemical Theory and Computation* 5.12 (2009). PMID: 20011626, pp. 3150–3160. DOI: `10.1021/ct900373z`. eprint: `http://dx.doi.org/10.1021/ct900373z`. URL: `http://dx.doi.org/10.1021/ct900373z`.

(*168*) B. Widom. “Some Topics in the Theory of Fluids.” In: *The Journal of Chemical Physics* 39.11 (1963), pp. 2808–2812. DOI: `http://dx.doi.org/10.1063/1.1734110`.

(*169*) G. V. Wylen and R. Sonntao. *Fundamentals of classical thermodynamics*. 3rd edition, John Wiley & Sons, New York., 2013.

(*170*) V. I. Levenshtein. “On the Minimal Redundancy of Binary Error-Correcting Codes.” In: *Information and Control* 28.4 (1975), pp. 268–291. DOI: `10.1016/S0019-9958(75)90300-9`.

(*171*) S. Itzkovitz et al. “Coarse-graining and self-dissimilarity of complex networks.” In: *Physical Review E* 71.1 (2005), p. 016127.

(*172*) I. Albert and R. Albert. “Conserved network motifs allow protein–protein interaction prediction.” In: *Bioinformatics* 20.18 (2004), pp. 3346–3352.

(*173*) J.-R. Kim, Y. Yoon, and K.-H. Cho. “Coupled feedback loops form dynamic motifs of cellular networks.” In: *Biophysical Journal* 94.2 (2008), pp. 359–365.

(*174*) M. Jain et al. “The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community.” In: *Genome Biology* 17.1 (Nov. 2016), p. 239. ISSN: 1474-760X. DOI: `10.1186/s13059-016-1103-0`.

Curriculum Vitae

Name	Sven Frederik Jager
Date of Birth	08.09.1988
Place of Birth	Berlin

Education

08/1994–06/1998	Gutenberg Schule Dieburg
08/1998–06/2004	Goethe Schule Dieburg
08/1994–06/2007	Landrat Gruber Schule Dieburg (Abitur)
10/2007–06/2008	Study of Engineering at TU Darmstadt
10/2008–06/2010	Study of Biomolecular Engineering at TU Darmstadt (B.Sc.)
10/2010–11/2013	Study of Biomolecular Engineering at TU Darmstadt (M.Sc)
12/2013–3/2018	PhD at TU Darmstadt (Computational Biology and Simulation)

Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit entsprechend den Regeln guter wissenschaftlicher Praxis selbstständig und ohne unzulässige Hilfe Dritter angefertigt habe.

Sämtliche aus fremden Quellen direkt oder indirekt übernommenen Gedanken sowie sämtliche von Anderen direkt oder indirekt übernommenen Daten, Techniken und Materialien sind als solche kenntlich gemacht. Die Arbeit wurde bisher bei keiner anderen Hochschule zu Prüfungszwecken eingereicht.

Darmstadt, den 22. Januar 2018

Sven Jager

Danksagung

Zum Schluss möchte ich allen danken, die mich während meiner Dissertation unterstützt haben.

Als erstes danke ich meinem PI Kay Hamacher für die Betreuung meiner Arbeit und dafür, dass ich stets eigene Ideen in meine Forschung einbringen konnte. Ich habe bei dir sehr viel gelernt. Danke, dass du immer ein offenes Ohr für alles hattest. Ich bedanke mich auch für die Unterstützung und Hilfe beim Erstellen und dem Schreiben von Manuskripten.

Ein besonders großer Dank geht an meine Zweitgutachterin Beatrix Süß. Sie begeisterte mich für das Feld der RNA und hat mich stets unterstützt (besonders in den letzten Monaten meiner Arbeit). Danke!

Katja Schmitz Danke, dass du immer ein offenes Ohr sowie Labor für mich hattest.

Benjamin Schiller Dir möchte ich besonders danken. Unsere gemeinsame Forschung hat diese Arbeit sehr stark beeinflusst.

Sascha Hein Dir möchte ich danken für die ganze Hilfe im Labor sowie für alle wissenschaftlichen Diskussionen, die wir hatten.

Thea Lotz Danke, dass du mir immer geholfen hast und immer ein offenes Ohr für mich hattest. Ich finde es super, dass du iGEM so gut betreust und wünsche dir ganz viel Erfolg.

Heribert Warzecha und iGEM Wir haben zwei Jahre iGEM betreut und es waren schöne und zugleich erfolgreiche Jahre. Vielen Dank für deine Unterstützung.

Benjamin Mayer Ben, ich bedanke mich für die vielen Stunden an der Tafel sowie deine Hilfe bei iGEM und besonders für die Freundschaft während der letzten Jahre.

Oliver Buss Danke für die ganze Unterstützung, die Diskussionen sowie unsere gemeinsamen Arbeiten in der Promotion.

Daniel Bauer und Phillipp Babel An euch geht ein besonderer Dank! Einmal für den Kicker, den Bierkühlschrank sowie die MD Simulationen. Es war sehr cool mit euch in der AG.

Michael Schmidt Dir möchte ich für die vielen Diskussionen sowie für unsere Zusammenarbeit danken. Ich wünsche dir viel Erfolg bei deiner Karriere als Wissenschaftler.

Der AG Ein besonderer Dank geht natürlich auch an die AG. Es hat sehr viel Freude gemacht in diesem Umfeld zu promovieren. Dank insbesondere an: Sabine Knorr, Stefanie Weissgraeber, Patrick Boba, Martin Hess, Frank Keul, Jan Wissman, Sebastian Stammler, Felix Reinhardt, Caroline Pierre und Valerie Fehst.

Florian Groher Unsere Kooperation im letzten Jahr hat meine Arbeit stark beeinflusst. Dafür möchte ich dir danken.

Meine Studenten Natürlich geht ein Dank auch an meine Studenten: Thomas, Max, Tine und Sebastian. Ihr habt zu dieser Arbeit viel beigetragen. Tine und Max konnte ich sogar für die Computational Biology begeistern. Gerade so etwas hat mich sehr gefreut! Danke für tolle Paper und Tage im Labor.

Familienmitgliedern und Freunden Nicht zuletzt möchte ich allen meinen Familienmitgliedern und Freunden, insbesondere meiner Schwester, meinem Vater, meiner Mutter und meiner Großmutter, für ihre Unterstützung, ihr Essen und ihre Liebe danken.

Bianca Reisinger Bianca Reisinger, ich bedanke mich einfach für alles! Du hast mich so sehr unterstützt und einiges mit mir ausgehalten. Ich liebe dich.