

On Evolution, Structure and Dynamics in Potassium Channels



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Vom Fachbereich Biologie der Technischen Universität Darmstadt zur Erlangung des akademischen Grades eines *Doctor rerum naturalium* genehmigte Dissertation von

Dipl.-Biol. Frank Keul aus Arnstadt

Berichterstatter: Prof. Dr. Kay Hamacher

Mitberichterstatter: Prof. Dr. Gerhard Thiel

Eingereicht am: 12.04.2016

Mündliche Prüfung am: 12.06.2017

Darmstadt 2017

D17

Keul, Frank: On Evolution, Structure and Dynamics in Potassium Channels
Darmstadt, Technische Universität Darmstadt,
Jahr der Veröffentlichung der Dissertation auf TUprints: 2018
Tag der mündlichen Prüfung: 12.06.2017
Veröffentlicht unter CC BY-NC 4.0 International
<https://creativecommons.org/licenses/>

To my wife and son.



Summary

Ion channels play fundamental roles in living organisms ranging from propagation of action potentials to chemotaxis. While the sequences of ion channels can differ greatly, their function and their structure share great similarities. Structurally, the here investigated channels follow a basic, structural composition with inner and outer helices entrenching a pore helix and an selectivity filter. These homotetrameric structures form water filled pores through which ions can pass upon opening of the channel. To investigate structure-function correlates within this broad group of proteins we derive in this thesis novel approaches to capture evolutionary substitution behaviors, investigate co-evolutionary complexity and analyze structural relations.

The first chapter captures an overview of stand-alone scientific contributions with a brief introduction to their methodological aspects as well as their results. Within the following chapters of this thesis we will integrate and expand parts of these publications to further the understanding in interdependencies within ion channels.

The second chapter revolves around a novel type of substitution matrices derived from Pfam alignments of channel proteins. Using the novel PFASUM algorithm (Keul *et al.*, 2017) we generate family-specific substitution matrices and show their improved performance in comparison to BLOSUM matrices for channel protein sequences. We find that these novel PFASUM matrices cluster amino acids with diverse physico-chemical properties stronger than their BLOSUM counterparts which in turn leads to improved alignments.

In the third chapter of this work we aim at capturing amino acid sequence inherent information from two major channel protein families through information theoretical methods. Here, we devise the measure of co-evolutionary complexity ΔC_{MI} which measures the deviation of observed evolutionary mechanics from a reference model. When applied to potassium channel sequences we find large differences in complexity of the evolutionary mechanism within the outer helix between channels with two transmembrane domains and six. We attribute the substantially different evolutionary behavior of six transmembrane domain channels to the interaction between the so called voltage sensing domain and outer transmembrane helix.

In the last chapter we use coarse-grained elastic network models to compare protein dynamics between open and closed channel structures through analyzing changes in the free energy of the fold. Furthermore, we investigate relationships between the regions within channel pore of differently sized channels from numerous organisms by examining the influence of perturbations between the regions. Hereby, we focus on changes in free energy in a subspace of the molecular Hamiltonian when subjected to perturbation while still considering the entirety of all interactions within the Hamiltonian. This allows us the comparison of perturbation thought experiments on differently sized, but similarly organized structures. Through this we are able show that the selectivity filter of ion channels is decoupled from the all other pore forming segments. Furthermore, we find that – in the context of elastic network

models – fold dynamics within the analyzed subspace of the Hamiltonian are independent of sequence information.

Zusammenfassung

Ionenkanäle sind beteiligt an essentiellen Abläufen in lebenden Organismen von der Weitergabe von Aktionspotentialen bis hin zu chemotaktisch induzierten Bewegungen. Während die Aminosäuresequenzen von Ionenkanälen sich stark voneinander unterscheiden kann, besitzen Kanäle sehr große Ähnlichkeiten auf Struktur- und Funktionsebene. Strukturell folgen die in dieser Arbeit untersuchten Kanäle einem grundlegenden Muster in welchem innere und äußere transmembranhelikale Bereiche eine Porenhelix und ein Selektivitätsfilter einrahmen. Die homo-tetrameren Strukturen besitzen eine wassergefüllte Pore, durch die Ionen beim Öffnen des Kanals hindurchtreten können. Um Struktur-Funktions Korrelate innerhalb dieser breiten Gruppe von Proteinen zu untersuchen, leiten wir in dieser Arbeit neue Ansätze her, um die co-evolutionäre Komplexität, evolutionäres Substitutionsverhalten und strukturelle Beziehungen zu untersuchen.

Das erste Kapitel gibt einen Überblick über eigenständige wissenschaftliche Beiträge mit einer kurzen Einführung in deren methodischen Aspekte und Ergebnisse. In den folgenden Kapiteln dieser Arbeit integrieren und erweitern wir Teile dieser Publikationen, um das Verständnis der Zusammenhänge innerhalb der Ionenkanäle zu vergrößern.

Das zweite Kapitel dreht sich um eine neue Arten von Substitutionsmatrizen, welche auf Pfam-Alignments von Kanalproteinen basieren. Mit dem neuen PFASUM-Algorithmus (Keul *et al.*, 2017) erzeugen wir hier familienspezifische Substitutionsmatrizen und zeigen deren verbesserte Alignment-Qualitäten für Kanalproteinsequenzen im Vergleich zu BLOSUM-Matrizen. Wir finden hier, dass diese neuartigen PFASUM-Matrizen Aminosäuren mit verschiedenen physikalisch-chemischen Eigenschaften stärker als ihre BLOSUM-Pendants zusammenfassen, welche wiederum zu verbesserten alignments führt.

Das dritte Kapitel dieser Arbeit befasst sich mit der Erfassung von inhärenten Informationen aus Aminosäuresequenzen zwei großer Kanal-Protein-Familien durch informationstheoretische Methoden. Hier wird das Maß der co-evolutionären Komplexität ΔC_{MI} hergeleitet, welches die Abweichung der beobachteten evolutionären Mechanismen von einem Referenzmodell misst. Bei Anwendung auf Kaliumkanalsequenzen finden wir große Unterschiede in der Komplexität des Evolutionsmechanismen innerhalb der äußeren Helix beim Vergleich von TM2- und TM6-Kanalproteinen. Wir führen dieses wesentlich unterschiedliche, evolutionäre Verhalten von TM6-Kanälen auf deren Wechselwirkung zwischen der sogenannten voltage sensing domain und der äußeren Transmembranhelix zurück.

Im letzten Kapitel verwenden wir elastische Netzwerkmodelle (ENM), um die Proteindynamik zwischen offenen und geschlossenen Kanalstrukturen zu analysieren, in dem wir Änderungen Faltung durch Unterschiede in deren freien Energie messen. Darüber hinaus untersuchen wir die Beziehungen zwischen den einzelnen Regionen innerhalb der Poren von Kanälen. Hier untersuchen wir den Einfluss von Störungen zwischen den Regionen in unterschiedlich großen Kanälen. Dabei konzentrieren wir uns auf Änderungen der freien Energie in einem Unterraum der ENM Hesse-Matrix, wenn dieser einer Störung ausgesetzt wird, aber die Gesamtheit aller Wechselwirkungen immer noch Berücksichtigung

findet. Dies erlaubt uns den Vergleich von Störungsexperimente in unterschiedlich großen, aber strukturell ähnlich organisierten Kanälen. hierdurch können wir zeigen, dass der Selektivitätsfilter von Ionenkanälen von allen anderen porenbildenden Segmenten entkoppelt ist. Weiterhin finden wir, dass im Rahmen von der elastischen Netzwerkmodelle die Faltungsdynamik im analysierten Teilraum des molekularen Hamiltonischen Mechanik unabhängig von Sequenzinformation ist.

Contents

Introduction into Potassium Channels	1
1. Contributions	5
1.1. Visual analysis of patterns in multiple amino acid mutation graphs	5
1.2. Addressing inaccuracies in BLOSUM computation improves homology search performance	7
1.3. Consistent Quantification of Complex Dynamics via a Novel Statistical Complexity Measure	9
1.4. PFASUM: A substitution matrix from Pfam structural alignments	12
2. Channel-specific Substitution Matrices	15
2.1. Background	16
2.1.1. Evolutionary Models	16
2.1.2. Multiple Sequence Alignments	19
2.2. Method	20
2.2.1. Pfam Database	20
2.2.2. Generation of Substitution Matrices	20
2.2.3. Comparison of Substitution Matrix Composition	21
2.2.4. MSA Generation	22
2.3. Results	23
2.3.1. Substitution matrix comparison	23
2.3.2. Matrix performance generating alignments	26
2.4. Discussion	29
3. Evolutionary Information in Sequence Data	31
3.1. Background	31
3.1.1. Mutual Information normalizations and corrections	33
3.1.2. Direct Coupling Analysis	36
3.1.3. Statistical Complexity	38
3.2. Methods	39
3.2.1. Datasets	39
3.2.2. Comparison of information theoretical measures	41
3.3. Results	42
3.3.1. Properties of Co-evolutionary Complexity	43
3.3.2. Insights into ion channel evolution	47
3.4. Discussion	51

4. Dynamics of Channel Proteins	53
4.1. General channel organization	53
4.1.1. Two TMD Ion Channels	55
4.1.2. TM6 Ion Channels	59
4.2. Background	63
4.2.1. Partition Function	64
4.2.2. Free Energy perturbation	65
4.2.3. Domain Interaction Perturbation	66
4.3. Methods	67
4.4. Results	69
4.4.1. Conformational Changes in Channels upon Opening and Closing	69
4.4.2. General Dynamics of the Channel Pore Module	73
4.5. Discussion	78
Conclusion	81
Bibliography	83
Appendices	97
A. Acronyms	99
B. Channel specific substitution matrices	103
C. Evolutionary Information in Sequence Data	105

Introduction into Potassium Channels

Biological cells are the smallest functional element of any known, living organism. Complex chemical reactions within cells are only possible due to a strict structural separation between cell interior and its surrounding environment through phospholipid bilayer membranes. While small, uncharged molecules such as gases can pass through biological membranes via diffusion, larger and/or charged molecules require protein-based assistance to cross the hydrophobic boundary. Amongst other classes of proteins, ion channels are capable of moving charged ion across membranes and act as pores within these.

Channels are ubiquitously present in cells of higher order eukaryotic organisms and are found to play central roles in many forms of diseases. A condition known as episodic ataxia, where diseased suffer from severe symptoms of discoordination, is caused by missense mutations in the gene encoding the human Kv1.1, a voltage-gated potassium channel (Browne *et al.*, 1994). Other potassium channels such as the ATP-sensitive Kir6.2 are indirectly influenced by mutation. Here, alterations within its regulatory subunit SUR1 lead to hyperinsulinemic hypoglycemia, a condition often found in type II diabetes (Aittoniemi *et al.*, 2009). The rare chanelopathic disease hypokalemic periodic paralysis, in which affected experience muscle weakness and paralysis, is caused by mutations in genes for the voltage-gated sodium channel Nav1.4 and the voltage-gated calcium channel Cav1.1 (Matthews *et al.*, 2009; Matthews and Hanna, 2014). While the aforementioned diseases are caused by mutations in genes for cation channels, cystic fibrosis originates from mutations in the chloride channel cystic fibrosis transmembrane conductance regulator protein (O'Sullivan and Freedman, 2009).

In contrast to animal cells, plant cells possess significantly different cellular organization due vacuoles, chloroplasts and a cell wall matrix surrounding the membrane. Despite these differences, ion channels play major roles in key functions of higher plants. In *Dionaea muscipula* (Venus flytrap) action potential is propagated through ion channels and results in increased permeability of aquaporins with subsequent rapid closing of the plant's leaves (Volkov *et al.*, 2008). Other more common behaviors of plant cells, such as the opening and closing of stomata, are also regulated through alterations in the turgor pressure through ion flux across membranes. In case of the stomatal guard cells, opening of the hyperpolarization activated KAT1 and KAT2 potassium channels leads to ion influx into the cytoplasm, osmotic water influx, consequential swelling of the guard cells and opening of the stomatal pore (Assmann, 1993). In smaller, singled-celled organisms, such as the green alga *Chlamydomonas reinhardtii*, ion channels are quintessential for orientational movements. In *C. reinhardtii* phototaxis is achieved through sensation of light in channelrhodopsin-1 and -2, consequential depolarization of membrane potential and opening of Ca²⁺ channels in the flagella (Nagel *et al.*, 2003).

In bacterial organisms comparably small channels can react on changes in pH level and induce ion flux across the hydrophobic boundary. In *Streptomyces lividans* the proton-dependent opening of the potassium channel KcsA generates an electrochemical gradient (Doyle *et al.*, 1998; Zimmer *et al.*, 2006). Other channels, such as the mechanosensitive MscL and MscS of *Escherichia coli* are capable to transform the information of applied tension on the membrane to an electrochemical signal, preventing *E. coli* from lysis through hypoosmotic shock (Haswell *et al.*, 2011). Investigation of

bacterial channels has resulted in valuable insights into the structural foundation of their eukaryotic counterparts. Prokaryotic channels are often smaller and more simplistic, while sharing functional similarities to their larger eukaryotic siblings (Hille, 2001).

While all aforementioned ion channels can be found in living organisms, numerous channel proteins were found in viruses in recent years. Spanning different groups of viruses ion channels can be found in, e.g., influenza A and B (Holsinger *et al.*, 1994), the human immunodeficiency virus (Schubert *et al.*, 1996) or the *Paramecium bursaria* chlorella virus 1 (Plugge *et al.*, 2000). Virus channels play crucial roles in promotion of virus uncoating in endosomes (Mould *et al.*, 2000), in regulation of virus release (Schubert *et al.*, 1996) or in the injection of the virus genome into the host by overcoming pressure barriers (Neupärtl *et al.*, 2008).

Structurally, most ion channels share a similarly organized, central ion conducting pore element which can be surrounded by, e.g., cytoplasmic vestibules (Kuo *et al.*, 2003), ligand binding (Clayton *et al.*, 2008), voltage sensing (Long *et al.*, 2005) or light sensing domains (Nagel *et al.*, 2003). Channel gating and ion conductance across membranes usually follows a direct stimulus and moves along electrochemical gradients. Along the vertical axis of the proteins, ions can travel through a water filled pore at up to 10^6 ions per second for only a few milliseconds (Aldrich *et al.*, 1983; Sanchez *et al.*, 1986), while channels retains a highly selective nature with an ion specific filter sequence. While opening of channels can originate from opposing events such as hyper- or depolarization of membrane potentials, the channels involved in mediation of these events share significant structural similarities.

Due to their involvement in numerous diseases and their function as molecular switches that can transform different stimuli into electrical current, research on ion channels gained popularity in recent years. Hence, the following work will examine evolutionary and structural foundations of channel proteins to further elucidate dependencies within sequence and structure space. Therefore the work is divided as followed:

In the first chapter of this thesis we examine publications I contributed to and which find application in later chapters. The articles summarize approaches to improve sequence comparison and alignment as well as novel algorithms to extract information from multiple sequence alignments. We present here two publications in which we first correct a decade old miscalculation in the BLOSUM matrix code and, second, present a whole new substitution matrix from structural alignments. Furthermore, we show results from a novel pattern analysis algorithm on basis of mutation graphs and recapitulate the principles of the statistical complexity within multiple sequence alignments to characterize evolutionary information.

The second chapter of this thesis delves into the essential task of generating multiple sequence alignments. Here, we derive novel family-specific substitution matrices based on alignments for larger six and smaller two transmembrane domain channels. Our goal here is to improve upon traditional, generalized evolutionary models by generating case specific substitution models. Ideally, these models inherit unique properties of the underlying alignment and allow us to align specific sequences with fewer sequential similarities with higher accuracy. Hence, we compare the performance of these novel substitution matrices to comparable BLOSUM matrices by using different alignment algorithms on test datasets.

Many different information theory based approaches are capable to derive structural and evolutionary knowledge from these multiple sequence alignments. In the third chapter, we derive a novel method to measure the intricate effects of evolutionary dynamics within proteins and compare the results to traditional measures of co-evolution. Here, we closely examine differences in the evolutionary dynamics between eukaryotic and prokaryotic potassium channels to gain insights for future usage in design of channel proteins.

In the final chapter, we then focus our attention on the structural similarities of different ion conducting proteins. To this end, we derive a novel approach to compare inherent dynamics of differently sized channel proteins through the application of elastic network models and Schur complementation. Here, we aim at clarifying whether or not the filter region of ion channels is structurally decoupled from other regions of the protein. While most structures used in this chapter conduct potassium ions, we deliberately extended the channel selection to include sodium channels to conclude in a more generalized statement.



1 Contributions

During my research and work on this thesis I contributed to several projects. Manuscripts which have been published and which are still in preparation are listed below with a short summary of methodology and results as well as a description of my contribution. The corresponding publication is referenced.

1.1 Visual analysis of patterns in multiple amino acid mutation graphs

We developed a fast pattern search algorithm aimed at finding biologically relevant patterns in many mutation graphs. Through the visualization system interactive exploration and comparison of the found patterns of these acyclic directed graphs is possible. Furthermore, through location of found patterns in the three dimensional structure, the found patterns presents a starting point for biological analysis. The detailed description of the pattern search algorithm and the visualizations have been published under:

Lenz, O; Keul, F; Bremm, S; Hamacher, K and von Landesberger, T (2014) Visual Analysis of Patterns in Multiple Amino Acid Mutation Graphs, *In Proc. VAST*, pp. 93-102

1.1.1 Background

Modern drug design focuses on developing novel pharmaceuticals with increased impact and lessened unwanted side effects. Binding sites of drugs can be found within active sites, on the surface or between proteins. While in eukaryotic organisms the application of drugs can affect unwanted pathways and proteins, anti-viral drugs face different challenges. Here, the target protein can, upon administration of the drug, exhibit compensatory mutations and render the (novel) drug ineffective, due to the quasi-species nature of viral load (Eigen *et al.*, 1988; Martell *et al.*, 1992; Handel *et al.*, 2006; Metzner *et al.*, 2009).

In light of this, understanding the role of compensatory mutations has moved into the focus of scientist. Here, trying to predict possible mutations that can reduce the effectiveness of drugs is one of the main goals as these events can strongly influence large parts of a protein. Hence, we focused in our study on such linked chain mutations, represented as mutation graphs.

1.1.2 Methods

In order to obtain mutation graphs, we were inspired by Bleicher *et al.* (2011). By using a multiple sequence alignment of HIV1 protease, we derived correlated mutations through testing whether the occurrence of amino acid a in column i of the alignment significantly increased or decreased the probability to find amino acid b in column j . Here, we focused on all n_a subset sequences containing

amino acid a . The conditional probability of observing at least $n_{b|a}$ observations is given via cumulative binomial distribution if the expected occurrence holds $n_{b|a} < n_a \cdot \frac{n_b}{N}$, with n_b being the number of observations of amino acid b in column j :

$$p(X \geq n_{b|a}) = 1 - \sum_{k=0}^{n_{b|a}} \frac{n_a!}{k!(n_a - k)!} \left(\frac{n_b}{N}\right)^k \left(1 - \frac{n_b}{N}\right)^{n_a - k} \quad (1.1)$$

If $n_{b|a} > n_a \cdot \frac{n_b}{N}$, we test for opposite tail of the binomial distribution through $p(X < n_{b|a}) = 1 - p(X \geq n_{b|a})$.

Based on this, we derived a directed unweighted graph. In this graph, every vertex was labeled with the respective amino acid/position pair. Edges connecting vertices were drawn whenever the p-value (i.e. $p(X < n_{b|a})$ or $p(X \geq n_{b|a})$) was below a significance threshold. We set this threshold at 0.01 prior to applying the conservative Bonferroni-correction for multiple testing. Hence, the effective p-value threshold was $P < 10^{-8}$ for the alignment of the HIV1 protease. The resulting graph was then subject to the novel pattern recognition algorithm.

Further details can be found in Lenz *et al.* (2014).

1.1.3 Results

Our pattern search algorithm was able to detect 101 patterns in the HIV protease mutation graph. We were able to show that within the HIV dataset of treated and untreated patients, certain evolutionary patterns serve as bottleneck for cascading evolutionary effects. These bottleneck patterns presented themselves predominantly within mutation graphs and were thus evolutionary pathways that had to be crossed. The most outstanding bottleneck patterns involved the residues 57, 61, 65 and 71. Here, a mutation at position 61 to glutamine induced occurrence of specific amino acids in some positions (position 65 and 72), whereas in position 57 the opposite was observed. Altogether, mutation of the glutamine at position 61 influenced multiple residues in the so-called cantilever region (position 65 and 72) and the flaps of HIVP (position 57). The found evolutionary pattern was thus connecting these two regions. Furthermore, the unique property of the exo region near the position 61 was used in a HIVP inhibitor design (Chang *et al.*, 2010; Kunze *et al.*, 2014), further emphasizing the need to recognize evolutionary dependencies and patterns.

Furthermore, our analysis focused on the most commonly observed pattern which was found in over 58% of all mutation graphs. Here, occurrence of an asparagine at position 88 within an α -helical structure, applies both an inhibitory and an enhancing evolutionary effect on position 30. By finding an asparagine at position 88, the probability of finding an aspartic acid in position 30 is increased, whereas the structurally similar asparagine is disfavored in this position. Hence, the flexibility of the loop region connecting the HIVP active site (Asp25 – Thr26 – Gly27) with the subsequent β -sheet could be directly influenced through this change in net charge.

1.1.4 Contributions

OL and TL devised the pattern search algorithm. FK prepared the datasets and derived the graph generating algorithm. KH and TL devised the study. All authors prepared the manuscript.

1.2 Addressing inaccuracies in BLOSUM computation improves homology search performance

Since its release in 1991, the BLOSUM matrix – or in general BLOSUM-type matrices – are amongst the most prominent and heavily used substitution matrices, despite known errors in the matrix generation (Styczynski *et al.*, 2008). In this work, we present an additional coding error which when corrected and in combination with the correction proposed by Styczynski *et al.* yields the novel CorBLOSUM substitution matrix. In a large scale study, we reveal that the CorBLOSUM matrix outperforms its predecessors significantly. The complete analysis with the entirety of the results has been published under:

Hess, M; Keul, F; Goesele, M and Hamacher, K (2016) Addressing inaccuracies in BLOSUM computation improves homology search performance, *BMC Bioinformatics*, Volume 17:189

1.2.1 Background

An elementary procedure in modern computer-based bioinformatics is the searching of databases for homologous sequences. Fine-tuning parameters of modern day database search algorithms is far from a trivial task and evolutionary models – as substitution matrices – can be varied freely by the user. Hence, investigating the performance of substitution matrices has played a crucial role in earlier studies (Brenner *et al.*, 1998; Price *et al.*, 2005; Styczynski *et al.*, 2008; Song *et al.*, 2015). Here, the ASTRAL database, the *de facto* gold standard for homologous sequence search, has been used as benchmark dataset (Brenner *et al.*, 1998; Price *et al.*, 2005; Styczynski *et al.*, 2008).

In this context, BLOSUM-type matrices are often used as reference, as database search programs such as NCBI BLAST (Altschul *et al.*, 1990) and SSEARCH (Pearson, 1991) present these matrices as default parameters. Nevertheless, the BLOSUM-type matrices originate from a code with at least two coding errors, one of which has been presented in an earlier study (Styczynski *et al.*, 2008). Correcting this error in the cluster weighting procedure results in a new matrix, referred to as RBLOSUM. While the RBLOSUM corrected a coding error, the performance of this new substitution matrix did not improve the homologous sequence search performance when compared to the BLOSUM matrix – at least on available benchmarks at the time.

In our study, we showed that another easily missed coding error persisted in the original BLOSUM code. Here, typecasting of the calculated similarity threshold – used for clustering of the sequences – leads to faulty clustering of sequence close to the user preset similarity threshold. Correcting this error yields a from BLOSUM systematically different substitution matrix – the CorBLOSUM. Through exhaustive analysis on all available ASTRAL releases at different maximal sequence identities we

conducted the largest performance assessment of BLOSUM-, RBLOSUM- and CorBLOSUM-type matrices to date.

1.2.2 Methods

In order to obtain comparable substitution matrices, we derived RBLOSUM and CorBLOSUM matrices with similar relative entropy to the original BLOSUM50 and BLOSUM62 matrices. Since the database for the generation of the original BLOSUM matrix (BLOCKS 5) has been updated numerous times, we expanded our analysis to the BLOCKS13+ and BLOCKS14.3 releases. Hence, we derived 18 different substitution matrices of which a set of three (BLOSUM, RBLOSUM and CorBLOSUM) show similar relative entropies and hence are comparable.

To perform a thorough performance evaluation, we used all available ASTRAL releases (version 1.55 to 2.06) at three different similarity thresholds (20%, 40% and 70%). This wide variety of databases allowed us to assess the effect of ever-improving sequence space coverage and differing database compositions on the performance of different substitution matrices. Additionally, we varied the gap open and gap extension penalties in SSEARCH between 5 and 20, and 1 and 2, respectively, hence covering the most used gap parameter space.

The assessment of search performance was conducted with coverage measure \mathcal{Q} (Green and Brenner, 2002; Price *et al.*, 2005) at exactly 0.01 errors per search query. Here, we developed a new CoverageCalculator tool to improve upon the performance of existing tools (Green and Brenner, 2002). In this tool, sequence relations are based on the ASTRAL superfamily annotation and the resulting coverage is normalized to account for superfamily sizes. To assess the significance of the found coverage differences, we employed the Bayesian bootstrap proposed earlier (Price *et al.*, 2005) to obtain Z-scores at a 2σ -level.

1.2.3 Results

Our results showed, that implementing both the RBLOSUM-correction and our found correction resulted in a substitution matrix that outperformed its BLOSUM and RBLOSUM counterparts in $\sim 75\%$ of all test scenarios. On current SCOPE based ASTRAL releases CorBLOSUM matrices fared even better and achieved a higher maximum coverage in $\sim 86\%$ of all scenarios in comparison to the faulty BLOSUM-type matrices.

Additionally, we were able to show, that composition of the benchmark database severely influences the performance of the substitution matrices (see Fig. 1.1, exemplarily). For state-of-the-art databases with a broader structure and sequence space coverage, all substitution matrices increased in performance.

Furthermore, our exhaustive study showed that the results for the RBLOSUM-BLOSUM comparison published earlier (Styczynski *et al.*, 2008) were a mere snapshot as the RBLOSUM correction resulted in a performance gain in other setups – especially when varying both ASTRAL release and BLOCKS version. Nevertheless, since CorBLOSUM-type outperformed either in for the better part of all test scenarios, we advise updating current sequence search tools to incorporate CorBLOSUM matrices instead of the erroneous BLOSUM- and RBLOSUM-type matrices.

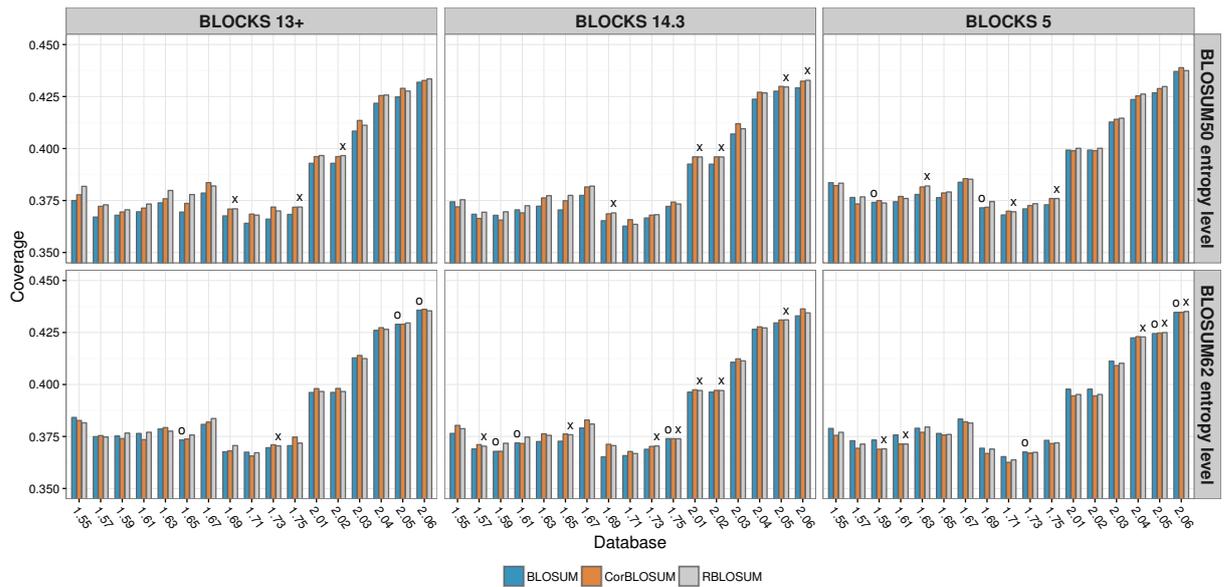


Figure 1.1.: Progression of the maximum achieved coverage of CorBLOSUM-, RBLOSUM- and BLOSUM-type matrices for all ASTRAL40 test databases. The upper row shows the results for the respective BLOSUM50 entropy level, the lower row for BLOSUM62 entropy level. Insignificant coverage differences between CorBLOSUM and BLOSUM are indicated by an O and between CorBLOSUM and RBLOSUM by a small an X above the bars.

For further details, see Hess *et al.* (2016).

1.2.4 Contribution

MH, FK, MG and KH jointly conceived the concept of the paper. MH and FK performed the data acquisition, analysis and manuscript and supplement preparation. KH and MG supervised the work and edited the manuscript. Code implementation was performed by MH and FK. FK prepared the figures. All authors discussed the results and implications and commented on the manuscript on all stages. All authors read and approved the final manuscript.

1.3 Consistent Quantification of Complex Dynamics via a Novel Statistical Complexity Measure

We developed a novel method to measure the complexity (difference) of biological systems. By comparing sequences of HIV positive patients who have been subject to protease inhibitor treatment to sequences from untreated patients, we gained insights on the reaction of the evolutionary complexity upon inhibitor treatment. The results will be published under:

Keul, F and Hamacher, K (2017) Consistent Quantification of Complex Dynamics via a Novel Statistical Complexity Measure, manuscript in preparation

1.3.1 Background

Information theory has found a wide range of applications in numerous fields, ranging from telecommunications to medicine by analyzing distributions of variables. After the introduction of the information theory through Shannon (1948), measuring correlated structures of a random system has come into focus recently (Lopez-Ruiz *et al.*, 1995; Feldman and Crutchfield, 1998). Here, systems which possess a very high (uniform distributed random variables) or a very low (one realization in the random variables) degree of randomness are not considered complex systems. Nevertheless, an wide variety of diversely structured processes can be found in between the limits of the distribution. Hence, statistical complexity measures have been developed to detract randomness and present insights into structure and regularity of a distribution.

The statistically complexity, as proposed by Lopez-Ruiz *et al.* (1995), can be summarized in its generalized form as $C(\mathbf{p}) = H(D_{\text{KL}})D(\mathbf{p}, \mathbf{q})$ with $H(\mathbf{p})$ being the Shannon entropy of the distribution \mathbf{p} . $D(\mathbf{p}, \mathbf{q})$ is a distance between the observed (\mathbf{p}) and a reference (\mathbf{q}) distribution. It is understood that the reference distribution captures the dynamics of the system without the complex relationships present in \mathbf{p} . The choice of the distance measure $D(\mathbf{p}, \mathbf{q})$ has been discussed elsewhere (Feldman and Crutchfield, 1998; Martin *et al.*, 2006).

1.3.2 Methods

In order to capture the complexity of dynamical systems, we derived a novel approach to the computation of statistical complexity. While the Kullback-Leibler divergence (D_{KL}) has been proposed as a distance measure (Feldman and Crutchfield, 1998; Martin *et al.*, 2006), we focused on symmetric Jensen-Shannon divergence in Eqn. 1.2.

$$D_{\text{JS}}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} (D_{\text{KL}}(\mathbf{p}, \mathbf{m}) + D_{\text{KL}}(\mathbf{q}, \mathbf{m})) \quad (1.2)$$

with $\mathbf{m} = \frac{1}{2} (\mathbf{p} + \mathbf{q})$

Uniquely, the square root of the Jensen-Shannon divergence converts the measure to a metric and presents a basis for important convergence theorems (Endres and Schindelin, 2003).

Through extension of the original concept of statistical complexity to *statistical complexity difference* we obtain a ΔC value as in 1.3

$$\begin{aligned} \Delta C(\mathbf{p}||\mathbf{q}) &= C(\mathbf{p}||\mathbf{q}) - C(\mathbf{q}||\mathbf{p}) \\ &= (H(\mathbf{p}) - H(\mathbf{q})) \sqrt{D_{\text{JS}}(\mathbf{p}||\mathbf{q})} \end{aligned} \quad (1.3)$$

Here, Shannon's entropy is normalized to its respective maximum to ensure comparability of the distributions \mathbf{p} and \mathbf{q} . Furthermore, we replaced the distribution of the reference system \mathbf{q} – typically the uniform distribution – with the distribution of the unperturbed system. Hence, ΔC measures the complexity difference upon perturbation of the system.

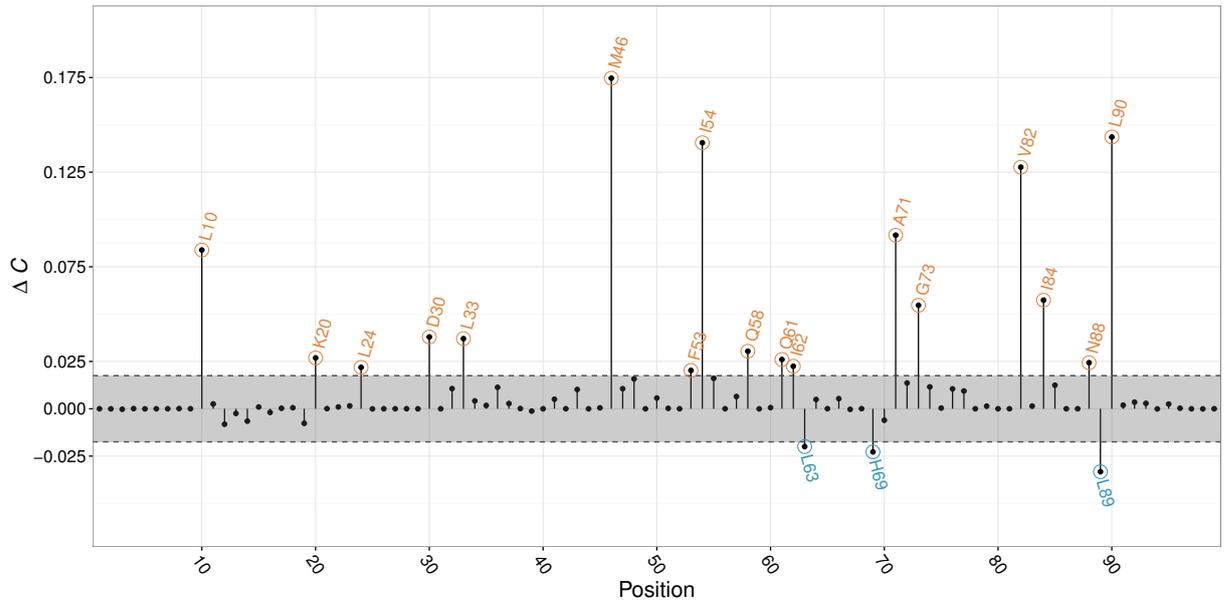


Figure 1.2.: Plot for the ΔC values for the comparison of the treated and untreated datasets of the HIV1 protease. Highlighted in orange are positions where the evolutionary complexity of the HIV1 protease increases upon drug treatment. Blue positions experience the reversed effect.

Here, we applied the complexity difference to multiple sequence alignments of the HIV1 protease (HIVP). The HIVP sequences were obtained from the HIV drug resistance database (Rhee *et al.*, 2003). All sequences consisting of less than 99 amino acids or which contained ambiguous, non-canonic amino acids or stop codons were removed. Hence, each HIVP sequence consists of 99 amino acids was considered to be able to produce a functional protein. Thus, it was not necessary to perform heuristic-driven alignments, not introducing bias through these procedures. We computed the ΔC between the perturbed system (i.e. sequences from patients treated with protease inhibitor) and the reference system (all untreated patients).

1.3.3 Results

Our results show that ΔC measure is capable to identify numerous positions in the alignment where the statistical (evolutionary) complexity between sequences from treated and untreated patients differs (see Fig. 1.2). Due to the properties of ΔC it is possible to attribute directionally to the change in complexity based on the algebraic sign of ΔC . Here, at positive ΔC values, the statistical complexity of evolution is higher for the sequence of treated patients than compared to the complexity of untreated patients. Furthermore, mutations at almost all position with $|\Delta C| \geq 0.1 \max(\Delta C)$ have been shown to induce reduced susceptibility to HIVP inhibitors. This leads us to assume that at positions with $\Delta C > 0$ (and $|\Delta C| \geq 0.1 \max(\Delta C)$) selective pressure through the application of drugs induces an increased complexity of evolution. Likewise, positions at which $\Delta C < 0$ (and $|\Delta C| \geq 0.1 \max(\Delta C)$) signalize reduced statistical complexity of evolution and hence can give insights into new combination therapies with traditional drugs.

For further details, see Keul and Hamacher (2017).

1.3.4 Contribution

FK and KH devised the study. FK prepared the datasets, computations, illustrations and implemented the code. FK and KH jointly wrote the manuscript.

1.4 PFASUM: A substitution matrix from Pfam structural alignments

Substitution matrices are the foundation of modern bioinformatics with the selection of appropriate models describing evolutionary events being of the utmost importance. Here, substitution matrices based on outdated and (most likely) inaccurate databases are widely used. In light of this, we derived a novel substitution matrix from Pfam structural alignments and showed that this new matrix series outperforms most commonly used substitution matrices for homologous sequence searches and multiple sequence alignments. The complete analysis can be found in:

Keul, F; Hess, M; Goesele, M and Hamacher, K (2017) PFASUM: A substitution matrix from Pfam structural alignments, *BMC Bioinformatics*, *in submission*

1.4.1 Background

Over the last years, several substitution matrices families have been developed, differing in construction, underlying data bases and fields of application. Two of the most commonly used representatives for substitution matrices are the Point Accepted Mutation matrices(PAM) (Dayhoff and Schwartz, 1978) and the BLOcks SUBstitution Matrix (BLOSUM) (Henikoff and Henikoff, 1991). While in PAM amino acid mutation probabilities for specific evolutionary distances are generated through Markov chain models, BLOSUM-type matrices are generated from information of highly conserved and distantly related amino acid blocks. Furthermore, the chosen approach to substitution matrices of Dayhoff and Schwartz (1978) is the foundation for other matrices such as VTML (Müller and Vingron, 2000), while matrices such as OPTIMA (Kann *et al.*, 2000) and PBLOSUM (Song *et al.*, 2015) are based on BLOSUM matrices.

Even though, these matrices are widely used in various applications ranging from homologous sequence search over multiple sequence alignments to machine learning, the foundation for most substitution matrices are relatively old databases. While this is not necessarily a disadvantage, old databases tend to represent an outdated and incomplete picture of the sequence and structure space. Contrarily, databases such as Pfam (Sonnhammer *et al.*, 1997; Finn *et al.*, 2016) are updated on a regular basis, present hand curated seed alignments and package ever improving sequence and structure space coverage. Hence, we developed a novel algorithm capable to process gapped alignments and account for redundancies to derive so called PFASUM substitution matrices from Pfam seed alignments.

1.4.2 Methods

While BLOSUM and BLOSUM derived algorithms are based on ungapped multiple sequence alignments (MSAs), the PFASUM algorithm can summarize evolutionary substitution events on any alignment. Hereby, pair frequency counts are obtained similar to BLOSUM. To compensate for datasets with oversampled data and the introduction of bias thereof, the PFASUM algorithm includes the clustering method described by Henikoff and Henikoff (1991); Hess *et al.* (2016) which clusters sequences based on their similarity within a family of homologous sequences. Additionally, the PFASUM algorithm includes a novel methodological approach to process and account for ambiguous amino acids.

In order to evaluate substitution matrix performance, we assessed matrix performance for the novel PFASUM matrices as well as commonly used matrices found in SSEARCH (Pearson, 1991) on two levels. First, a comparison between these matrices on basis of their homologous sequence search results was conducted, using methods and tools published earlier (Brenner *et al.*, 1998; Price *et al.*, 2005; Hess *et al.*, 2016). Additionally, we benchmarked substitution matrices on their ability to reproduce reference alignments from the alignment benchmark datasets of BALiBASE 3.0, OXBench and SABmark 1.65 using MUSCLE. Hereby, all MUSCLE alignments use guide tree construction via *k*-mer clustering and subsequent generation of the MSA using substitution matrices. With this setup, the guide tree construction is independent of the substitution matrix and allows us to properly assess the MSA building qualities for each matrix without additional influence of the guide tree. By using three distinctively different MSA benchmark databases we are also capable to assess the performance of the substitution matrices on varying degrees of sequence similarity within alignments.

1.4.3 Results

Our results revealed that PFASUM-type matrices significantly outperform comparable default matrices when conducting homologous sequence searches. Especially for sequences with remote relation, PFASUM matrices showed increased coverage of sequence space and were able to identify more reliably these relations. While PFASUM matrices performed the best for remotely homologous sequences, their performance on datasets with closely related sequences also improved upon performance of existing matrices, albeit at a much lower level.

For the generation of MSAs, Fig. 1.3 shows an excerpt of our results. For MSA datasets of with relatively high sequence similarity, we find the tested PFASUM matrices only marginally bested by VTML200. On dataset with lower sequence similarity (BALiBASE 3.0 and SABmark 1.65) PFASUM matrices achieve significantly higher average alignment quality scores. Especially for alignments with very low sequence similarity PFASUM-type matrices showed a remarkable increase in alignment quality.

For further details, see Keul *et al.* (2017).

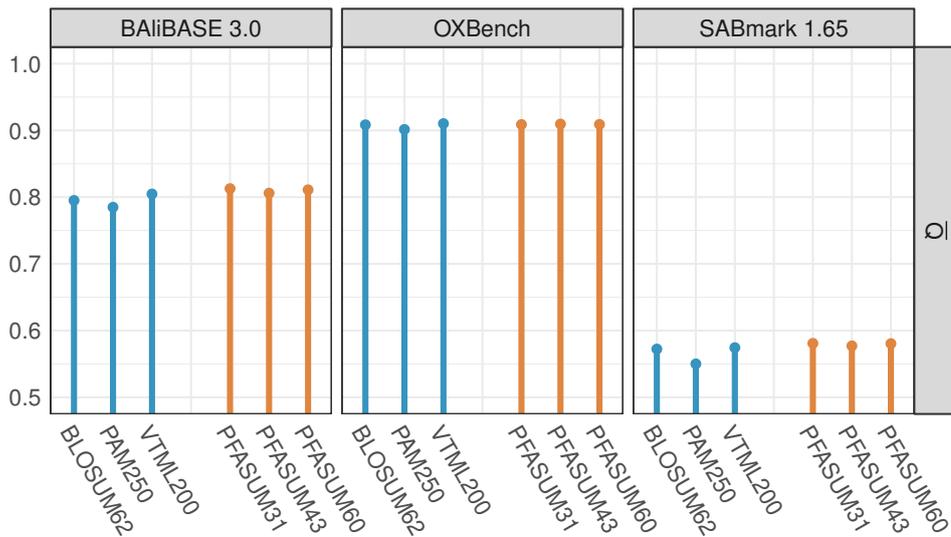


Figure 1.3.: Comparison of PFASUM-type and commonly used default matrices on the alignment benchmark datasets BAliBASE 3.0, OXBench and SABmark 1.65. PFASUM matrices (orange) outperform widely used standard substitution matrices (blue) based on the average pairwise accuracy (\bar{Q}) when aligning remote homologs.

1.4.4 Contribution

FK, MH, MG and KH jointly conceived the concept of the paper. MH and FK performed the data acquisition, analysis and manuscript as well as supplement preparation. KH and MG supervised the work and edited the manuscript. Code implementation was performed by MH and FK. FK prepared the figures. All authors discussed the results and implications and commented on the manuscript on all stages.

2 Channel-specific Substitution Matrices

In the post-genomic era, high throughput sequencing enables us to obtain fully sequenced genomes fast and relatively cheap. The number of annotated sequences in databases is ever-increasing (Thompson *et al.*, 2005; Andreeva *et al.*, 2008; O’Leary *et al.*, 2015) and sophisticated heuristics have been developed to search through large databases to find homologous sequences given a specific query sequence (Pearson and Lipman, 1988; Altschul *et al.*, 1990; Li *et al.*, 2012). While pairwise sequence comparison can annotate unknown sequences, insights on evolutionary and structural dependencies require higher numbers of sequences, aligned in so called multiple sequence alignment (Lockless and Ranganathan, 1999; Weil *et al.*, 2009; Boba *et al.*, 2010; Morcos *et al.*, 2011; Jones *et al.*, 2012; Ekeberg *et al.*, 2014).

Generating these multiple sequence alignments (MSAs) is a non-trivial task and in fact an NP-complete problem (Wang and Jiang, 1994). Additionally, factors such as evolutionary distance of the originating species and unknown evolutionary relation between protein families increase the difficulty to determine (quasi-)correct evolutionary relations. In order to capture evolutionary relationships on an amino acid level, simple machine learning approaches were developed such as substitution matrices (Dayhoff and Schwartz, 1978; Henikoff and Henikoff, 1992; Müller and Vingron, 2000; Hess *et al.*, 2016; Keul *et al.*, 2017).

The most popular sequence search tools and MSA algorithms use generalized substitution matrices in order to rate arbitrary input sequences (Pearson and Lipman, 1988; Altschul *et al.*, 1990; Thompson *et al.*, 1994; Katoh *et al.*, 2002; Edgar, 2004; Li *et al.*, 2012). These generalized evolutionary models are derived from databases consisting of a multitude of protein families, with different protein folds and functions (Dayhoff and Schwartz, 1978; Henikoff and Henikoff, 1992). Assuming universal evolutionary effects, these matrices aim at finding a single, average evolutionary model – essentially averaging evolution over all fold-, function- or family-specific models. More recent studies suggested that context specific approaches to homology search and multiple sequence alignments can yield more accurate results (Ng *et al.*, 2000; Vilim *et al.*, 2004; Leelananda *et al.*, 2016). Furthermore, a substitution matrix fitted to the compositional bias in Mollicutes had been shown to improve the prediction of homologous relationships for this organism class (Claire *et al.*, 2011).

Here, we will present a novel, family-specific class of substitution matrices derived from the Pfam database (Sonnhammer *et al.*, 1997; Finn *et al.*, 2016) using our PFASUM algorithm (Keul *et al.*, 2017). Based on the protein families for ion transport proteins and ion channels, we will compare the accuracy of these new and very specific substitution matrices to well-established and widely used BLOSUM-type matrices (Henikoff and Henikoff, 1992). We propose that these new substitution matrices generate alignments closer to ground truth alignments than their BLOSUM counterparts.

2.1 Background

Based on general principles of genetics, evolution of organisms is closely linked to the evolution of proteins in function, structure and the respective amino acid sequences. Changes in the coding nucleic acid sequence can originate from external, environmental factors – such as radiation – or can stem from the error prone amplification of the genome, to name two examples. Following genetics we can subdivide the term "mutation" in two different classes of events – large scale alterations and small scale variation of and within a DNA sequence. On the one hand, large scale mutations include phenomena such as entire gene duplications, transposition of chromosomal regions or deletions of these. On the other hand, we consider substitutions of nucleic acids, base triplet deletions and nucleic acid duplications as small scale changes in the DNA as these events only result in substitution, deletion or insertion of amino acids.

In an alignment of two closely related, homologous sequences, both deletion and insertion are represented by a missing amino acid (a so called *gap*). While we can stipulate that evolutionary distant sequences share fewer amino acids as two closely related sequences, we cannot determine the evolutionary age on the raw information of an amino acid sequence. Hence, we cannot distinguish between amino acid insertions in one and deletions in the other sequence since it is unclear which of the two sequences has to be considered ancestral to the other. Nevertheless, given enough data points, we can derive *evolutionary models* from these small scale evolutionary events to summarize observed changes and tendencies in the sequences of related and homologous proteins.

2.1.1 Evolutionary Models

Since the term *evolutionary model* is broad, we will focus on the two major models used to describe evolutionary behavior in the next paragraphs, as examples. Generally, evolutionary models "learn" evolutionary relationships on basis of a training database. These training datasets can range from rather limited evolutionary observations of closely related proteins (Dayhoff and Schwartz, 1978; Sonnhammer *et al.*, 1997) to using multiple structure based alignments (Hess *et al.*, 2016; Keul *et al.*, 2017) as well as conserved blocks within protein families (Henikoff and Henikoff, 1992). The resulting models can then be used to classify sequences in existing classes (Sonnhammer *et al.*, 1997; Finn *et al.*, 2016), to directly compare two or more sequences (Smith and Waterman, 1981; Altschul *et al.*, 1990; Pearson, 1991; Brenner *et al.*, 1998; Price *et al.*, 2005; Hess *et al.*, 2016) or to construct multiple sequence alignments (Edgar, 2004, 2009; Sievers *et al.*, 2011). While other methods generate evolutionary models through support vector machines (Cai *et al.*, 2003; Liao and Noble, 2003; Mohabatkar *et al.*, 2011), the most widely adopted concept for capturing these relations are the so called *substitution matrices*.

Foundation of Substitution Matrices

In general, substitution matrices are derived from databases of (somehow) aligned sequences (or sequence segments) as otherwise corresponding positions could not be matched. Two of the most prominent substitution matrices are PAM and BLOSUM. PAM-type matrices, introduced by Dayhoff

and Schwartz (1978), use a Markov chain model to express the relationship of more distant protein sequences. Here, the basis for the PAM1 matrix is the initial transition probability matrix of a Markov chain model. Probabilities for this initial transition matrix are derived from 1572 amino acid changes between very closely related sequences. PAM matrices with higher matrix numbers (e.g. PAM120 or PAM250) can easily be calculated by formulating the n -step transition probabilities (or 120-step and 250-step probabilities) based on this initial matrix following a Markov chain rule. In order to obtain the PAM n matrix, rounded log-odds ratios between the n -step transition probability matrix and the frequency of the observed amino acid in the underlying database are computed. PAM n entries represent substitution events after n evolutionary steps (n transitions) in relation to the initial probability of independent evolution. Through this, PAM matrices describe larger evolutionary events as product of minor evolutionary events. As a central property of regular Markov chains, PAM matrices converge to a limiting probability distribution.

The BLOSUM-type matrices on the other hand, are based on the BLOCKS database, a set of conserved domains of protein families (Henikoff and Henikoff, 1991). From these ungapped blocks the BLOSUM algorithm derives so called log-odd scores by relating the observed substitution frequencies $p(\alpha, \beta)$ of the amino acid α and β (with $\alpha, \beta \in$ canonic amino acids \mathcal{A}) to the statistically independent distribution of both amino acids (i.e. the product of the marginals $p(\alpha)$ and $p(\beta)$) as shown in Eqn. 2.1. Here, the probabilities $p(\alpha, \beta)$ are derived by counting all $\alpha\beta$ pairings ($n(\alpha, \beta)$) and relating this count to all counted pairs ($N = \sum_{\alpha, \beta} n(\alpha, \beta)$).

$$S_{\alpha, \beta} = \log_2 \frac{p(\alpha, \beta)}{p(\alpha) \cdot p(\beta)} \quad (2.1)$$

The marginals are derived by summing the probability of observing amino acid conservation $p(\alpha, \beta)$ and the sum of all off-diagonal entries ($\sum_{\beta, \beta \neq \alpha} p(\alpha, \beta)$) for a given residue α . To reduce the computational complexity, the resulting log-odd score $S_{\alpha, \beta}$ is then rounded to the next integer value. Forming the expectation value over all possible substitution events yields the relative entropy $I = \sum_{\alpha, \beta} S_{\alpha, \beta} p(\alpha, \beta)$ on basis of the non-rounded log-odds scores.

In order to reduce bias introduced through oversampling of "biologically interesting" protein sequences, Henikoff and Henikoff introduced a similarity based clustering of the sequences within sequence blocks. Before calculating the observed amino acid frequencies for a given block, sequences with a user selectable minimum similarity value/threshold such as 62% are clustered. Within each cluster, sequences are considered as a single sequence for counting pairs of α and β . Furthermore, observed acid substitutions $n(\alpha, \beta)$ between sequences of different clusters are reweighted by the corresponding cluster sizes (Henikoff and Henikoff, 1992). Hence, counts of amino acid pairs ($n(\alpha, \beta)$) include the weight of the cluster sizes instead of counting every amino acid fully.

Since its release in 1992, the BLOCKS database grew significantly and received its final update in 2007. Even though the database is outdated and it had been shown that the BLOSUM implementation by Henikoff and Henikoff (1992) exhibits influential coding errors (Styczynski *et al.*, 2008; Hess *et al.*, 2016), BLOSUM matrices on basis of the BLOCKS 5 (Henikoff and Henikoff, 1992) are accepted as default evolutionary model and are widely used by numerous applications.

Table 2.1.: Table of ambiguous amino acids in their one letter representation with their designated canonic amino acids.

Ambiguous amino acid	B	Z	J	X
canonic amino acid	N, D	E, Q	I, L	all

Family-Specific Substitution Matrices

While generalized substitution matrices, such as PAM- and BLOSUM-type, are widely used, they are based on the *average* amino acid substitution rate over a large set of sequences from a variety of protein families. In the recent years, performance improvements of these matrices have been investigated multiple times, ranging from compositional adjustment of the matrices (Altschul *et al.*, 2005; Lemaitre *et al.*, 2011) to generation of fold specific substitution matrices (Vilim *et al.*, 2004; Leelananda *et al.*, 2016). In the following, we will introduce the protein-family-specific substitution matrix computation, based on the generation of the PFASUM matrix (Keul *et al.*, 2017). Whereas the PFASUM matrix was generated on basis of all Pfam seed alignments, and thus represents the *average* substitution matrix for all hand curated alignments, the family-specific PFASUM matrices aim at capturing family-specific evolutionary tendencies, similar to substitution matrices generated from membrane bound proteins (Ng *et al.*, 2000).

As a proof on concept, we derived substitution matrices for two closely related protein families, the ion transport (PF00520) and the ion channel proteins (PF07885) and compare their performance to comparable BLOSUM-type matrices. Whereas the computation of PFASUM matrices is similar to BLOSUM, Pfam alignments make it necessary to account for gaps. Columns with high gap content can introduce a bias on amino acid substitution events and as such, the PFASUM computation omits all gaps for the calculation of column-wise pair frequencies. Furthermore, as only single protein families are used here, reweighting of amino acid pair counts based on the size of the protein family is not necessary.

While the BLOSUM computation neglects all occurring ambiguous amino acid characters – such as B, J, Z and X – PFASUM-type matrices include these in the matrix generation. Since ambiguous amino acids one letter codes encode at least two canonic amino acids, the counts of any found ambiguous amino acid is distributed equally in the PFASUM algorithm. We use a simple translation scheme for the four ambiguous amino acids (see Tab 2.1). Hence, each amino acid is represented by a set Θ of encoding symbols. The set of amino acids representing asparagine (N) is $\Theta_N = \{B, N\}$ while the set symbolizing alanine (A) is $\Theta_A = \{J, A\}$. Pairs of amino acids including one or two ambiguous amino acids take into account the cardinality of the encoded amino acid sets into account to evenly reweight the count for this found pairing.

With this, we can reformulate the pair frequency computation to this equation:

$$p(\Theta_i, \Theta_j) = \sum_k \frac{1}{Z_k} \left[\left(1 - \frac{1}{|\Theta_i||\Theta_j|} \right) n_k(\Theta_i, \Theta_j) + \frac{1}{|\Theta_i||\Theta_j|} \sum_{\alpha \in \Theta_i} \sum_{\beta \in \Theta_j} n_k(\alpha, \beta) \right] \quad (2.2)$$

Here, Z_k is $\sum_{\Theta_i, \Theta_j} n_k(\Theta_i, \Theta_j)$ with $i, j \in \mathcal{A}$ and $n_k(\alpha, \beta)$ describes the amino acid pair counts of the k th family of proteins.

The normalization based on the cardinality of the sets Θ , enables the PFASUM computation to effectively handle non-canonic amino acids (first summand) and their ambiguous character. For example, every found amino acid pair AB is counted with 0.5 to the pairings AN and AD as well as one full AB . As the actual frequencies within each set Θ_i is unknown, we chose to distribute the pair counts equally.

Similar to the generation of BLOSUM matrices, the PFASUM computation allows to cluster similar sequences based on a preset similarity score. By using the corrections proposed by Styczynski *et al.* (2008) and Hess *et al.* (2016), the PFASUM matrix generation calculates the pairwise sequence similarity based on the shortest length of both ungapped sequences. Hence, sequence fragments are always assigned to the cluster of the full length sequence, through the amended BLOSUM clustering algorithm (Hess *et al.*, 2016; Keul *et al.*, 2017). Pair counts derived from clustered sequences follow the BLOSUM computation (Henikoff and Henikoff, 1992). Log-odds scores are computed as described above by using Eqn. 2.2.

2.1.2 Multiple Sequence Alignments

Generating alignments of protein sequences is one of the most important tasks in modern day molecular biology and can be found in most computational methods (Pazos *et al.*, 1997; Al-Lazikani *et al.*, 2001; Pazos and Valencia, 2001; Gloor *et al.*, 2005; Pierri *et al.*, 2010; Morcos *et al.*, 2011; Jones *et al.*, 2015). As such, the quality of multiple sequence alignments (MSAs) underlying these methods critically influences the accuracy and results of these methods. Finding the *optimal* solution to aligning multiple protein sequence presents an important and difficult task and benchmarks of existing tools are performed on a regular basis (Blackshields *et al.*, 2006; Thompson *et al.*, 2011; Pais *et al.*, 2014). In general, all algorithms attempt the calculation of near optimal MSAs in manageable time using different approaches to the generation of MSAs such as progressive alignment procedures (Thompson *et al.*, 1994), profile Hidden Markov Models (Eddy, 1995) or evolutionary algorithms (Notredame and Higgins, 1996; Gondro and Kinghorn, 2007). In the upcoming section we will focus our attention on two different MSA algorithms. Whereas the theoretical background for both algorithms differs greatly, both using substitution matrices to generate MSAs with respect to this evolutionary substitution model. Furthermore, both algorithms have been reported to handle large sequence numbers well (Katoch *et al.*, 2002; Edgar, 2004).

MUSCLE and MAFFT

In 2004, Edgar presented the MUSCLE algorithm which achieved comparable accuracy to programs such as T-COFFEE (Notredame *et al.*, 2000) or MAFFT (Katoch *et al.*, 2002) while being significantly faster. The MUSCLE algorithm follows three essential steps: initial alignment, improving of the alignment and refinement of the alignment. At the start, alignments are generated progressively through a guide tree obtained via k-mer clustering. Furthermore, MUSCLE presents us the option to further refine the alignment through additional profile alignments, and therefore improving accuracy in comparison to purely progressive methods such as CLUSTAL-W (Thompson *et al.*, 1994). Contrary

to other algorithms, MUSCLE optimizes a log-expectation objective function which uses frequencies derived from substitution matrices.

Similar to MUSCLE, MAFFT uses progressive, dynamic programming to obtain multiple sequence alignments that can be refined iteratively (Kato *et al.*, 2002). Nevertheless, MAFFT employs the Fast Fourier Transform (FFT) to generate so called group-to-group alignments. For this, amino acids are translated into numeric vectors of physico-chemical properties of volume and polarity. By applying a sliding window approach to the FFT analysis of the correlation between the numeric property vectors of two sequences, MAFFT is able to consistently find homologous segments. Progressive alignment methods based on these homologous segments reduces computational time drastically and achieves similar or higher accuracy than purely progressive methods (Edgar, 2004).

2.2 Method

2.2.1 Pfam Database

The Pfam Database comprises of a multitude of MSAs for large numbers of protein families and super-families (Sonnhammer *et al.*, 1997; Finn *et al.*, 2016) which share similar fold characteristics and functional likeness. For each family at least two alignments are provided – the seed and the full alignment. The seed alignment represents a set of representative and manually aligned sequences for one protein family (Finn *et al.*, 2008). Full alignments are obtained after three steps using hidden Markov models (HMM) (Sonnhammer *et al.*, 1997):

- (i) Generation of profile HMM from seed alignments with HMMER (Durbin *et al.*, 1998).
- (ii) Searching the SwissProt database with the generated profile HMMs (Bairoch, 2004).
- (iii) Aligning the found sequences to the profile HMM to obtain the full alignment (Eddy, 1995).

In addition to seed and full alignments, Pfam provides so called RP dataset based on representative proteomes (Chen *et al.*, 2011). These datasets consist of a representative, yet more diverse subset of sequences of the full alignment.

2.2.2 Generation of Substitution Matrices

Evolutionary processes within membrane bound proteins typically differ from those observed in soluble proteins due to, e.g., varying entropic surface effects (Ng *et al.*, 2000). We derived family-specific substitution matrices of membrane bound proteins, evaluated their properties and compared their performance to traditional (family-unspecific) substitution matrices. These family-specific evolutionary models were, exemplarily, generated for the ion transport protein family (PF00520, 22625 sequences) and ion channel family (PF07885, 7664 sequences). Full alignments of the Pfam 29.0 dataset (the most recent at the time) were used at clustering threshold of 100 to obtain these matrices. To distinguish these matrices from the generalized PFASUM matrices based upon the entire Pfam seed database, we labeled them PFASUM00520 and PFASUM07885 accordingly.

Table 2.2.: Table of the used substitution matrices with their respective relative entropy.

substitution matrix	relative entropy	substitution matrix	relative entropy
PFASUM00520	0.2457 bit	BLOSUM38	0.2494 bit
PFASUM07885	0.4056 bit	BLOSUM47	0.4109 bit

In order to appropriately compare substitution matrices, both entry- and performance-wise, Altschul (1991) suggested that comparable substitution matrices inevitably possess similar matrix entropy. Hereby, the matrix entropy I is defined as the expectation value of the unrounded log-odds ratios between actually observed substitution events ($p(\theta_1, \theta_2)$) and independent evolution ($p(\theta_1) \cdot p(\theta_2)$). Hence, the relative entropy proposed by Altschul represents the Kullback-Leibler divergence between constrained and independent evolution of amino acids in the underlying data set ($I = D_{\text{KL}}(p(\theta_1, \theta_2) || p(\theta_1)p(\theta_2))$). To compare substitution matrices derived from different data sets, it is elementary to ensure that I for matrices is similar and adjust/chose parameters for the matrix generation accordingly (e.g. adjustment of clustering coefficient in BLOSUM or matrix number in PAM). This prevents introduction of artificial bias through different underlying database compositions.

Hence, we generated BLOSUM-type matrices with comparable relative entropy to both PFASUM-type matrices from on the BLOCKS 5 database (Henikoff and Henikoff, 1991) using the original code supplied by Henikoff and Henikoff (1992). The relative entropies for the obtained matrices are shown in Tab. 2.2. We note here, that comparison of substitution matrices with large relative entropy differences can lead to biased results. Hence, we refrain from the direct comparison of PFASUM00520 and PFASUM07885. Additionally, we restrict the comparison of the matrices to the 20 canonic amino acids as the calculation of ambiguous amino acids entries (such as Z or J) differs greatly between BLOSUM and PFASUM.

2.2.3 Comparison of Substitution Matrix Composition

To gain insights into the structure of the substitution matrices, we used two clustering algorithms to find groups of easily substitutable amino acid. By setting $k = 5$ in the k-means clustering algorithm (MacQueen *et al.*, 1967; Hartigan and Wong, 1979) we aimed at obtaining five distinct clusters of amino acids from the substitution matrices. By using the Fruchterman-Reingold algorithm for force-directed graph drawing (Fruchterman and Reingold, 1991) only based on positive substitution matrix entries, we achieve clustering of amino acids based on preferential substitution events. Fruchterman and Reingold developed an algorithm which use an edge-weighted adjacency matrix and reduces the distance between the nodes based on the magnitude of the weights connecting the nodes.

Theoretically, substitution matrices consist of log-odds scores which describe the relationship between observed and expected, independent substitution events. Positive scores represent often observed, likely substitutions and result in increased column entropy in MSAs while preserving block characteristics of domains. Negative scores represent unlikely substitutions and their influence is easily affected by the gap model chosen. Altogether, substitution matrices with few positive, non-diagonal log-odds scores assume – by principle – less amino acid substitution than matrices with more $S_{\alpha\beta} > 0$ for $\alpha \neq \beta$. Due to this we omitted these in the analysis of the substitution matrix composition. Here, evolutionary

interchangeability between physico-chemically similar amino acids should be found within substitution matrices as these aim at encapsulating physico-chemical and evolutionary dependencies. This allows us to evaluate the applicability of these matrices on basis of their amino acid cluster tendencies.

2.2.4 MSA Generation

In order to evaluate the performance of substitution matrices, we chose to generate MSAs from unaligned sequences of an existing MSA and compare the resulting MSAs with the original MSA. We used the Pfam RP15 datasets for the families PF00520 and PF07885 as reference alignments and their unaligned sequences as input for the two MSA generating tools MUSCLE (Edgar, 2004) and MAFFT (Kato *et al.*, 2002). MSAs were generated using family-specific substitution matrices, as well as their BLOSUM counterparts at a comparable entropic level under variation of the gap opening penalty (between 5 and 20) and the gap extension penalty (between 1 and 5). As MAFFT distinguishes between global and local gap penalties, we set gap penalties with `--op` and `--ep` options to the aforementioned values. We chose to rebuild the guide tree once (`--retree 1`) to improve the program speed. For all other parameter default values were used. In MUSCLE we adjusted the number of iterations per alignment step to 2 (`-maxiter 2`). All other parameter were left at their default value.

RP15 datasets represent a subset of sequences from full Pfam alignments used to construct both PFASUM matrices. These datasets are based on a set of representative proteomes (RPs) which reflect all other proteomes in their representative proteome group (RPG) and thus representing the majority of the sequence space. By reducing the so-called co-membership threshold to 15%, proteomes with at least 15% similarity are clustered in an RPG and representatives are selected subsequently. Hence, representative sequences from each RPG (in our case the RP15 dataset) represent sequences from distant RPGs and possess less similarity than the sequences of the full alignment.

To compare alignment algorithms and parameters, several measures have been proposed earlier. We use three distinct alignment quality measures, focusing on correctly aligned amino acids pairs as well as correctly aligned MSA columns.

Q-score The PREFAB quality score (Q-score) counts the number of correctly aligned pairs of residues and relates these to the number of residue pairings in the reference alignment (Edgar, 2004). Hence, this score type indicates the number of correctly aligned pairwise alignments of sequences.

TC score The BALiBASE total column score (TC score) relates the total number of correctly aligned MSA columns to the total number of columns in the reference alignment (Thompson *et al.*, 2005). Thus, this score evaluates the global alignment but is also prone to underestimating the alignment quality.

Cline score The Cline shift score measures the average – per pairwise alignment – positional shift necessary to convert the generated MSA to the reference MSA (Cline *et al.*, 2002), with higher number of shifts reducing the score.

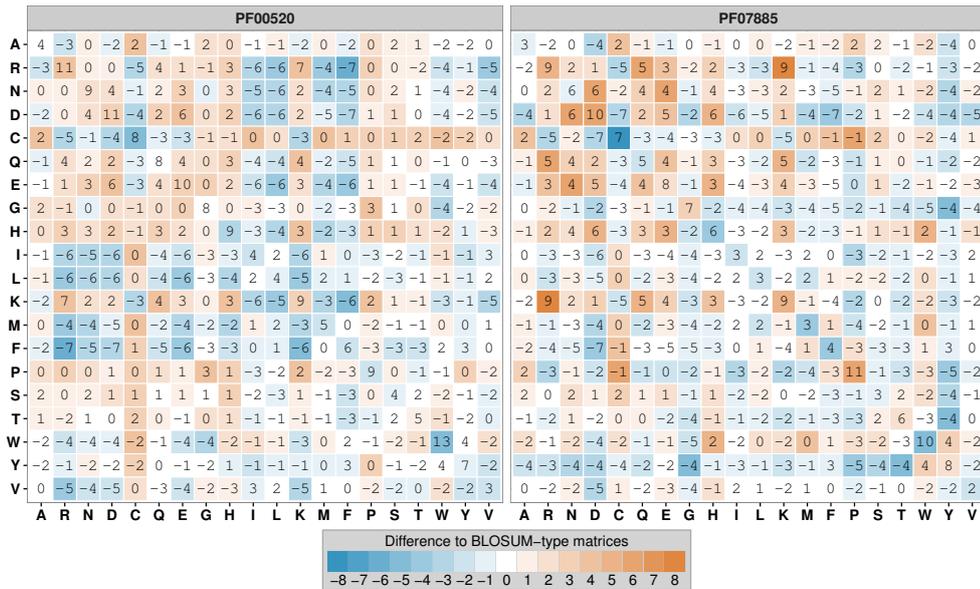


Figure 2.1.: Comparison of substitution matrix entries for PFASUM00520 (left hand side) and PFASUM07885 (right hand side) with BLOSUM38 and BLOSUM47, respectively. PFASUM-entries are displayed through numeric values while the differences between the rounded log-odd scores of the PFASUM-type matrix with its corresponding BLOSUM counterpart is represented through color-coding of the matrix entries. While the relative matrix entropy difference between both family-specific PFASUM-type matrices and their respective BLOSUM counterpart is small ($\ll 0.006$ bit) the entry-wise comparison between these differs to a great extend (81% of matrix entries)

By considering all three score types, we can gain insights into specific attributes of the generated alignments when compared to their respective reference.

2.3 Results

2.3.1 Substitution matrix comparison

To compensate for data set variances between the BLOCKS database (Henikoff and Henikoff, 1991, 1992) and the Pfam 29.0 release (Sonnhammer *et al.*, 1997; Finn *et al.*, 2016), we adjusted the BLOSUM clustering coefficient in concordance to the PFASUM100 matrices for both ion channel families (PF00520 and PF07885). Hence, we will focus the analysis of the substitution matrices and their performance on the two following matrix pairings:

- (i.) PFASUM00520 versus BLOSUM38, and
- (ii.) PFASUM07885 versus BLOSUM47.

Even though the PFASUM algorithm allows to adjust clustering coefficients similar to the BLOSUM algorithm, we limit our analysis to the PFASUM100 matrices for the two families. It is worth noting, that the relative entropy of PFASUM00520 is with 0.2457 bit significantly smaller than the relative entropy of PFASUM07885 at 0.405584 bit, indicating more independent substitution events (see Tab. 2.2).

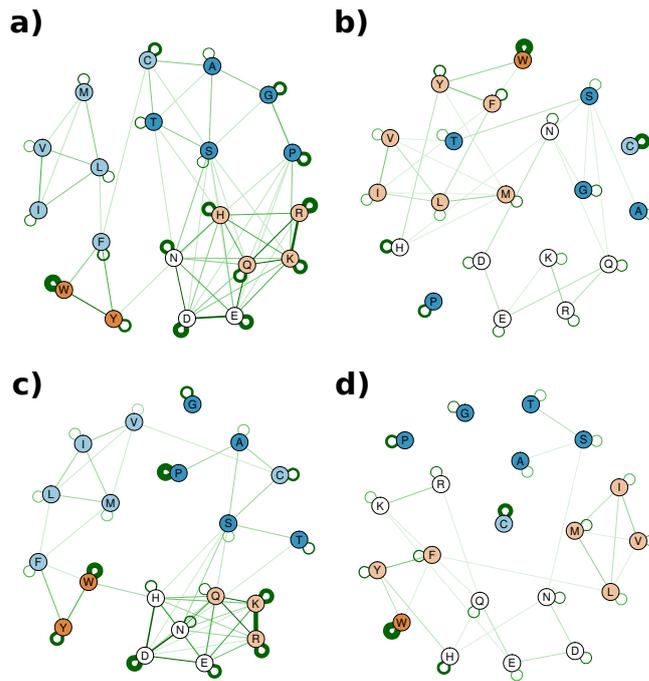


Figure 2.2.: Higher order structural organization within the substitution matrices. Color-coded are the five clusters found via k-means (5-means). Spatial organization of the nodes is obtained by applying the Fruchterman-Reingold algorithm for force-directed graph drawing on the substitution matrices by only considering positive matrix entries. **a)** shows the matrix organization for the substitution matrix obtained from the full alignment of PF00520 (PFASUM00520). **b)** depicts the BLOSUM38. **c)** shows PFASUM07885 derived from the full alignment of PF07885. **d)** shows the BLOSUM counterpart of PFASUM07885 – namely BLOSUM47. It is noteworthy that the both k-means and Fruchterman-Reingold algorithms yield similar clusters for the PFASUM-type matrices, whereas the BLOSUM matrices display structurally different entry organization.

Differences in Matrix Entries

Fig. 2.1 shows the differences between the two analyzed PFASUM-type matrices and their BLOSUM counterparts. While the relative matrix entropy between comparable matrices is similar (see Tab. 2.2), the entry-wise comparison of the substitution matrices reveals large differences varying from -8 to 8 . Hence, certain substitution events – such as the substitution of arginine to phenylalanine – occur multiple orders of magnitude less frequent in the sequences of 6 transmembrane domain channels than in the BLOCKS 5 database. We also find highly favored substitution events – such as the substitution of lysine to arginine in PF07885– where the matrix scores for certain evolutionary amino acid exchanges are substantially higher than the other scores. Overall, the PFASUM substitution matrices differ to their BLOSUM counterparts in $> 81\%$ of their entries, indicating substantially different matrices. Please note, the direct comparison between PFASUM00520 and PFASUM07885 cannot be made due to significantly different relative entropies.

Substitution Matrix Organization

In order to properly assess the differences in the composition of the substitution matrices in Fig. 2.1, we performed a cluster analysis by employing two different clustering algorithms. Via k-means clustering

Table 2.3.: Table of the clusters found with the k-means and the Fruchterman-Reingold algorithms.

Matrix	k-means clusters	Fruchterman-Reingold clusters
PFASUM00520	(AGPST) (CFILMV) (WY) (DEN) (HKQR)	(ACGPST) (DEHKNQR) (FWY) (ILMV)
PFASUM07885	(AGPST) (CFILMV) (WY) (DEHN) (KQR)	(ACPST) (DEHKNQR) (FWY) (ILMV) (G)
BLOSUM38	(AGPST) (C) (W) (DEHKNQR) (FILMVY)	(ACGSN) (FHILMTVWY) (DEKQR) (P)
BLOSUM47	(AGPST) (C) (W) (DEHKNQR) (FILMVY)	(ATV) (C) (P) (G) (ILMV) (FWY) (KR) (DEHNQ)

we grouped the amino acid types according to their matrix scores in five clusters (Hartigan and Wong, 1979) and visualized the distribution of positive log-odd scores using the Fruchterman-Reingold algorithm for force-directed graph drawings (Fruchterman and Reingold, 1991). Fig. 2.2 shows these results for all four analyzed substitution matrices and Tab. 2.3 summarizes the found clusters.

Based on the k-means results, both PFASUM-type matrices are similarly organized, while having distinctively different relative entropies. With the exception of the grouping of histidine, PFASUM07885 and PFASUM00520 can be clustered identically in 5 clusters. The Fruchterman-Reingold algorithm when applied on all positive matrix entries, organizes the matrices into similar clusters – with the lone exception being the cluster assignment of glycine. This results from the absence of positive (and hence favorable) substitution scores for glycine in the Pfam family PF07885. The clustering found by k-means comprises of clusters for hydrophobic amino acids (cysteine, phenylalanine, isoleucine, leucine, methionine and valine) and aromatic amino acids (tryptophan and tyrosine). Furthermore two clusters for polar and charged amino acids (aspartic acid, glutamic acid, histidine, lysine, asparagine, glutamine and arginine) are found for both matrices. The last cluster is formed from the amino acids alanine, glycine, proline, serine and threonine for PF00520. At first glance, amino acids within this cluster share few commonalities as glycine and alanine are considered small amino acids, while serine and threonine are members of the polar amino acid group. Interestingly, all amino acids in this cluster effect α -helical structures in protein with alanine possessing a high helix propensity (Pace and Scholtz, 1998), whereas glycine, proline, serine and threonine exhibit helix bending properties (Ballesteros *et al.*, 2000).

The k-means results for both BLOSUM-type matrices are identical. While we again find the cluster of helix influencing amino acids (alanine, glycine, proline, serine and threonine), the BLOSUM matrices exhibit a different matrix structure. Here, we find that all positively and negatively charged amino acids form a single cluster. Through the structure of the matrix both cysteine and tryptophan are forming their own cluster, as apparently only few substitutions of these two amino acids can be found in the BLOCKS database. Consequently, the last cluster consists of hydrophobic amino acids (isoleucine, leucine, methionine and valine) as well as the two aromatic amino acids (phenylalanine and tyrosine).

The clusters generated with the Fruchterman-Reingold force-directed graph drawing algorithm are similar for both PFASUM-type matrices. Here, we only considered positive entries in the substitution matrix for positioning of the vertices. Both PFASUM-matrices exhibit a high degree of connectivity between the charged amino acids, asparagine and glutamine, indicating evolutionary interchangeability between these. Similarly, a cluster of the aliphatic amino acids isoleucine, leucine, methionine and valine can be found. In general, PFASUM00520 and PFASUM07885 display similar matrix organizations, even though both matrices derive from different datasets and possess different relative

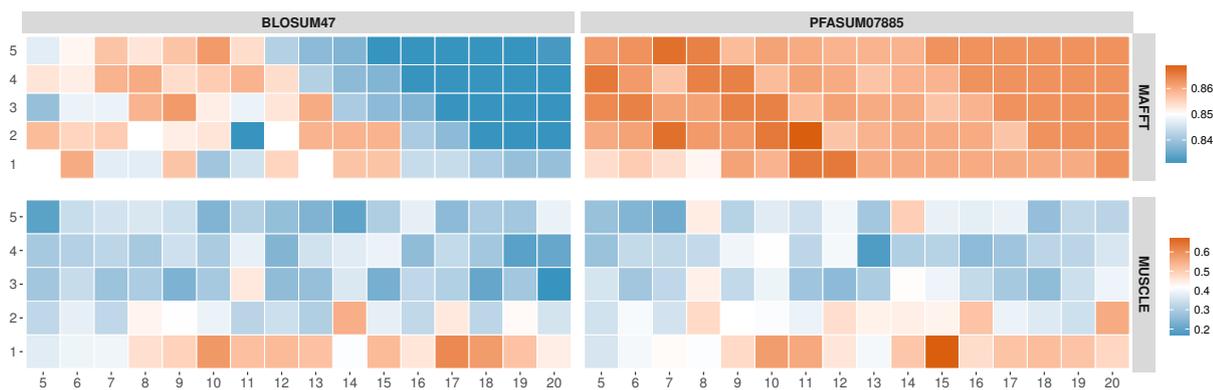


Figure 2.3.: Performance comparison via the Q-score of the BLOSUM47 and the family-specific PFASUM07885 substitution matrices on the RP15 dataset of small ion channel proteins (PF07885). Gap opening costs are shown on the abscissa, while the gap extension costs are shown on the ordinate. Q-score values are color-coded.

For the MAFFT multiple sequence alignment algorithm, PFASUM07885 outperforms its BLOSUM counterpart at any gap opening and extension combination. Similarly for the MUSCLE algorithm, alignments made with the BLOSUM47 matrix are less similar than those constructed using the PFASUM07885 matrix.

entropies. The clusters built via the Fruchterman-Reingold algorithm closely resemble the clusters found by the k-means algorithm and form groups of amino acids with similar physico-chemical properties.

Contrary to both PFASUM matrices, BLOSUM38 and BLOSUM47 present less propensity to form physico-chemical clusters. Both matrices experience similar clustering properties when subject to k-means clustering, forming three major clusters while two amino acids (see Tab. 2.3). They also combine positively and negatively charged amino acid in one cluster. Furthermore, the cluster of aliphatic amino acids found in both PFASUM matrices is extended with phenylalanine, while tryptophan – another aromatic amino acid – is not included in any other cluster. Likewise, we only found distorted clusters with the Fruchterman-Reingold algorithm when compared to the PFASUM matrices for both BLOSUM matrices. Here, as both BLOSUM only possess few positive off-diagonal entries and thus we can only determine very few clusters for BLOSUM47, while the clusters of BLOSUM38 show little resemblance to k-means clusters in PFASUM matrices.

Interestingly, the results for k-means and the Fruchterman-Reingold clustering differ drastically, even though family-specific PFASUM matrices and their BLOSUM counterparts possess similar relative entropies. The only cluster found in all four substitution matrices is formed by alanine, glycine, proline, serine and threonine, and appears to be a cluster of structurally influential residues.

2.3.2 Matrix performance generating alignments

PFASUM07885 outperforms its counterpart

While BLOSUM-type matrices are conventionally used as default parameter for finding or aligning amino acids sequences, our results for the RP15 dataset of PF07885 (ion channel family) show that

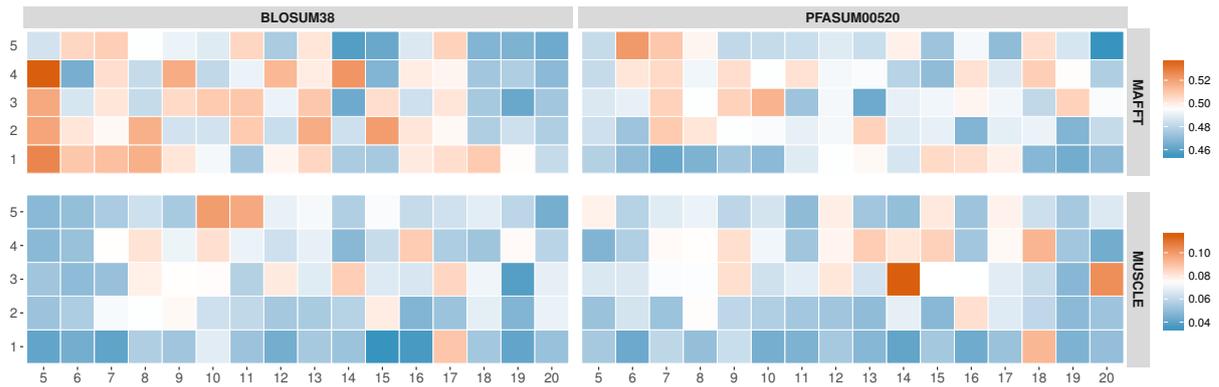


Figure 2.4.: Performance comparison through the Q-score of the BLOSUM38 and the family-specific PFASUM00520 substitution matrices on the RP15 dataset of ion transport proteins (PF00520). Gap opening costs are shown on the abscissa, while the gap extension costs are shown on the ordinate. Q-score values are color-coded.

For MAFFT generated MSA, BLOSUM38 outperforms PFASUM00520 in terms of maximal Q-score. Nevertheless, the opposite can be reported for MUSCLE generated alignments. It is noteworthy though, that MUSCLE generated MSAs showed significantly lower alignment quality than MAFFT based alignments, independent of the substitution matrix.

this choice can be suboptimal. The results for the PREFAB quality score (see Fig 2.3) and the Cline shift-score (see Fig. B.1) show that PFASUM07885 receives higher scores when compared to its BLOSUM counterpart for both alignment method investigated. In detail, MAFFT based alignments obtain substantially higher Q-score or Cline shift scores over a range of numerous gap opening and extension penalty combinations than their MUSCLE generated counterparts, independent of the chosen substitution matrix. In MUSCLE, gap opening and extension penalties strongly influences substitution matrix performance, again regardless of the used evolutionary substitution model. While MUSCLE appears less robust than MAFFT, our results for MUSCLE indicate a consistently higher performance of PFASUM07885 over BLOSUM47 for a wide variety of gap penalty combinations depicted through a more robust performance with Q-scores ≥ 0.4 . Observably, alignments generated with MAFFT present a higher residue-wise alignment precision than alignments built with MUSCLE, This fact is reflected by Q-score and Cline score value ranges above 0.84 for either substitution matrix, whereas the maximum for MUSCLE based Q-scores is at ~ 0.67 or at ~ 0.75 for the Cline shift score.

Interestingly, while we find high pairwise conformity of amino acids between the generated MSA and the reference alignment, only a small fraction of columns are correctly aligned (see Fig. B.2). Here, MUSCLE generated alignments show even higher agreement to the reference than alignments obtained from MAFFT. Nevertheless, both alignment tools and both substitution matrices fail to correctly align more than $\sim 9\%$ of all columns (BLOSUM47). These results can be attributed to the sensitivity of the BALiBASE total column score, as one misaligned residue in an entire MSA column results in a completely misaligned column. Hence, minor disarrangements in an MSA can reduce the BALiBASE total column score drastically.

Table 2.4.: Table of the best matrix and gap parameter combination for each score type, test dataset and alignment method.

Method	Family	score type	value	substitution matrix	gap parameter (gap open + gap extension)
MUSCLE	PF00520	Q-score	0.115	PFASUM00520	14 + 3
		Cline shift score	0.1	PFASUM00520	14 + 3
		Total Column score	0.0591	PFASUM00520	12 + 2
	PF07885	Q-score	0.674	PFASUM07885	15 + 1
		Cline shift score	0.748	PFASUM07885	15 + 1
		Total Column score	0.087	BLOSUM47	10 + 1
MAFFT	PF00520	Q-score	0.537	BLOSUM38	5 + 4
		Cline shift score	0.598	BLOSUM38	5 + 4
		Total Column score	0.0518	BLOSUM38	19 + 2
	PF07885	Q-score	0.869	PFASUM07885	11 + 2
		Cline shift score	0.903	PFASUM07885	11 + 2
		Total Column score	0.0598	BLOSUM47	15 + 5

Gap-rich PF00520 presents a complicated alignment challenge

Whereas the results for the RP15 sequences of PF07885 present a clear picture of the substitution matrix performance, aligning the sequences of the RP15 dataset of PF00520 appears to be more difficult. The Q-score for MUSCLE generated alignments ranged from 0.0335 to 0.115 (see Fig. 2.4). This marks a significantly performance drop for BLOSUM- and PFASUM-type substitution matrices when compared to the RP15 dataset of PF07885 (see Fig. 2.3, lower panel). These results indicate that MUSCLE generated PF07885 RP15 alignment share little to no similarity with the reference alignment when considering all amino acid pairings. The Q-score for in MAFFT generated MSAs lies in the interval (0.454, 0.537) and shows lower alignment quality than any PF07885 based alignment. Cline shift score evaluation of the generated alignments shows similar results for both alignment algorithms when compared to the Q-score results (see Fig. B.3). Again, the total column scores for MUSCLE and MAFFT showed that only few entire MSA columns were correctly aligned by both algorithms.

It is noteworthy, that the highest score for the PF00520 RP15 sequences was achieved through using the BLOSUM38 in combination with MAFFT. The best score for MUSCLE generated alignments was obtained with the family-specific PFASUM00520. Nevertheless, our results show that regardless of the substitution matrix, both alignment algorithms exhibited problems correctly aligning sequences of the PF00520 RP15 dataset when compared to the gap-rich reference alignment.

Overall performance evaluation

PFASUM substitution matrices when used in combination with MAFFT achieved the highest Q-score and Cline shift score for both test datasets (see Tab. 2.4). Furthermore, PFASUM07885 outperformed

its BLOSUM counterpart on the PF07885 RP15 dataset while PFASUM00520 was surpassed in performance by BLOSUM38 on RP15 sequences of PF00520.

Overall, for Q-score alignment evaluation, PFASUM-type matrices performed $\sim 66\%$ of all parameter sets better than the comparable BLOSUM-type matrix. A similar statement can be made for the Cline shift score. Roughly 69% of the 160 MAFFT and MUSCLE alignments generated with a PFASUM-type matrix achieved a higher score than their BLOSUM-based counterparts. Here, over a wide variety of parameter combinations Pfam-based and thus family-specific substitution matrices outperformed their respective BLOSUM counterparts. When considering only the performance on the well-behaved (i.e. alignments with gap scarcity), the performance improvements over BLOSUM-type matrices occur in over 96% of cases for Q-score and over $\sim 64\%$ of cases for Cline shift score for both alignment algorithms.

MAFFT based alignments achieved higher scores for both local alignment evaluation scoring algorithms. In 73.75% of all tested gap penalty models, PFASUM-type matrices outperformed their BLOSUM counterpart on basis of the Q-score. Subsequently, the number of correctly aligned residue pairs for PFASUM-based alignments is higher than those from BLOSUM-based alignments. When considering the Cline shift score, PFASUM originating alignments reach in 78.75% of parameter combinations a higher score than their counterparts. For the MUSCLE algorithm the BLOSUM-type matrices BLOSUM38 and BLOSUM47 are still outclassed by their respective counterpart in 58.75% and 60% of the parameter combinations – for Q-score and Cline shift score, respectively.

2.4 Discussion

As a proof of principle we derived family-specific PFASUM substitution matrices from structure based alignments of the Pfam database. We directly compared these new substitution matrices to matrices derived from the BLOCKS 5 database (BLOSUM38 and BLOSUM47). In general, we showed that PFASUM matrices are significant different in terms of matrix entry composition and organization. Family-specific PFASUM matrices deviate in numerous entries from the *generic* BLOSUM-type matrices. These findings indicate vastly different evolutionary processes happen in the two Pfam families used here – namely the ion transporter family (PF00520) and the ion channel family (PF07885).

In the PFASUM07885 substitution matrix, we find glycine without positive substitution score (other than to itself). This is very different to PFASUM00520 where three other amino acids can be substituted for glycine – proline ($S_{GP} = 3$), alanine ($S_{GA} = 3$) and even serine ($S_{GS} = 3$). While both substitution matrices originate from seemingly similar protein families, we can observe a large difference between these two evolutionary models for glycine. This disparity points at a different role for glycines in PF07855 than in PF00520, with a higher evolutionary pressure on the preservation of glycine. In smaller ion conducting structures (PF07885), glycines are irreplaceable for both structure and function, indicated by an log-odds ratio score below 0 and thus disfavoring substitution events altogether. Overall, both PFASUM matrices present clear cluster of amino acids similar physico-chemical properties, when compared to their respective BLOSUM counterpart. Hence, we can assume that in both ion conducting protein families evolutionary effects are governed by these properties in contrast to a less stringent evolutionary pressure when averaged over large numbers of proteins we observe in BLOSUM-type matrices.

Furthermore, we compared generated alignments from MAFFT and MUSCLE on basis of RP15 sequences to the aligned RP15 sequences of Pfam. Generally, Pfam alignments are obtained through alignment with a profile Hidden Markov Models (HMM) based on hand curated Pfam seed alignments (Sonnhammer *et al.*, 1997; Eddy, 1998). To this end, the Pfam full alignments used to derive the substitution matrices originate from sequences found through a search with the profile HMM and subsequent alignment with the profile HMM. Since the RP15 dataset represents a more diverse subset of full alignment sequences with reduced average sequence similarity, alignments for learning of the model (full alignment) and test alignment (RP15) originate from the same model – the profile HMM of the seed alignment. Even though this relationship exists, it is surprising that the PF00520 based PFASUM matrix failed to improve upon the performance of its BLOSUM counterpart. Nevertheless, significantly lower scores for alignments obtained from the MUSCLE and MAFFT algorithms indicate that both alignment tools struggle with the sequence diversity in the RP15 dataset. Hence, neither algorithm was able to consistently achieve high alignment evaluation scores when comparing newly generated alignments to the gap-rich reference alignment.

Our study shows that – at least for the ion transducing protein families PF00520 and PF07885 – MAFFT generated alignments resemble the reference alignment closer than comparable alignments from MUSCLE. While both algorithms use substitution matrices to generate multiple sequence alignments, their matrix usage is very different. In MAFFT, substitution matrices are used to refine the initial alignment obtained through group-to-group alignment by Fast Fourier Transform. On the one hand, the MUSCLE algorithm integrates the substitution matrix beginning with the initial alignment based on the k-mer derived guide tree. When compared to MAFFT, the basis for MUSCLE appears to be an error prone initial alignment which is dependent on the substitution matrix. Due to the structural basis of the Pfam datasets, MAFFT performs better at the stage of the unrefined, initial alignment for such, regardless of the used substitution matrix.

In summary, for reasonably sized and well-behaved sequence datasets, we were able to show that PFASUM-type matrices improve upon the performance of BLOSUM-type matrices in overall amino acid pairing accuracy.

3 Evolutionary Information in Sequence Data

Since the introduction of the concept of information theory through Hartley (1928) and later definitions through Shannon (1948), information theory found its way into various fields of research ranging from data compression algorithms (Ziv and Lempel, 1977) and cryptography (Shannon, 1949; Stinson, 2005) to analysis of co-evolution in protein sequence datasets (Gloor *et al.*, 2005; Hamacher, 2011; Morcos *et al.*, 2011). Here, information theory tries to capture patterns and transmission events through analysis of discrete probability distributions of finite-sized data. In the following chapter, we introduce a novel approach to capture evolutionary behavior in proteins. With this measure we are able to directly compare the complexity of intricate, evolutionary mechanisms. We investigate the difference in co-evolutionary complexity between two different ion transporting protein families. Furthermore, we reveal differing patterns of co-evolution in prokaryotic and eukaryotic potassium channels as well as compare these results to commonly used information theoretical measures. Additionally, we examine the robustness of this novel and other approaches when the underlying amino acid distributions are perturbed.

Throughout the chapter we will use boldfaced typeset for vectors \mathbf{v} and matrices \mathbf{M} . $|C|$ denotes the cardinality of C .

3.1 Background

The foundation of information theory was set by Shannon in 1948 who formulated the general principle to measure uncertainty through entropy H . The so called Shannon entropy is the expectation value of information of a discrete random variable X . Hence, the entropy $H(X)$ can be described as:

$$H(X) := \left\langle \log_2 \frac{1}{p(X)} \right\rangle_p = - \sum_{x \in \mathcal{X}} p(x) \log_2 p(x) \quad (3.1)$$

Here, $p(x)$ stands for the probability mass function of X in the alphabet \mathcal{X} and is derived from the counts $n(x)$ in relation to the sum of all counts (i.e. $p(x) = n(x)/\sum n(x)$). The logarithmic base determines the unit of the entropy with a base of 2 resulting in *bit* as a unit. Shannon's entropy can be understood as the expectation value of the information contained within the random variable X given its probability mass function p , with the information defined as $I(X) = -\log_2 p(X)$. Conventionally, as the probability mass function can be 0 for certain realizations in the alphabet, we will use $0 \log_2 0 \equiv 0$ which directly follows from L'Hôpital's rule for limits of intermediate forms.

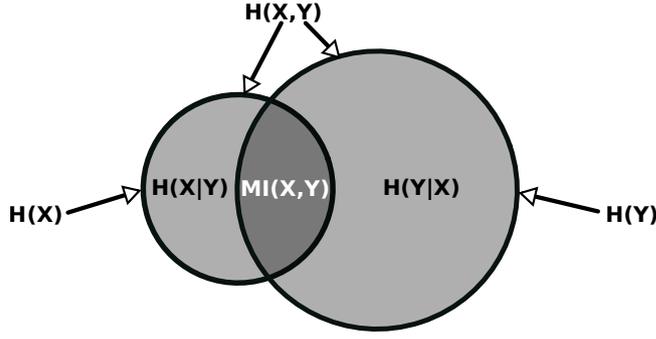


Figure 3.1.: Relationships between Shannon's entropy, joint entropy, conditional entropy and mutual information.

Likewise, a basic extension of the single random variable entropy (see Eqn. 3.1) is the so called joint entropy for a pair of discrete random variables X and Y . Comparable to Shannon's entropy the joint entropy reads:

$$H(X, Y) := - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 p(x, y) \quad (3.2)$$

One can easily see a close relation between Shannon's entropy and the joint entropy. Here, the joint entropy $H(X, Y)$ can be obtained from the entropy $H(X)$ through the addition of the conditional entropy $H(Y|X)$ as $H(X, Y) = H(X) + H(Y|X)$. The conditional entropy $H(Y|X)$ is the necessary information to describe Y when the probability distribution for X is already known.

While Shannon entropy measures the uncertainty of a random variable (or a pair of random variables), relative entropies such as the Kullback-Leibler divergence (D_{KL}) measure the distance between two distributions \mathbf{p} and \mathbf{q} (Kullback and Leibler, 1951). Here, \mathbf{p} is the short form of $p(X)$. The D_{KL} is defined as:

$$D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) := \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{p(x)}{q(x)} \quad (3.3)$$

As one can clearly see, the D_{KL} is an asymmetric measure of the distance between \mathbf{p} and \mathbf{q} with additive properties. As such, we can obtain $H(\mathbf{q})$ through $H(\mathbf{p}) + D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q})$. Hence, the D_{KL} measures the additional information necessary to arrive at the distribution \mathbf{q} when starting with the distribution \mathbf{p} . Technically, if there is any symbol $x \in \mathcal{X}$ such that $p(x) > 0$ while $q(x) = 0$, $D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) = \infty$. While often used as distance measure, the D_{KL} does not fulfill the metric criteria of symmetry ($D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) \neq D_{\text{KL}}(\mathbf{q} \parallel \mathbf{p})$) and subadditivity ($D_{\text{KL}}(\mathbf{p} \parallel \mathbf{q}) \leq D_{\text{KL}}(\mathbf{p} \parallel \mathbf{m}) + D_{\text{KL}}(\mathbf{m} \parallel \mathbf{q})$).

In order to compensate for these short-comings of the D_{KL} Lin (1991) derived a divergence measure similar to the D_{KL} on basis of Shannon's Entropy and Jensen's inequality (Jensen, 1906). This so called Jensen-Shannon divergence measures the similarity between two distributions by comparing each to average distribution.

$$D_{\text{JS}}(\mathbf{p}, \mathbf{q}) := \pi_1 D_{\text{KL}}(\mathbf{p} \parallel \mathbf{m}) + \pi_2 D_{\text{KL}}(\mathbf{q} \parallel \mathbf{m}) \quad (3.4)$$

with $\mathbf{m} = \pi_1 \mathbf{p} + \pi_2 \mathbf{q}$

Here, π_k corresponds to the weight of the k th probability distribution and follows $\sum_k^n \pi_k = 1$ and $\forall k \in n : \pi_k \geq 0$ with n being the number of distributions compared. Conventionally, as the mixture of all distributions introduces additional parameters, π_k is set to form the arithmetic mean of all distributions ($\pi_k = 1/n$). Hence, for $n = 2$ with $\pi_1 = \pi_2 = 0.5$, D_{JS} as formulated above is bound within the interval $[0, 1]$ and achieves its upper boundary *iff* the distributions \mathbf{p} and \mathbf{q} are completely divergent. Furthermore, Endres and Schindelin (2003) showed that $\sqrt{D_{JS}}$ possesses properties of a true metric, a property we will exploit later on.

The mutual information (MI, $I(X, Y)$) – another often used and well known information theoretic measure – is based on the Kullback-Leibler divergence. The MI can be directly derived from the D_{KL} and measures the information content one discrete random variable X holds about another discrete random variable Y .

$$I(X, Y) := \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)} \quad (3.5)$$

Here, $p(x, y)$ represents the probability mass function of the joint distribution of the random variables X and Y , while $p(x)$ and $p(y)$ represent the marginal distributions. Due to this, that MI measures the divergence between the observed joint distribution and the product distribution $p(x)p(y)$, i.e. the stochastic independence of both random variables (see Eqn. 3.4 and Fig. 3.1). Effectively, the MI represents the decrement of uncertainty in one random variable through knowledge of another. Due to the properties of the probability mass function, the boundaries of MI are directly linked to the number of discrete bins found (the alphabet size $|\mathcal{X}|$). Here, we find that the mutual information lies within the interval $[0, \log_2 |\mathcal{X}|]$.

3.1.1 Mutual Information normalizations and corrections

Mutual information is heavily employed in various fields ranging from image processing (Maes *et al.*, 1997; Suri and Reinartz, 2010), machine learning (Battiti, 1994; Liu *et al.*, 2009), to neuroscience (Borst and Theunissen, 1999; Szczepanski *et al.*, 2011) and structural biology (Shackelford and Karplus, 2007; Gomes *et al.*, 2012). Instances where it is necessary to compare MI results give rise to normalizations of MI as the boundaries of MI are directly linked to the number of discrete bins (i.e. alphabet size). Additionally, introduction of error or bias corrections due to database composition or size, caused the implementation of MI corrections. In the following we will introduce all MI normalizations and corrections that are used later on.

Entropy-based normalizations

Entropy-based normalization schemes use, in general, upper bounds of the mutual information. By relating MI to a function of the single random variable entropies, estimations about the magnitude are possible. One such approach is the *redundancy* ($R(X, Y)$). The redundancy measure is closely

related to the MI normalization in respect to the arithmetic mean of both entropies (see 3.6) (MI_{arith} or *uncertainty*).

$$MI_{\text{arith}}(X, Y) = 2R(X, Y) := 2 \frac{I(X, Y)}{H(X) + H(Y)} \quad (3.6)$$

Others have proposed using the geometric mean of both single point entropies to normalize MI, similarly to normalization of covariances and in correspondence to the inner product in Hilbert space (Strehl and Ghosh, 2002; Zaki, 2015). Hence, the MI normalization is:

$$MI_{\text{geom}}(X, Y) := \frac{I(X, Y)}{\sqrt{H(X)H(Y)}} \quad (3.7)$$

Alternatively, Rueckert *et al.* (1998) and Studholme *et al.* (1999) proposed to normalize the MI through the joint entropy in their *normalized mutual information* (NMI) measure. This normalized MI variant is:

$$\begin{aligned} NMI(X, Y) &:= \frac{H(X) + H(Y)}{H(X, Y)} = \frac{I(X, Y)}{H(X, Y)} + 1 \\ &\approx \frac{I(X, Y)}{H(X, Y)} = MI_{2P}(X, Y) \end{aligned} \quad (3.8)$$

Without loss of generality, we can omit the constant to obtain MI_{2P} on basis of the proposed NMI which is used heavily in image registration and alignment (Rueckert *et al.*, 1998; Studholme *et al.*, 1999; Cahill, 2010). While $H(X, Y)$ presents a natural, upper boundary of MI, we can easily see from Eqn. 3.5 that the mutual information is also bound by the minimum of both entropies $H(X)$ and $H(Y)$. Logically, the MI normalization MI_{min} is:

$$MI_{\text{min}}(X, Y) := \frac{I(X, Y)}{\min(H(X), H(Y))} \quad (3.9)$$

Since the relationship between this minimum of entropies and all other normalization denominators is $\min(H(X), H(Y)) \leq \sqrt{H(X)H(Y)} \leq 0.5(H(X) + H(Y)) \leq H(X, Y)$ as per Vinh *et al.* (2010), all MI normalizations listed here represent upper boundaries of MI with $\min(H(X), H(Y))$ being the closest upper boundary.

Row-Column Weighting and Average Product Correction

While the aforementioned MI normalizations aim at unifying the range of MI through consideration of different upper boundaries, Row-Column Weighting of mutual information (RCW) proposed by Gouveia-Oliveira and Pedersen (2007) focuses on eliminating bias introduced through sequences sampling of pharmaceutical relevant (human) sequences and phylogeny.

$$RCW(X, Y) := \frac{2I(X, Y)}{\bar{I}(X, \cdot) + \bar{I}(\cdot, Y)} \quad (3.10)$$

Here, we relate the observed mutual information $I(X, Y)$ between two distributions X and Y from a set of distributions $\xi = \{X, Y, \dots\}$ to the mean of the average mutual information of X ($\bar{I}(X, \cdot)$) and Y ($\bar{I}(\cdot, Y)$). These average mutual informations runs over all distributions in ξ and is formulated as $\bar{I}(X, \cdot) = \frac{1}{|\xi|-1} \sum_{P \in \xi, P \neq X} I(X, P)$. Hence, row-column weighting of MI relates the mutual information to the arithmetic mean of the average MI for the distributions X and Y .

In similar fashion, Dunn *et al.* (2007) derived the average product correction of MI through determining background MI values. Dunn *et al.* (2007) assumed that MI itself is influenced by an additive term of background MI. By removing this residual background MI from the actual mutual information between two distributions, the APC was able to estimate accurate co-evolutionary signals. Contrary to Dunn *et al.* (2007), we denote the average product corrected MI as APC:

$$\text{APC}(X, Y) := I(X, Y) - \frac{\bar{I}(X, \cdot)\bar{I}(\cdot, Y)}{\bar{I}(\cdot, \cdot)} \quad (3.11)$$

Similarly to the RCW, $\bar{I}(X, \cdot)$ denotes the mean MI for distribution X averaged over all other distributions (except X). $\bar{I}(\cdot, \cdot)$ describes the average mutual information over all distribution combinations (without the combination X and Y).

Z-score statistics

In 2009, Weil *et al.* investigated the influence of finite size effects of available data, as well as the influence of phylogenetic related sequences and overall sequence conservation. Through shuffling the data and without affecting the underlying distributions X and Y the authors generated a reference MI, independent of co-evolutionary relations. This MI *nullmodel* contained only residual background joint-entropy $H(X, Y)$ as X and Y were shuffled independently, while the single point entropies $H(X)$ and $H(Y)$ remained unchanged. Through this procedure Weil *et al.* (2009) were able to calculate Z-score statistics for MI by repetitive shuffling. Through this, it was possible to compute the average MI for the distributions X and Y ($\bar{I}(X, Y)$) as well as the corresponding variance $\sigma_{\text{MI}(X, Y)}^2$. By assuming an underlying Gauss distribution for the shuffled mutual information one can easily compute significance of an observed MI value in units of standard deviations through Z-score statistics. The generalized form for these is:

$$Z_f(X, Y) = \frac{f(X, Y) - \bar{f}(X, Y)}{\sqrt{\sigma_{f(X, Y)}^2}} \quad (3.12)$$

The Z-score statistics can then be computed as in see Eqn. 3.12 for any given function $f(X, Y)$, an average function $\bar{f}(X, Y)$ and the variance $\sigma_{f(X, Y)}^2$ of $f(X, Y)$. Through these, it is possible to measure the divergence (in standard deviations) of $f(X, Y)$ to an independent nullmodel $\bar{f}(X, Y)$. Here, $f(X, Y)$ represents any function which uses both distributions X and Y . Due to this, we are able to extend the Z-score computation for MI (Z_{MI}) to APC (Z_{APC}) and RCW (Z_{RCW}) by replacing the generic function $f(X, Y)$ with $\text{APC}(X, Y)$ and $\text{RCW}(X, Y)$, respectively.

3.1.2 Direct Coupling Analysis

Direct coupling analysis (DCA) represents a methodological approach inspired through statistical mechanics which was successfully applied to determining spatial residue-residue interactions based on alignments of protein sequences (Weil *et al.*, 2009; Morcos *et al.*, 2011; Baldassi *et al.*, 2014; Asti *et al.*, 2016). In general, DCA relies on the inference of a probabilistic model fitted on sequence data, which are able to describe variability in sequence space of large multiple sequence alignments (MSAs).

In DCA, the maximum-entropy principle is applied, leading to the statistical model $P(S_1, \dots, S_N)$ based on the Boltzmann distribution for observing a particular amino acid sequence S_1, \dots, S_N . Here, the partition function is formed in respect to pairwise couplings $\mathbf{J}_{ij}(S_i, S_j)$ and local fields $\mathbf{h}_i(S_i)$ of these columns and reads:

$$P(S_1, \dots, S_N) = \frac{1}{Z} \exp \left\{ \sum_{i < j} \mathbf{J}_{ij}(S_i, S_j) + \sum_i \mathbf{h}_i(S_i) \right\} \quad (3.13)$$

with

$$Z = \sum_{S_1, \dots, S_N} \exp \left\{ \sum_{i < j} \mathbf{J}_{ij}(S_i, S_j) + \sum_i \mathbf{h}_i(S_i) \right\}$$

Note, indices run from 1 to N . Through small coupling extension and mean-field approximation (Georges and Yedidia, 1991; Morcos *et al.*, 2011), the coupling strengths in 3.13 can be derived directly from the correlation matrix $\mathcal{C}_{ij}(\alpha_l, \alpha_k) = f_{ij}(\alpha_l, \alpha_k) - f_i(\alpha_l)f_j(\alpha_k)$ through $\mathbf{J}_{ij}(\alpha_l, \alpha_k) = -(\mathcal{C}^{-1})_{ij}(\alpha_l, \alpha_k)$ and are dependent on the amino acid frequencies $f_i(\alpha_l)$ of the amino acid α_l in the i th alignment column. Here, we find $k, l \in \{1, \dots, |Q|\}$ with Q being the alphabet. Furthermore, the marginal counts of $P(S_1, \dots, S_N)$ are necessarily determined by the empirical (pair) counts of the MSA so that $f(\alpha_i) = \sum_{S_k | k=i} P(S_1, \dots, S_N)$ and $f(\alpha_i, \alpha_j) = \sum_{S_k | k=i, j} P(S_1, \dots, S_N)$.

The inversion of the empirical correlation matrix based on frequency counts for single site and pair frequencies ($f_i(\alpha_l)$ and $f_{ij}(\alpha_l, \alpha_k)$ respectively) yields the $|Q| \times |Q|$ coupling matrix \mathbf{J}_{ij} (with $|Q|$ being the alphabet size). Additionally, through addition of a pseudo-count, the DCA includes a uniform prior distribution. For each $|Q| \times |Q|$ coupling matrix \mathbf{J}_{ij} direct couplings can be derived, effectively compressing the direct information for each amino acid pair coupling to a scalar position coupling value (Morcos *et al.*, 2011).

Another DCA based approach, the GaussDCA, uses multivariate Gaussian modeling of the multiple sequence alignments to obtain information on contact pairs (Baldassi *et al.*, 2014). Here, an *a priori* distribution is used to compensate for sampling bias of the sequence data in form of a pseudo-count, mirroring a uniform distribution. In general, GaussDCA aims at fitting a model derived from multivariate Gaussian modeling on sequence data in form of the MSA through maximizing the likelihood of probability distribution ($P(S_1, \dots, S_N)$) given the empirically observed sequence data (S_1, \dots, S_N). Hence, the GaussDCA approach results in a model most likely representing the MSA data.

Within the GaussDCA, amino acid sequences are converted to a binary representation from which an empirical covariance matrix (\mathcal{C}) is derived. Here, each residue α_l^m (l th residue of the m th sequence) is converted to Q binary variables $\mathbf{x}_{l,(1,\dots,Q)}^m$. Here, each binary variable represent one canonic amino acid. Hence, this vector of binary variables takes on 1 if the amino acid is present in the m th position of the l th sequences and zero otherwise. Consequently, if for the l th amino acid $\forall i \in 1, \dots, Q : \mathbf{x}_i^m = 0$, we observed a gap at residue α_l^m . Hence, \mathcal{C} is derived as:

$$\mathcal{C}_{ij} = \frac{1}{M} \sum_{m=1}^M (\mathbf{x}_i^m - \bar{\mathbf{x}}_i) (\mathbf{x}_j^m - \bar{\mathbf{x}}_j) \quad (3.14)$$

with

$$\bar{\mathbf{x}}_i = \frac{1}{M} \sum_{m=1}^M \mathbf{x}_i^m$$

Here, $\bar{\mathbf{x}}_i$ represents the empirical mean over all M sequences of the alignment for the i th entry in the binary encoded alignment, with the covariance matrix \mathcal{C} being a $(L \cdot Q) \times (L \cdot Q)$ -matrix with L being the length of all sequence in the multiple sequence alignment.

The multivariate Gaussian distribution is defined for the vector of N variables \mathbf{x} , the vector of N means $\boldsymbol{\mu}$ and the $N \times N$ covariance matrix Σ as:

$$P_G(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^N |\Sigma|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\} \quad (3.15)$$

Here, $|\Sigma|$ represents the determinant of the covariance matrix. The parametrization of the exponent $E(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$ can be transformed into $E(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{2} \mathbf{x}^T \mathbf{J} \mathbf{x} - \mathbf{h}^T \mathbf{x}$ with the interaction matrix $\mathbf{J} = -\Sigma^{-1}$ and the local fields $\mathbf{h}^T \mathbf{x} = \Sigma^{-1} \boldsymbol{\mu}$ (Asti *et al.*, 2016). When Σ possess full rank, the maximum likelihood of the model representing the data is achieved at $\Sigma = \mathcal{C}$ and $\boldsymbol{\mu} = \bar{\mathbf{x}}$. Nevertheless, due to sampling and subsequent degenerateness of the empirical covariance matrix, it is necessary to include a prior distribution. Here, we define \mathcal{U} and $\boldsymbol{\eta}$ as covariance and mean of a uniformly distributed sample which are included to \mathcal{C} and $\bar{\mathbf{x}}$. Hence, we set $\boldsymbol{\eta} = \boldsymbol{\eta} = 1/Q+1$ as constant. Since \mathcal{U} consists of $L \times L$ number of matrices with the dimensions $Q \times Q$, $\mathcal{U}_{ii} = \mathbf{U}$ with the diagonal elements of \mathbf{U} being $U_{kk} = Q/(Q+1)^2$. The off-diagonal elements of \mathbf{U} are $U_{kl, k \neq l} = -1/(Q+1)^2$. Hence, the modified empirical covariance matrix (\mathcal{C}^*) can be expressed as in Eqn. 3.16 and used as described above.

$$\Sigma = \mathcal{C}^* = \lambda \mathcal{U} + (1 - \lambda) \mathcal{C} + \lambda(1 - \lambda) (\bar{\mathbf{x}} - \boldsymbol{\eta})^T (\bar{\mathbf{x}} - \boldsymbol{\eta}) \quad (3.16)$$

λ represents the influence of the prior distribution on \mathcal{C}^* and influences $\bar{\mathbf{x}}$ similarly as $\boldsymbol{\mu} = \bar{\mathbf{x}}^* = \lambda \boldsymbol{\eta} + (1 - \lambda) \bar{\mathbf{x}}$. Similar to DCA, inversion of Σ yields the coupling matrix \mathbf{J} from which couplings can be derived for each \mathbf{J}_{ij} submatrix through Direct Information (Morcos *et al.*, 2011) or Frobenius Norm ($F_{ij} = \sqrt{\sum_{kl} \mathbf{J}_{ij}(\alpha_k, \alpha_l)^2}$). Due to this, aforementioned DCA and the GaussDCA differ in the integration of prior distributions in the disentangling of direct couplings from background noise.

In the following sections we will use the GaussDCA to detect direct couplings of positions in MSA as earlier results (Baldassi *et al.*, 2014) attested this method efficient computation since most expressions can be derived analytically. Additionally, we refer to the GaussDCA as DCA for simplicity's sake.

3.1.3 Statistical Complexity

While classic information theoretical measures, such as Shannon's entropy $H(X)$ and mutual information $I(X, Y)$, capture randomness and unpredictability of a system, these methodologies fail to adequately comprise complex relationships within these systems (Feldman and Crutchfield, 1998). In general, statistical complexity deals with intricate structures within dynamics emanating from a much simpler system (Kantz *et al.*, 1998). Hence, maximally random (e.g. white noise) and completely ordered systems (e.g. periodic motion) inherit no structure and possess no complexity. Oppositionally, mutational patterns in viruses or electroencephalograms of brain activity can be considered systems in *off-equilibrium* with intricate mechanics.

To quantify the statistical complexity (not to be confused with the Big-O notation or computational complexity in general) a number of approaches have been taken. Generally, the statistical complexity is an amalgamation of an entropy \mathcal{H} and a distance measure \mathcal{D} so that the complexity C of a random variable X is expressed as $C(X) = \mathcal{H}(X, \dots) \cdot \mathcal{D}(X, \dots)$. While not all entropic and distance measures need additional arguments represented by the term \dots , this generic form accounts for possible entropies and distances.

Recently, we showed that statistical complexity calculated from Shannon's entropy $H(X)$ (see Eqn. 3.1) and the Jensen-Shannon divergence (see Eqn. 3.4) can give insights in mutation patterns within sequences of the HIV1 protease (Keul and Hamacher, 2017). Here, we will now extend the difference in statistical complexity (ΔC) to incorporate co-evolutionary influences in form of a normalized mutual information (see Eqn. 3.5).

Complexity of Co-evolution

Mutual information (MI) shares similar characteristics to Shannon's entropy as zero MI ($I(X, Y) = 0$) is the result of no relation between the distributions X and Y , whereas the maximum MI is only dependent on the maximum entropy of either distribution ($I(X, Y) = \log_2(N_{\min})$ with $N_{\min} = \min(N_X, N_Y)$). The mutual information of two distributions X and Y is maximal when the joint entropy for both is minimal, i.e. exhibits minimal entropy. Nevertheless, a completely ordered two point distribution can be considered, in fact, as non-complex, as can complete randomness. Hence, neither case can reveal significant insights on intrinsic complexity of co-evolutionary processes. In order to assess these complexities we define the co-evolutionary complexity as:

$$C_{\text{MI}}(X, Y) = I(X, Y) \cdot \sqrt{D_{\text{JS}}(X, Y | \hat{X}, \hat{Y})} \quad (3.17)$$

Here, we define $D_{\text{JS}}(D_1 | D_2) = D_{\text{JS}}(D_1, D_2)$ to avoid confusing notations. $C_{\text{MI}}(X, Y)$ measures the co-evolutionary complexity of the distribution pair X and Y in relation to the reference distribution pair \hat{X} and \hat{Y} on basis of the mutual information. Similar to Keul and Hamacher (2017), we chose to

assess the difference in complexity $C_{\text{MI}}(X, Y)$ in reference to the co-evolutionary complexity of the reference system $C_{\text{MI}}(\hat{X}, \hat{Y})$. Thus, we obtain the co-evolutionary complexity difference ΔC_{MI} :

$$\begin{aligned}\Delta C_{\text{MI}}(X, Y) &= I(X, Y) \cdot \sqrt{D_{\text{JS}}(X, Y | \hat{X}, \hat{Y})} - I(\hat{X}, \hat{Y}) \cdot \sqrt{D_{\text{JS}}(\hat{X}, \hat{Y} | X, Y)} \\ &= [I(X, Y) - I(\hat{X}, \hat{Y})] \sqrt{D_{\text{JS}}(X, Y | \hat{X}, \hat{Y})}\end{aligned}\quad (3.18)$$

From Eqn. 3.18 we can see that $\Delta C_{\text{MI}}(X, Y)$ measures the difference in co-evolutionary complexity upon deviation from the reference distributions. Therefore, it is necessary to obtain appropriate reference distributions as the (often used) uniform distribution for $P(\hat{X})$, $P(\hat{Y})$ and $P(\hat{X}, \hat{Y})$ results in $MI(\hat{X}, \hat{Y}) = 0$. While using $C_{\text{MI}}(X, Y)$ straightforward could indicate deviations between the distributions X and Y and their references, information on which distributions inherit stronger co-evolutionary complexity would be lost.

To properly compare MI from two different distributions (such as $MI(X, Y)$ and $MI(\hat{X}, \hat{Y})$) it is imperative to normalize both measures in relation to their respective number of filled bins (i.e. the used alphabet). Here, we normalize MI with its upper bound, the entropy of a uniform distributed $N_X \times N_Y$ contingency table. Therefore we obtain for the co-evolutionary complexity difference ΔC_{MI}

$$\begin{aligned}\Delta C_{\text{MI}}(X, Y) &= \Delta I^{(n)} \sqrt{D_{\text{JS}}(X, Y | \hat{X}, \hat{Y})} \\ \text{with } \Delta I^{(n)} &= \frac{I(X, Y)}{\log_2(N_X \cdot N_Y)} - \frac{I(\hat{X}, \hat{Y})}{\log_2(N_{\hat{X}} \cdot N_{\hat{Y}})}\end{aligned}\quad (3.19)$$

Of all here listed methods, normalizations and corrections, only DCA and ΔC_{MI} use prior distributions to correct or relate observed frequency counts. Hence, these methods include correction terms to account for incompleteness and finite size effects of the data. Additional corrections for finite sample size can be found in for example Grassberger (1988) or Dudík *et al.* (2005).

3.2 Methods

3.2.1 Datasets

Throughout this chapter we will use various datasets to conduct performance analysis of information theoretical measures and to gain insight into evolutionary characteristics of specific protein families. Here, we will explain how each dataset was obtained.

HIV1 protease

We obtained the sequences of the HIV1 protease (HIVP) from the HIV drug resistance database (Rhee *et al.*, 2003). Since functional HIVP sequences possess a length of 99 amino acids (Shafer *et al.*, 2001), we omitted all sequences containing more or less than this length. Furthermore, we removed sequences with gap characters, ambiguous amino acids and non-canonical amino acid symbols. Hence, our full HIVP dataset consists of 34,747 sequences of which 6,447 sequences were from patients treated with at least one HIV protease inhibitor. Our dataset of sequences from patients without administration

of HIV protease inhibitors was 28,300 sequences strong. In each dataset, patient data spanned from 1992 to 2010, at least, and consisted of sequences from all continents. Due to the large size of the entire dataset and the global distribution of HIVP sequences, we can assume that this dataset covers a high percentage of the HIVP quasi-species sequence space.

Ion conducting proteins

In order to obtain insights into the co-evolution within potassium channels, we used alignments of protein families obtained from Pfam database (Finn *et al.*, 2008, 2016). Here, we chose the full alignments for PF00520 (Pfam: ion transporter proteins) and PF07885 (Pfam: ion channels) to investigate the co-evolutionary differences between these families.

To generate alignments with corresponding position pairs necessary for the direct comparison with f.e. ΔC_{MI} measure, we performed a three step method to reduce the alignments:

- (i) We selected one sequence of each alignment with a known crystal structure.
- (ii) We performed a structural alignment with the `TM-align` plug-in (Zhang and Skolnick, 2005) in `pymol` (Schrödinger, LLC, 2015) of the two monomeric structures.
- (iii) The resulting structural alignment was then used to reduce the full alignments of either protein family. Here, positions with a gap in either sequence of the structural alignment were omitted from the full alignment of both families.

Based on this work flow, we are able to compare the two protein families with the structures of the potassium channel from the chlorella virus PBCV-1 (Kcv) – for PF07885 – and the human voltage-gated potassium channel Kv1.2 – for PF00520 – (see. Tab. 4.1 for reference). As we are mainly interested in potassium channel sequences, we screened all sequences of both families for the minimalistic filter motives *GFG* and *GYG*. Sequences, containing neither motive were omitted. Visual inspection of the alignments showed two absolutely conserved columns ($p(G) = 1$) surrounding a column with mixed phenylalanine and tyrosine for all remaining sequences of both alignments. This allows us to compare evolutionary effects within the pore region of the protein families PF00520 and PF07885. The resulting alignments for PF00520 and PF07885 contained 5,344 and 6,661 sequences, respectively. Noteworthy is here, that the first 13 residues of the Kcv sequence in the PF07885 alignment are missing when compared to its crystal structure.

Furthermore, we extended our analysis of co-evolutionary differences and structural relations for sequences originating from eukaryotic and prokaryotic organisms. Here, we focused solely on potassium channel sequences of PF07885, as the number of prokaryotic sequences in PF00520 was almost two orders of magnitude smaller than its eukaryotic counterpart, with a high possibility of distorting results due to size effects within the data. From the PF07885 potassium channel sequences, 4,813 were annotated as eukaryotic and 864 sequence as prokaryotic. In order to appropriately visualize the results, we reduced the alignments of both domains. Here, we used the sequence of KcsA (Doyle *et al.*, 1998) and omitted all columns with a gap symbol in the KcsA sequence of PF07885. Noteworthy is here, that the first ten and the last three residues of the KcsA sequence in the crystal structure (PDB: 1BL8, refined by S. Tayefeh) are missing from the sequence in PF07885.

3.2.2 Comparison of information theoretical measures

For the comparison of matrices and vectors we use here Spearman's rank correlation coefficient $\rho_{\mathbf{x},\mathbf{y}} = \text{cov}(\mathbf{r}_{\mathbf{x}}, \mathbf{r}_{\mathbf{y}}) \sigma_{\mathbf{r}_{\mathbf{x}}}^{-1} \sigma_{\mathbf{r}_{\mathbf{y}}}^{-1}$. With this measure we are able to compare distributions with non-linear relations. Here, $\mathbf{r}_{\mathbf{x}}$ denotes the vector of rank based on the raw data vector \mathbf{x} when ordered. $\text{cov}(\mathbf{r}_{\mathbf{x}}, \mathbf{r}_{\mathbf{y}})$ is the covariance between the rank vectors for \mathbf{x} and \mathbf{y} . $\sigma_{\mathbf{r}_{\mathbf{x}}}$ represents the square root of the variance of $\mathbf{r}_{\mathbf{x}}$.

Variance of a matrix

In order to compare two-dimensional measures, such as the MI, DCA or the ΔC_{MI} , to one-dimensional ones (like Shannon's entropy or ΔC), we use the single value decomposition (SVD) to obtain the principal components of any matrix \mathbf{M} . The SVD of \mathbf{M} can then be written as $\mathbf{M} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T = \sum_i^{N-d} \mathbf{u}_i \lambda_i \mathbf{v}_i^T$ (Wall *et al.*). Here, \mathbf{u}_i and \mathbf{v}_i represent the i th column of the left and right singular vectors matrices \mathbf{U} and \mathbf{V} , respectively. λ_i is Λ_{ii} of the diagonal singular value matrix. The scalar d represents the multiplicity of zero singular values. Based on this, the contributing percentage of the i th singular vector/value combination to \mathbf{M} can be assessed through

$$V_i^2 = \frac{\|\mathbf{u}_i \lambda_i \mathbf{v}_i^T\|^2}{\|\mathbf{M}\|^2} = \frac{\lambda_i^2}{\sum_{j=1}^{N-d} \lambda_j^2} \quad (3.20)$$

In the computation of V_i^2 , matrix norms (denoted here as $\|\mathbf{X}\|^2$) are Frobenius norms. Hence, we are comparing a partially reconstructed matrix \mathbf{M}_i (i.e. $\mathbf{u}_i \lambda_i \mathbf{v}_i^T$) with the original matrix \mathbf{M} . Through this, we obtain the percentage of variance in \mathbf{M} that is explained through the i th singular vector.

Computation of measures

We used the `BioPhysConnector` package (Hoffgaard *et al.*, 2010) in R (R Core Team, 2015) to conduct the computations for all MI based methods. Implementation of RCW and APC followed the instructions from (Gouveia-Oliveira and Pedersen, 2007) and (Dunn *et al.*, 2007). Implementation of Z_{RCW} and Z_{APC} was derived from the `BioPhysConnector` implementation of Z_{MI} (Weil *et al.*, 2009). DCA computations for $\theta = 0$ and automatically computed θ were performed in `julia` using the `GaussDCA` package (Baldassi *et al.*, 2014).

Robustness of performance

The assessment of co-evolutionary measures has been performed on artificial datasets (Dunn *et al.*, 2007; Gouveia-Oliveira and Pedersen, 2007) and has just recently been investigated on state-of-the-art databases (Mao *et al.*, 2015). Nevertheless, these conventional approaches to asserting performance of co-evolutionary measures assume data completeness while "real life" applications might not warrant such assumptions. Therefore, we estimated the robustness of the aforementioned methods through using the HIV protease dataset described earlier. This dataset enables us to analyze the robustness

information generated by information theoretic measures uncoupled from bias introduced through alignments and phylogeny.

In order to conduct a thorough assessment for all methods, we derived the following workflow. First we computed all k information theoretic measures ($I_k^{(100)}$) on basis of the dataset $D^{(100)}$ with all $n_{D^{(100)}}$ sequences. Then we randomly selected $p\%$ of the sequences from the HIV protease dataset, with p in the interval $[0.001, 0.999]$ (with the number of selected sequences being $n_{D^p} = \lceil p \cdot n_{D^{(100)}} \rceil$). Based on this new dataset D^p we computed all measures ($I_k^{(p)}$) and compared each measure to $I_k^{(100)}$ through Spearman's rank correlation coefficient ρ . By repeating this procedure for $p \in \{0.001, \dots, 0.01\}$ 1000 times and for $p \in \{0.011, \dots, 0.999\}$ 100 times, we obtained the datasets D_i^p with i being the i th repetition for percentage of resampled sequences p . From these D_i^p we were able to estimate the mean correlations $\bar{\rho}$ with its variance σ_ρ^2 for each method for each p .

Similar to the common practice of discussing the relevance and explanatory power of the highest scoring results, we chose to investigate correlation of the highest 1% and 10% entries as well. Here, we computed an overlap percentage ω through relating the cardinality of overlapping entries in the highest 1% and 10% entries of $I_k^{(100)}$ ($I_{k,1\%}^{(100)}$ and $I_{k,10\%}^{(100)}$) to the respective 1% and 10% entries of $I_k^{(p)}$ ($I_{k,1\%}^{(p)}$ and $I_{k,10\%}^{(p)}$) as in Eqn. 3.21.

$$\omega(M, M^*) = \frac{|M \cap M^*|}{|M|} \quad (3.21)$$

We denote here, the set M which is either the highest ranked 1% or 10% entries for any method I_k . Consequently, M^* is the set for recomputed results of $I_k^{(p)}$ for a given percentage p . Similarly to Spearman's ρ we can now estimate mean overlap $\bar{\omega}$ as well as its variance σ_ω^2 for any percentage p and method I_k .

For ΔC_{MI} method, we separate each sequence subset in sequences from HIV protease inhibitor treated and untreated patients. Based on these two datasets, we compute ΔC_{MI} with the sequences from untreated patients representing the reference distribution. We take the absolute of any obtained values to include the maximum ΔC_{MI} values and account for the directionality of co-evolutionary complexity.

3.3 Results

The results are split into two parts. In the first part, we will investigate general properties of ΔC_{MI} in comparison to other information theoretical measures as well as the robustness of their results upon sampling of data. In the second part, we will derive novel insights into the evolution, fold and evolutionary dynamics of ion conducting channel proteins.

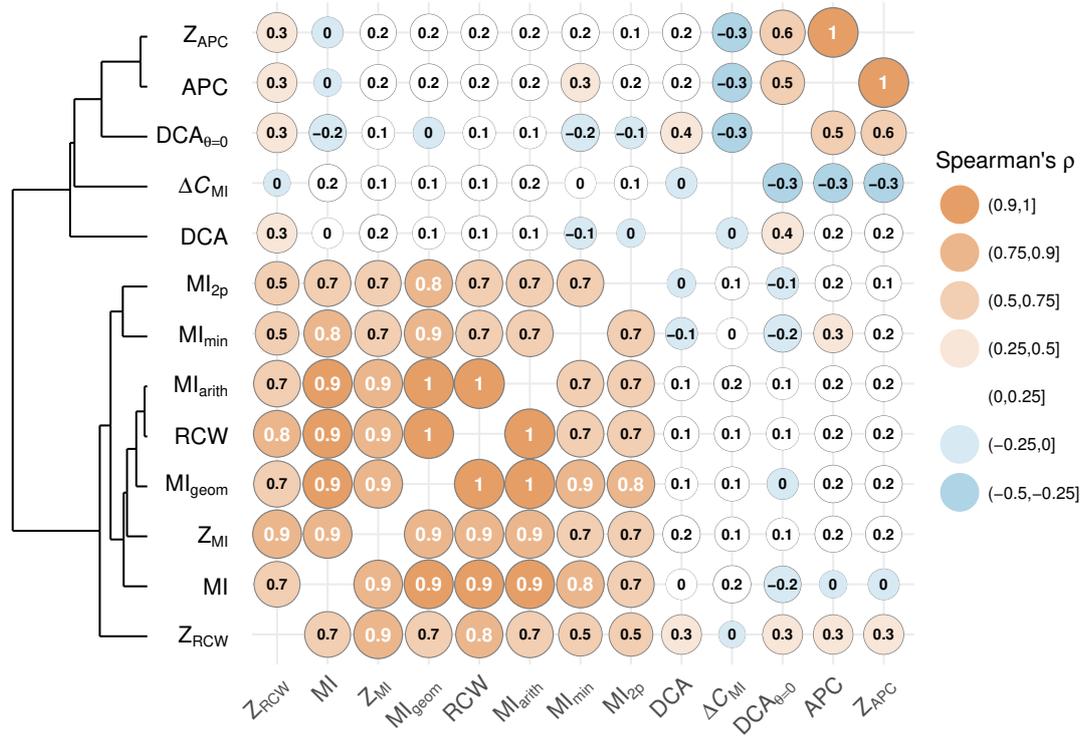


Figure 3.2.: Spearman’s rank correlation matrix of the here tested information theoretic measures on the basis of all HIVP sequences labeled as treated. Noteworthy is that the MI-based APC exhibits moderate correlation to the DCA (with $\theta = 0$) and displays high correlation to Z_{APC} . ΔC_{MI} (using the distributions obtained from the untreated HIVP sequences as reference) shows mostly moderate correlation to MI-based measures.

3.3.1 Properties of Co-evolutionary Complexity

ΔC_{MI} differs to traditional measures

While ΔC_{MI} presents a completely different theoretical approach to extract information on evolution from sequence data, it is necessary to investigate to which extend other, established measures capture similar events. Hence, we analyzed whether traditional measures describe some of the dynamics captured in ΔC_{MI} by comparing these traditional approaches to the novel co-evolutionary complexity on basis of the HIV protease dataset. Here, we only considered patients treated with protease inhibitors and performed a correlation analysis of aforementioned information theoretical measures and ΔC_{MI} . The results of the Spearman’s rank correlation coefficient are shown in Fig. 3.2. Furthermore, we used the resulting correlation matrix to cluster the methods. We can obtain a distance matrix through measuring the euclidean distance between the correlation vectors (i.e $d_{ij} = \sqrt{(\rho_i - \rho_j)^2}$ for the distance between method i and j). Based on this distance, we used the farthest neighbor clustering methods (Lance and Williams, 1967).

Based on the correlation matrix and the dendrogram derived therefrom, we can clearly identify 2 major cluster of methods (see Fig. 3.2). The first main cluster is populated with both DCA methods, APC and Z_{APC} and ΔC_{MI} . As the maximum entropy approach of DCA is conceptually different from traditional information theoretical measures, both DCA variants show only anecdotal correlation to all other MI-based methods. Surprisingly, the average product correction method proposed by Dunn *et al.* (2007) and its Z-score variant appear in the same cluster as both DCA measures, with APC and Z_{APC} showing good correlation to $DCA_{\theta=0}$ – and as such forming one sub-cluster. The here tested DCA variant which computes θ automatically and as such reduces the effective number of sequences in the dataset, shows no correlation to any other method. Similarly, ΔC_{MI} has no correlation to any of the here listed information theoretical measures, emphasized through the conceptually different approach to retrieving information from sequence datasets. These findings are unsurprising as ΔC_{MI} uses information of a background or reference distribution to relate the co-evolutionary complexity of the dataset to this reference complexity.

The other cluster of methods is formed by MI, MI_{arith} , MI_{geom} , RCW, MI_{2p} and MI_{min} , Z_{MI} and Z_{MI} . All of these methods show a minimum correlation of $\rho \geq 0.5$. Nevertheless, within this cluster of methods, Z_{RCW} differs the most from all other methods and shows to a lesser extend correlation to methods from the first cluster. Nevertheless, results similar to Z_{RCW} can be obtained with both the RCW and Z_{MI} method as all three methods possess at least 75% of similar interactions in their highest 10% matrix entries (see Fig. C.1).

Furthermore, we can observe additional similarities between closely related methods, such as MI_{arith} , RCW and MI_{geom} . High correlations between MI_{arith} and MI_{geom} are unsurprising as both methods represent normalization in respect of different types of averaging. Comparably, the row-column-weighting of mutual information presents a similar approach by normalizing the MI in respect to its row- and column-wise arithmetic means. Our results indicate that these three methods can be used interchangeably, at least on the HIVP dataset of patients treated with protease inhibitors.

Two other methods frequently used in image processing and machine learning show also strong correlation to one another. The MI-based methods MI_{2p} and MI_{min} relate the mutual information to the joint-entropy or the minimum column entropy. These two measures show moderate to high correlation to other MI based measures (apart from APC). Conceptually, both methods normalize the MI to theoretical upper boundaries, and unsurprisingly appear in one cluster.

Altogether, co-evolutionary dynamics captured by ΔC_{MI} are unique to this method. While almost all MI based approaches experience strong correlation to one another, ΔC_{MI} and DCA based methods present vastly different results.

Co-evolutionary Complexity partially derives from single point complexity

We presented in Section 3.1.3 the foundation of the two-dimensional co-evolutionary complexity on basis of the statistical complexity by (Keul and Hamacher, 2017). Here, we show the results of the co-evolutionary complexity difference ΔC_{MI} on the HIV1 protease dataset. In the context of ΔC_{MI} we compare the distribution of amino acids from sequences of HIV protease inhibitor treated patients with the sequence dataset from untreated patients.

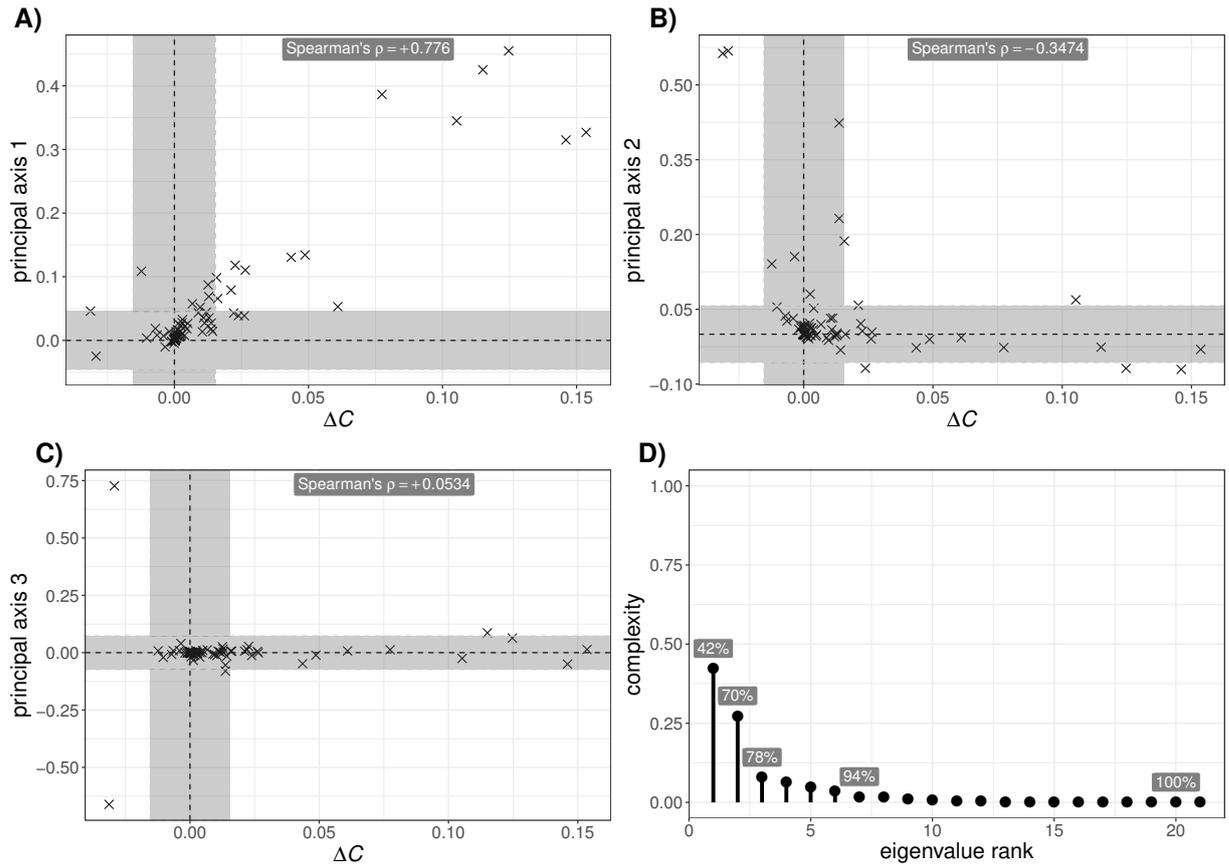


Figure 3.3.: Comparison of the first three principal axes of the co-evolutionary complexity ΔC_{MI} to the differential statistic complexity (ΔC) (A, B, and C). In A, B and C ΔC is plotted along the abscissa while the principal axes are plotted along the ordinate with gray areas highlighting vector entries below $0.1 \max(\Delta C)$ or for principal axis below $0.1 \max(\mathbf{v})$. The percentage of co-evolutionary complexity captured within each of the first principal axes is shown in D with their cumulative percentage highlighted in gray.

In order to investigate whether co-evolutionary complexity discloses intricate evolutionary processes different from the statistical complexity of Keul and Hamacher (2017), we used a dimensionality reduction approach similar to principle component analysis. Through singular value decomposition of ΔC_{MI} we can derive its principal axes, with principal axes (or eigenvectors) with large eigenvalue accounting for higher percentages of co-evolutionary complexity. Here, we focus on the predominant principal axes of ΔC_{MI} , compare these to ΔC and calculate their Spearman rank correlation coefficients between. We find that the first three principal axes of ΔC_{MI} account for $\sim 78\%$ of the complexity found in ΔC_{MI} (see Fig. 3.3D) and thus represent the majority of co-evolutionary complexity found. Fig. 3.3A-C show the comparison of these first three principal axis with the one-dimensional ΔC . While the first – and dominant – principal axis shows high correlation to ΔC (Fig. 3.3A) with Spearman's $\rho = + 0.776$, the second and third principal components show only weak or no correlation at all (Fig. 3.3B+C). From this, it is clear, that information on co-evolutionary complexity captured through ΔC_{MI} can only be partially reconstructed from ΔC . Local information from ΔC_{MI} can deviate from single-point statistical complexities.

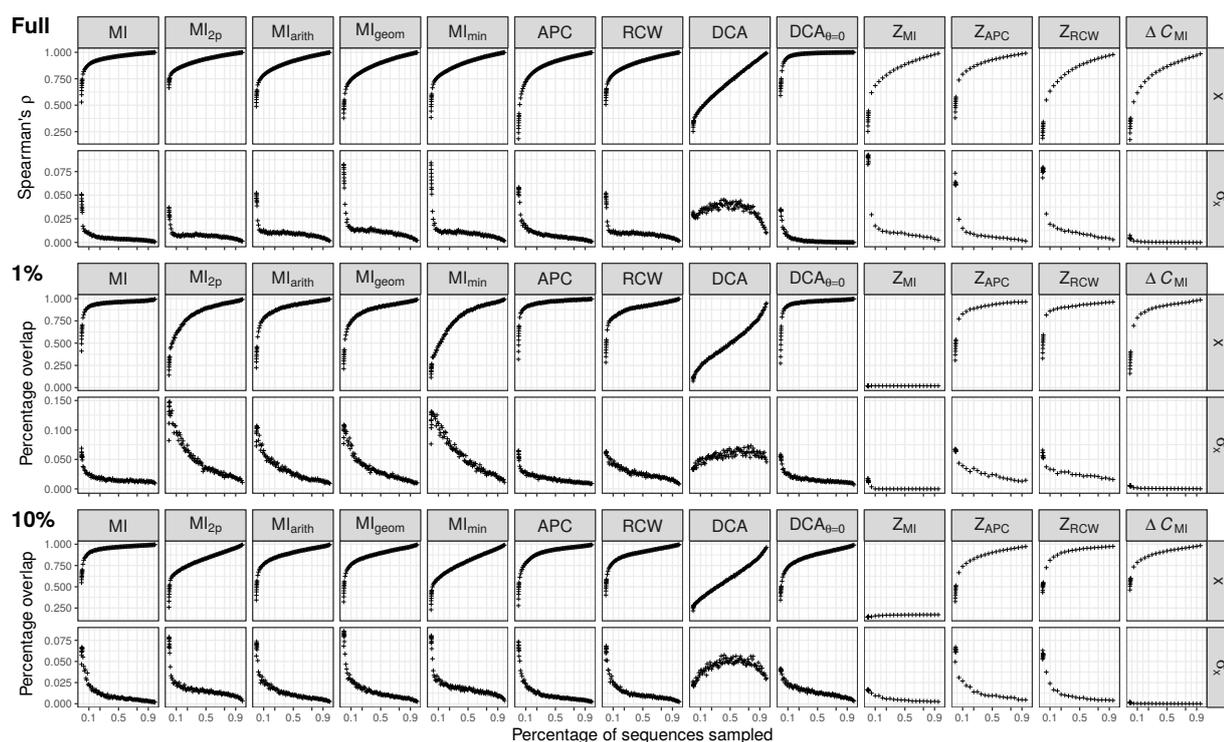


Figure 3.4.: Progression of the similarity between results generated from a subset of sequences and results for all measures based on the all HIVP sequences. The top panel shows the average Spearman rank correlation ($\bar{\rho}$) and its standard deviation (σ_{ρ}) on the full result matrices (labeled as "**Full**"). The second and third panel shows the similarity (and their standard deviation) between the top **1%** and top **10%** results based on the subset when compared to original, full dataset. On the abscissa of each plot the percentage of the sampled sequences is shown, with the resulting measure (Spearman's ρ or overlap in percent) shown on the ordinate. Here, we can clearly see that methods such as the DCA or MI_{2p} are heavily influenced by the sequence space composition of the dataset and appear less robust than, for example, RCW. Furthermore, MI-based Z-scores fail to reproduce top ranked contacts consistently, symbolized by similarities close to 0% on the top 1% and top 10% results. ΔC_{MI} shows consistent results for the top 1% and 10% even at low sequence numbers.

Consistency of information theoretical measures

Even though the advantages and results of information theoretical measures have been well documented (Gloor *et al.*, 2005; Gouveia-Oliveira and Pedersen, 2007; Dunn *et al.*, 2007; Weil *et al.*, 2009; Boba *et al.*, 2010), only few studies relate the performance of these measures to varying dataset sizes and compositions. To investigate the minimum number of sequence necessary to obtain consistent and reliable results we evaluated the performance of common information theoretical measures on sampled subsets of HIV protease sequences.

While it has been suggest elsewhere (Grassberger, 1988; Fernandes and Gloor, 2010) that measures such as the MI are sensible to correct (or wrong) prior assumptions, we extended our general assessment to a broader spectrum of information theoretical measures. On the basis of a HIV protease dataset, we

performed Jackknife resampling of the sequences by drawing between 1% and up to 99% of available sequences with up to 1000 repetitions. By treating results from the original dataset as *ground truth*, we are able to investigate the sensitivity of all above mentioned methods upon sampling and changes in the dataset. Ideally, all measures analyzed here, should present high correlation coefficients with low variances at very small (and realistic) sequence numbers.

Fig. 3.4 shows the results from our consistency analysis with each data point representing a set of at least 100 alignments on which all methods were computed. Furthermore, we subdivided the results into three categories, which are normally used when assessing the results of information theoretical measures. Similarities found in the 1% row of Fig. 3.4 are based on the ranks of the highest "1%" of values (i.e. the top 49 entries of any matrix). Similarly, the rows labeled with "10%" represent the similarity of the highest 10% values (i.e. 485 matrix entries). Panel rows labeled with "full" take the entire matrix into account when calculating ρ for a given method and Jackknife resampling percentage.

Unsurprisingly, correlations and similarity between results obtained from subsets and the original dataset increase with increasing number of sequences in the subset for all methods. While methods such as MI, RCW and APC display strong average similarities with even less than 10% of all available sequences used, MI_{2P} and DCA appear much more susceptible to changes in dataset composition. Furthermore, the performance of methods such as MI_{2P} , Z_{MI} , DCA and MI_{min} are reduced when only considering their top results. Contrarily, MI, APC, RCW and ΔC_{MI} increase their performance when taking only the top results into account. Z_{APC} and Z_{RCW} show increased precision when only considering the highest 1% values.

Interestingly, DCA performs consistently worse than all other here tested methods and shows even at high sequence sample percentages a large standard deviation in its results. Even though the standard deviation of Spearman's ρ is reduced when taking more matrix entries into account (i.e. from the highest 1% to 100% of the matrix entries), we still observe large variance in the correlation coefficients. Thus, DCA computations show the greatest sensibility on changes in the analyzed data set. One has to note, though, that the maximum effective alignment size was $M_{eff} \leq 53$ for DCA results obtained with automatically determined θ . Hence, this effective sequence number should be considered below the defined threshold of 1000 sequences by Morcos *et al.* (2011). Without the induced clustering through the automatic generated similarity threshold (i.e. $\theta = 0$) the DCA performance increases with higher average correlation to the DCA of the full dataset with reduced variance of correlation. Therefore, we can assume that due to relative high similarity of the HIVP sequences, the for sensitive DCA results necessary effective sequence count cannot be achieved.

Similar to other measures, robustness ΔC_{MI} results increases with the number of sequence included in the alignment. Within the tenth of a percent range results obtained from ΔC_{MI} show only marginal correlation to the *ground truth*. Nevertheless, the top 1% of the *ground truth* are consistently and correctly identified at low sequence numbers. While the overall correlation of all results increases slowly with additional sequence in the subset, high overlap in the top 1% and 10% can be observed.

3.3.2 Insights into ion channel evolution

We extended the computation of ΔC between sequences data set from the comparison of treated and untreated patients of Keul and Hamacher (2017) to the analysis of co-evolutionary complexity between

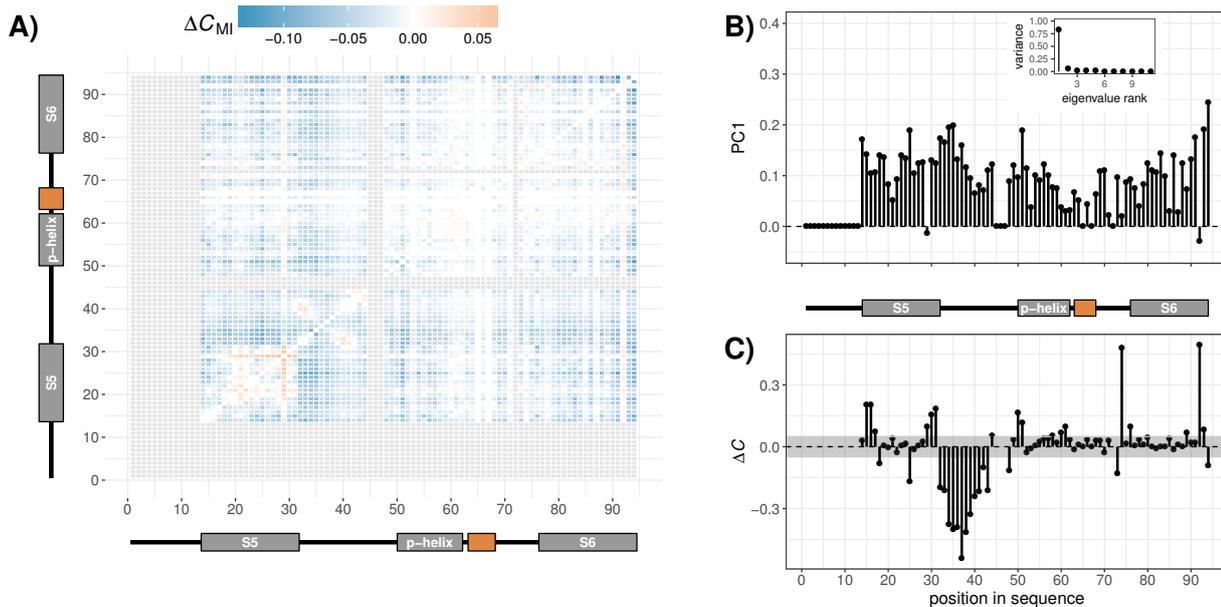


Figure 3.5.: Statistic complexity between the potassium ion channels in the families PF07885 and PF00520. Corresponding positions were obtained through structural alignment of Kv1.2 (member of PF00520) and Kcv (member of PF07885). In **A**), ΔC_{MI} is shown in respect to the Kcv sequence with gapped positions of the alignment yield an *NA* in the ΔC_{MI} computation. Co-evolutionary complexity of the ion channel family (PF07885) is increased in the outer transmembrane helix of the pore forming structure. Here, especially the N-terminal segment of S5 experiences increased complexity of co-evolution. **B**) shows the first principal component of ΔC_{MI} , comprising the highest ΔC_{MI} variance (see inlay) while **C**) shows ΔC between the potassium channels of families Values in gray (in **A**) and missing values in **C**) are attributed to gaps in Pfam sequences of Kcv and gaps in the pairwise alignment of Kcv and Kv1.2 to obtain corresponding positions.

these in an earlier subsection. In the following part of this work, we will focus on intricate mechanisms of ion channel evolution. Here, we will use the here introduced ΔC_{MI} to gain an understanding of the co-evolutionary complexity differences in respect to ion channels and in particular potassium channels. As foundation of all upcoming analyses we used the Pfam alignments of the Families PF00520 (Ion Transport family) and PF07885 (Ion channel family). Sequences from these families can be assigned features such as *potassium conducting*, *prokaryotic* or *eukaryotic* channel.

ΔC_{MI} reveals reduced evolutionary complexity in S5 of large ion channels

Through analyzing sequences containing the minimalistic potassium channel filter sequence *GY/FG* we are able to compare evolutionary and co-evolutionary differences between sequences from channels with six transmembrane domains (six TMD) and smaller two transmembrane domain (two TMD) channels on basis of the Pfam families PF00520 and PF07885. In this context, we use ΔC and ΔC_{MI} to investigate both single site and co-evolutionary complexity difference between these two families. The results of this comparison can be found in Fig. 3.5 and are separated into results for ΔC_{MI} (Fig. 3.5A), the first singular vector of ΔC_{MI} (Fig. 3.5B) and ΔC between the alignments of small and large potassium channels in Fig. 3.5C. Hereby, we chose to relate ΔC and ΔC_{MI} to sequences of PF07885 with all plots reduced to positions corresponding to residues found in the Kcv structure. Positive ΔC and ΔC_{MI} values indicate higher statistical complexity in the small ion channel dataset while negative

values represent higher complexity values in the PF00520 protein family. Additionally, positions without a corresponding amino acid in the pairwise alignment of Kcv and Kv1.2 were omitted in Fig. 3.5.

In Fig. 3.5A) we can observe drastic differences in co-evolutionary complexity within the outer helix (S5 helix) of the central pore regions of two TMD and six TMD ion channels. Here, we see a $\Delta C_{MI} > 0$, describing higher co-evolutionary complexity in two TMD potassium channels. Interestingly, we find the highest positive ΔC_{MI} value at position 29 of the Kcv sequence indicating a more intricate evolutionary process for this position. In Kcv, we find a positively charged lysine at position 29, exposed to the membrane environment. While this position represents a surface residue in Kcv, the distal region of the S5 helix presents an interaction interface between the voltage sensing domain (VSD) and the S5 helix in six TMD channels. We can therefore safely assume that large complexity difference can be attributed to increased conservation within S5 of six TMD channels. This finding is emphasized through high ΔC values in this region (see Fig. 3.5C)). Furthermore, we find high co-evolutionary complexity difference between the central part of the first turret loop (positions 39 to 41) with the proximal positions 32 and 33 close to the end of the S5 helix. Here, smaller two TMD channels experience more complex evolutionary relationships than their larger counterparts.

For $\Delta C_{MI} < 0$ we find the largest difference near the N-terminus of the Kcv sequence. A multitude of position pair in this region shares a more intricate evolutionary behavior in larger channels. Here, the most noteworthy co-evolutionary complexity change are large ΔC_{MI} values shared between the S5 helix and these distal S6 residues. Additionally, we find at positions 80, 89, 92, 93 and 94 a ΔC greater than the threshold of $0.1 \cdot \max(|\Delta C|)$ proposed by Keul and Hamacher (2017). The single site, statistical complexity in combination with the co-evolutionary complexity ΔC_{MI} reveals large differences in the complexity of the evolutionary mechanism within the S5 helix and at the N-terminus of the S6 helix. Additionally, our results show, that the majority of co-evolutionary complexity is contained within the first singular vector (see Fig. 3.5B)) and appears to show a partially different pattern than ΔC .

Prokaryotic channels show increased evolutionary complexity within the g-helix

While large ion channels can possess hundreds of amino acids within additional transmembrane and cytoplasmic domains, their essential core function is performed by a small number of residues. The structural and functional resemblance of this core unit throughout all domains of life – a fact we will exploit in Chapter 4 – allows us to investigate mechanism of evolutionary complexity for many different organisms. Focusing on the PF07885 alignment, which can be subdivided in sequence originating from eukaryotic and prokaryotic organisms, we aim at analyzing the intricate evolutionary mechanisms through mutual information, direct coupling analysis and co-evolutionary complexity. Here, we use sequence from Pfam full alignments and condense, for visualization purposes, the results on residues present in KcsA.

The comparison between prokaryotic and eukaryotic sequences through ΔC_{MI} is shown in Fig. 3.6A). Here, positive ΔC_{MI} indicate a more intricate co-evolutionary complexity in prokaryotic channels, while negative ΔC_{MI} values highlight more complex evolutionary mechanisms in eukaryotic channel sequences. Our results show that generally the co-evolutionary complexity in eukaryotic channels is higher than in prokaryotic for a majority of residue pairings in the S5 helix. Interestingly, at

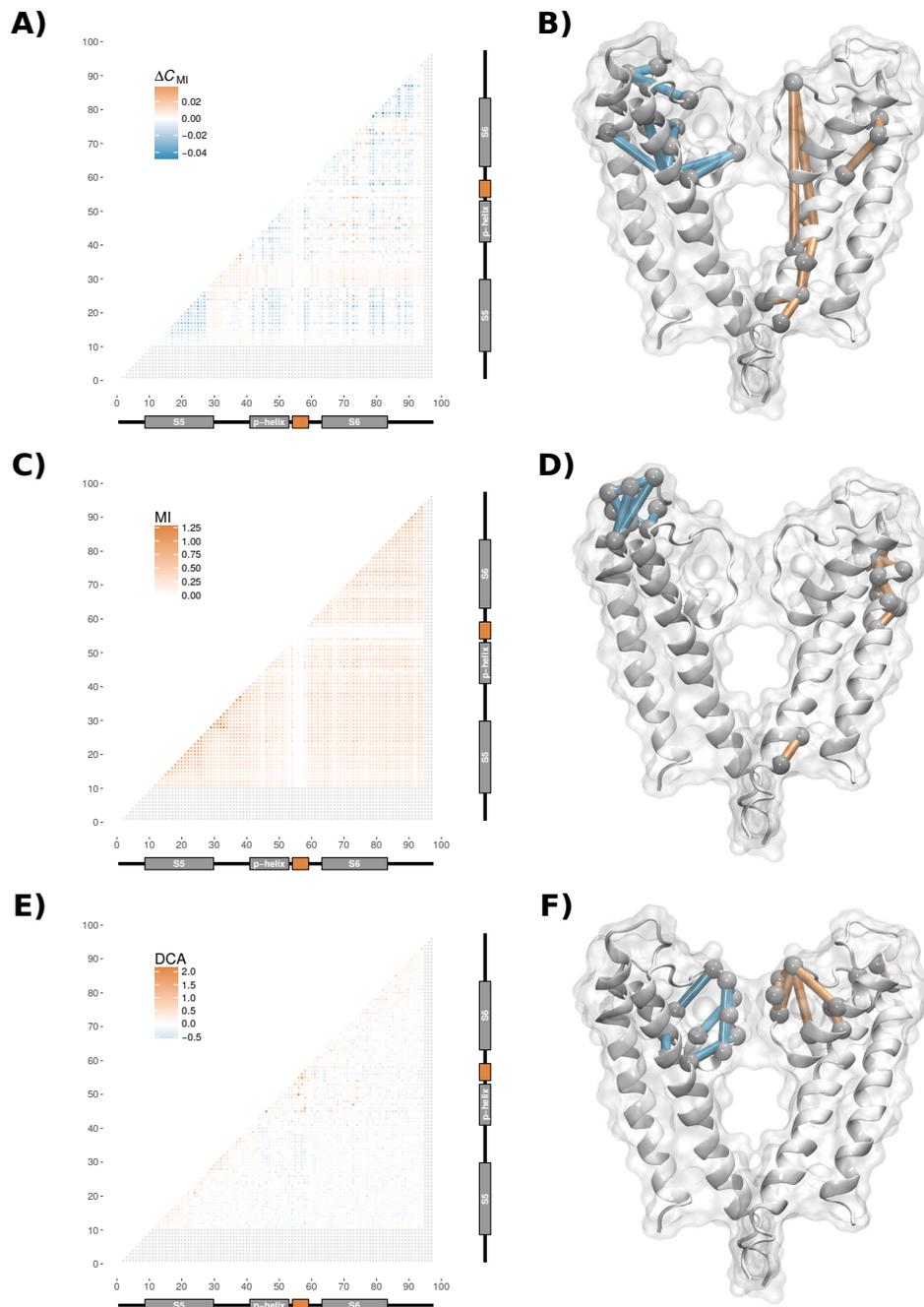


Figure 3.6.: **A)** Co-evolutionary complexity ΔC_{MI} between sequences from prokaryotic and eukaryotic organisms. **B)** Visualization of the highest 10 ΔC_{MI} pairs. **C)** Mutual information of the prokaryotic sequences. **D)** Visualization of the highest 10 MI pairs. **E)** Results from the Direct Coupling Analysis of the prokaryotic sequences. **F)** Visualization of the highest 10 DCA results. In **B)**, **D)** and **F)** blue sticks represent the top 10 eukaryotic pairings and orange sticks the top 10 prokaryotic pairings.

the N-terminal portion of the inner helix we observe the inverted effect with prokaryotic channels experiencing higher co-evolutionary complexity. Furthermore, the turret region connecting S5 with the pore helix experience a similar high complexity in prokaryotic channels. Additionally, we also find increased co-evolutionary complexity between the central residues of the S6 region and the N-terminal residues of KcsA. The highest ΔC_{MI} values connect predominantly residues found in the S6 helix to other regions, such as connecting position 74 (in S6) to position 25 (in S5). Contrarily, the lowest ten ΔC_{MI} values – i.e. higher co-evolutionary complexity found in eukaryotic sequences – are located near the selectivity filter (see Fig. 3.6B)).

Similarly, we find the highest 10 mutual information (MI) results for the eukaryotic dataset to be located at the turret region connecting S5 and the pore helix (see Fig. 3.6D)). For the prokaryotic sequences we find the top 10 pairings in a distinct co-evolution pattern in the upper region of the outer helix (S6 helix). The MI in Fig. 3.6C) shows the conservation of the filter sequence in prokaryotic sequences while the S5 helix shows a high degree of variability. The direct coupling analysis (DCA) results show very different patterns for the ten most important contacts when compared to MI and ΔC_{MI} . Here, for either alignment the top ten DCA values are located in the filter region, connecting the conserved filter sequence with the p-helix (see Fig. Fig. 3.6F)). Other important interactions for prokaryotic structures predicted by DCA are found in S5 helix as well as between the p-helix and the S6 helix (see Fig. 3.6E)).

3.4 Discussion

In this chapter we took an in depth look at the performance and robustness of well-established information theoretical measures. Here, our results show consistent and reliable results for MI and few MI-based methods when only very small fractions of the actual dataset were used. Thus, for analysis of smaller datasets, methods as RCW, APC and MI are less influenced by sampling bias and can already achieve reliable results. Furthermore, Z-score statistics for APC and RCW experience a higher degree in robustness than Z_{MI} . The novel ΔC_{MI} experiences consistent results for smaller alignments in comparison to the *ground truth* for the top interactions even at low sequence numbers.

Furthermore, this novel approach to identifying complex co-evolutionary relationships in proteins derived directly from (Keul and Hamacher, 2017) represents a unique way to extract information from sequence data. Comparing ΔC_{MI} results to other information theory based methods revealed that the intricate evolutionary mechanisms captured by ΔC_{MI} cannot be reproduced by other measures. By comparing potassium channel sequences from two different Pfam families (PF00520 and PF07885) we were able to identify large differences in the evolutionary mechanics in the S5 helical region. Here, larger potassium channels experience less complex evolutionary mechanism, resulting in more conservation of certain amino acids or physico-chemical properties. When considering structural aspects of channels, we find interactions between the voltage sensing domain and the N-terminal part of S5 within voltage-dependent six TMD. The close proximity of the voltage sensor increases the evolutionary pressure on residues found in the S5 helix and changes the evolutionary mechanics drastically.

Furthermore, we were able to show that MI, DCA and ΔC_{MI} document entirely different evolutionary relationships. While DCA reportedly covets co-evolving contacts important for folding of the protein

(Morcos *et al.*, 2011; Baldassi *et al.*, 2014), ΔC_{MI} reveals different types of interactions through describing changes in evolutionary complexity in reference to a different (background) model. For both investigated organism domains (prokaryotic and eukaryotic) DCA showed that contacts within the pore region were the most important for the structure of the protein. Contrarily, the greatest differences in co-evolutionary complexity were found close to the turret region and the N-terminal part of the outer helix for eukaryotic potassium channels. Prokaryotic channel sequences showed a more intricate evolutionary mechanism at the g-helix, indicating a significant influence of g-helix guided gating on the amino acid composition in this region.

4 Dynamics of Channel Proteins

The dynamics and behaviors of proteins is of fundamental interest in biophysics and molecular biology. Up to date various approaches have been developed to analyze protein dynamics. These range from pure experimental approaches such as nuclear magnetic resonance spectroscopy of proteins (protein NMR, e. g. Barbato *et al.* (1992); Ishima and Torchia (2000); Jarymowycz and Stone (2006)) through combined quantum mechanics and molecular dynamics simulations of proteins (QM/MM, e. g. Liu *et al.* (2001); Hensen *et al.* (2004); Steinbrecher and Elstner (2013)) to coarse grained models like Gaussian Network Models (GNM, e. g. Bahar *et al.* (1997); Yang *et al.* (2005); Hamacher (2011)). In general, motions within a protein can range from fast vibrations of atom bonds to the very slow folding of entire proteins (Lindahl, 2008).

In this chapter we will analyze the *general* behavior characteristics of channel proteins. Based on the hypothesis that potassium channels – and in a broader sense channels in general – have fundamentally similar dynamics, we stipulate that these dynamics are mainly attributed to the structural organization of functional blocks, so called regions. To this end, we will investigate any functional dependencies between the structural elements of channel proteins. Here, we will employ Anisotropic Network Models which allow us to implement distance and sequence dependent parametrization of coarse grained interaction networks. With these we will also try to elucidate whether any found structural dependency is linked to the amino acid composition or can be attributed mostly to spatial organization of the channel proteins. Furthermore, we will show a decoupling of the filter region from other regions based on free energy perturbation experiments.

4.1 General channel organization

In order to elucidate the aforementioned characteristics of channels, we will use a wide variety of channel structures obtained from classic crystallography methods or homology modeling paired with molecular dynamics simulations (see Tab. 4.1). we will focus on two main groups of ion channel architecture: the small and sometimes minimalistic two transmembrane domain channels (two TMD channels, with segment S5 and S6 in the literature) and the larger, sometimes voltage-dependent six transmembrane domain channels (six TMD channels, with segment S1 through S6 in the literature). In the upcoming paragraphs, we will take a closer look at each structure included in this analysis.

In general, all channels presented here possess a very similar structural organization. All channels are homotetramers with each chain consisting of at least one outer helix (S5), a pore helix, a selectivity filter and a inner helix. Here, we chose to subdivide the architecture of all channels in the following regions:

s-helix The slide helix (s-helix) has been suggested to be involved in the mechanical gating of potassium channels (Long *et al.*, 2005; Tao *et al.*, 2010; Zubcevic *et al.*, 2014; Lefoulon

Table 4.1.: Table of structures analyzed with their availability and reference listed.

	channel name	conformations	source	reference
2 TMD	Kcv	open & closed	personal communication	Tayefeh <i>et al.</i> (2009)
	KirBac1.1	closed	PDB: 1P7B	Kuo <i>et al.</i> (2003)
	KcsA	closed	PDB: 1BL8	Doyle <i>et al.</i> (1998)
	KirBac3.1	closed	PDB: 2WLJ	Clarke <i>et al.</i> (2010)
		open	PDB: 3ZRS	Zubcevic <i>et al.</i> (2014)
	MthK	open	PDB: 4HJO	Posson <i>et al.</i> (2013)
	GIRK2	closed	PDB: 3SYA	Whorton and MacKinnon (2011)
	NaK	closed	PDB: 2AHY	Shi <i>et al.</i> (2006)
		open	PDB: 3E86	Alam and Jiang (2009)
	TrkH		3PJZ	Cao <i>et al.</i> (2011)
6 TMD	Kv1.2	open	PDB: 3LUT	Chen <i>et al.</i> (2010)
	Kv1.2-2.1 chimera	open	PDB: 4JTA	Long <i>et al.</i> (2007)
	KvAP	open	PDB: 1ORQ	Jiang <i>et al.</i> (2002)
	KAT1	open & closed	personal communication	Lefoulon <i>et al.</i> (2014)
	NavAB	closed	PDB: 3RVY	Payandeh <i>et al.</i> (2011)
	MloK1	open & no cAMP bound	PDB: 4CHW	Kowal <i>et al.</i> (2014)
		open & cAMP bound	PDB: 4CHV	
	closed with no CNBD	PDB: 3BEH	Clayton <i>et al.</i> (2008)	

et al., 2014) and is located at the N-terminal end of the outer helix (S5). We chose to adopt the "s-helix" terminology for the S4-S5-linker helix of six TMD channels as well.

TM5 The region we consider to be the outer helix is named TM5 to avoid ambiguities with the canonically used term S5. The TM5 consists (mostly) of the outer helix (S5 segment) unless stated otherwise.

p-helix The pore helix (p-helix) belongs to the inner part of the channel and connects the selectivity filter with the outer helix. Through the positioning of the p-helix, cations in the central cavity are stabilized by dipole interactions (Doyle *et al.*, 1998; Jogini and Roux, 2005).

filter The selectivity filter presents a central, ion specific amino acid sequence with four selectivity filters building the highly hydrophobic permeation pore (Miller, 2000; Frank and Catterall, 2003; Frank *et al.*, 2005). Due to the selectivity filter, ions crossing the membrane and permeating the pore are stripped of their surrounding hydrating water molecules.

TM6 In some channels, the N-terminal part of the inner helix (the S6 segment) was reported as being involved in the mechanical gating (Kuo *et al.*, 2003; Tao *et al.*, 2009; Whorton

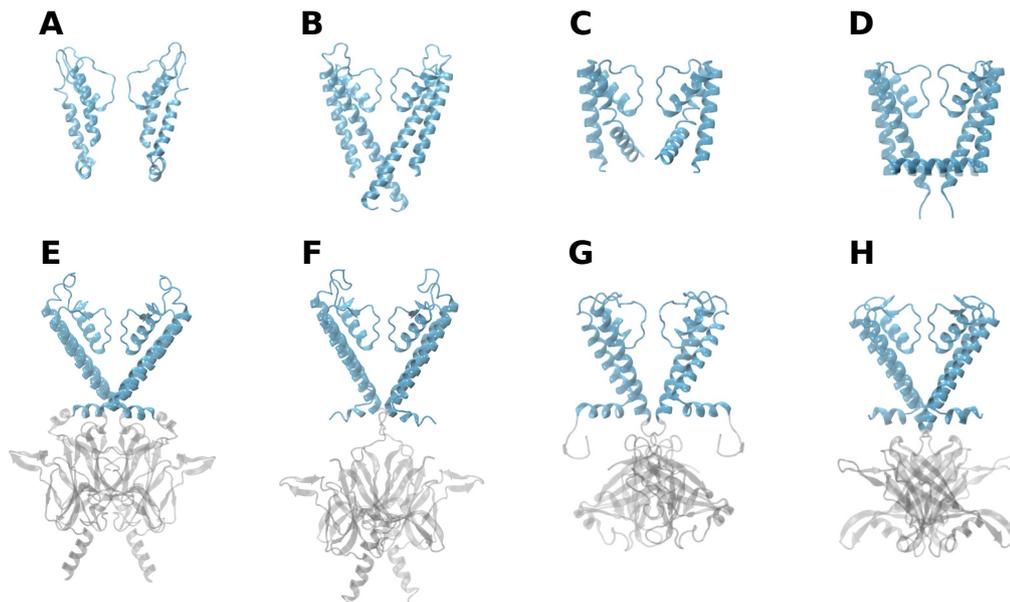


Figure 4.1.: Overview of the two transmembrane domain channels presented in this study. The central pore region is colored in blue with non-pore regions colored in gray. Here, the channels are: (A) Kcv (closed-state), (B) KcsA (closed-state), (C) MthK (open-state), (D) NaK (closed-state), (E) GIRK2 (closed-state), (F) Kir2.2 (closed-state), (G) KirBac1.1 (closed-state) and (H) KirBac3.1 (closed-state). Only two opposing monomers of the tetramers are shown.

and MacKinnon, 2011; Bavro *et al.*, 2012). Hence, we divide the inner helix into two functionally different regions. The TM6 part of the S6 segment is not involved in the gating of the channel.

g-helix The gate helix (g-helix) is the C-terminal part of the S6 segment and heavily involved in the mechanical gating of channels through bundle crossing (Jiang *et al.*, 2002; Thompson *et al.*, 2008; Payandeh *et al.*, 2011; Bavro *et al.*, 2012). We consider the g-helix region to be introduced by a kink in the S6 segment in open channel structures. Amongst others, these kinks can be produced by occurrences of the amino acids proline and glycine (Levitt and Greer, 1977) in the S6 segment after twelve to 15 residues.

Additionally, we find a voltage sensing domain (VSD) in the S4 segments of voltage-gated channels such as the here listed Kv, KAT1 and NavAB channels. For simplicities sake, we assume that in all six TMD channels the VSD – or a functional correlate thereof – is located in the fourth transmembrane segment.

4.1.1 Two TMD Ion Channels

The most simplistic organizational structure for ion-specific channel activity is presented by the tetrameric organization of channel monomers with two transmembrane domains (TMD). Generally, these channels possess at least two TMD, one pore helix and a channel specific filter sequence in each monomer. Major differences in architecture can be found at the N-terminus of these channels. Here, channels such as KirBac1.1 possess a so called slide helix (s-helix), which has been suggested

Table 4.2.: Listing of the range of regions of the two transmembrane domain channels. Closed-state structures are indicated with (-) and open-state with (+). In cases where both structures have differing indices for regions, superscript ⁽⁻⁾ and ⁽⁺⁾ denote the indices for closed- and open-state configurations respectively. Inapplicable regions are indicated with n/a.

	structure	VSD	s-helix	TM5	p-helix	filter	TM6	g-helix
2 TMD	Kcv ^{(-)/(+)}	n/a	1 – 13	14 – 32	50 – 62	63 – 68	76 – 86	87 – 94
	KirBac1.1 (-)	n/a	47 – 59	61 – 83	97 – 109	110 – 115	121 – 137	138 – 150
	KcsA (-)	n/a	1 – 8	9 – 30	41 – 53	54 – 59	63 – 83	84 – 103
	KirBac3.1 ^{(-)/(+)}	n/a	35 – 43 ⁽⁻⁾ 32 – 43 ⁽⁺⁾	47 – 71	83 – 95	96 – 101	107 – 123	124 – 136
	MthK (+)	n/a	n/a	18 – 43	46 – 58	59 – 64	70 – 83	84 – 99
	Kir2.2 (-)	n/a	70 – 79	80 – 108	130 – 142	143 – 148	156 – 169	170 – 183
	GIRK2 (-)	n/a	82 – 90	91 – 120	141 – 153	154 – 159	167 – 180	181 – 196
	NaK ^{(-)/(+)}	n/a	1 – 20 ⁽⁻⁾ n/a ⁽⁺⁾	21 – 45 ⁽⁻⁾ 23 – 45 ⁽⁺⁾	50 – 62	63 – 68	74 – 89	90 – 105
		n/a	n/a	65 – 90	98 – 109	110 – 115	125 – 139	140 – 145
	TrkH	n/a	n/a	178 – 202	207 – 218	219 – 224	237 – 250	253 – 261
		n/a	267 – 274	275 – 297	303 – 318	319 – 324	333 – 344	358 – 376
		n/a	n/a	393 – 419	424 – 435	436 – 441	454 – 469/a	472 – 479
n/a		n/a	393 – 419	424 – 435	436 – 441	454 – 469/a	472 – 479	
6 TMD	Kv1.2 (+)	289 – 311	312 – 324	325 – 351	361 – 373	374 – 379	387 – 405	406 – 421
	Kv1.2-2.1 chimera (+)	279 – 305	309 – 320	321 – 347	357 – 369	370 – 375	381 – 401	402 – 417
	KvAP (+)	114 – 133	135 – 147	148 – 172	183 – 195	196 – 201	207 – 225	226 – 240
	KAT1 ^{(-)/(+)}	160 – 181	182 – 195	196 – 220	247 – 259	260 – 265	273 – 293	294 – 307
	NavAB (-)	1096 – 1112	1116 – 1129	1130 – 1153	1163 – 1175	1176 – 1182	1194 – 1206	1207 – 1219
	MloK1 ^{(-)/(+)} & ^{(-)/((+))} ¹	98 – 111 ²	114 – 125	126 – 151	162 – 174	175 – 180	186 – 208	209 – 228 ⁽⁻⁾ 209 – 232 ⁽⁺⁾

¹ ((+)) and ((-)) denote the cAMP bound and free conformation of MloK1.

² Voltage-sensing properties of MloK1 S4 have not been reported.

to be involved in opening and closing of the channel, whereas other channels (such as KcsA) do not possess such a feature. For channels without a s-helix we subdivided the first TMD (TM5) into a short N-terminal segment – mirroring the s-helix in other channels – and the actual, transmembrane segment (TM5). The corresponding residue indices for all two TMD channels can be seen in Tab. 4.2.

Kcv

Kcv is the potassium channel from the chlorella virus PBCV-1 and presents a prime example for the smallest channel structure (see Fig. 4.1A). Since no crystal structure of Kcv is available, the structure of this viral potassium channel had been derived via homology modeling by Tayefeh *et al.* (2009). Here, KirBac1.1 was used as a template and the structures were validated via MD simulations in

a membrane environment. In these MD simulations ion transition through the pore was observed. This model – with a deprotonated Lysine 29 – serves us as an open-state Kcv. The protonated Kcv variant showed no ion conductivity can be measured and thus the channel can be considered in a closed-state. In Kcv, we opt for subdividing the inner helix at position 86, as the threonine found here can induce bending of alpha helices (Ballesteros *et al.*, 2000). From position 87 onwards we find residues belonging to the g-helix, a part we assume capable of forming a channel gate. If the g-helix of Kcv is indeed involved in gating of the channel, comparing open and closed structures should reveal differences on a structural level.

KirBac1.1

KirBac1.1 is a member of the inward rectifying potassium channel family (Kuo *et al.*, 2003; Cheng *et al.*, 2009). The channel structure consists of the α -helical and pore-forming component as well as an C-terminal, intracellular, mostly of β -sheets consisting part, the so called cytoplasmic vestibule (see Fig. 4.1G). Kuo *et al.* (2003) suggested that movement of the pore helix is involved in opening of the channel. Furthermore, interactions between the slide helix and the inner helix (TM6) were proposed as vital elements in opening the channel due to electrostatic interactions of side chains. For the channel to move into an open conformation, motion at the C-terminal part of the inner helix (TM6) was deemed necessary. Here, an indicative bend in the TM6 helix can be observed at position 137. The glycines found after this position initiate a bend in the inner helix (Levitt and Greer, 1977), and we chose to divide the inner helix at this position into the TM6 region and a proposed gate forming helical structure (g-helix).

KirBac3.1

The inward rectifying potassium channel KirBac3.1 of *Magnetospirillum magnetotacticum* as a prokaryotic homolog of the eukaryotic Kir channels. Similar to KirBac1.1, KirBac3.1 possesses a large C-terminal, cytoplasmic domain as well as an N-terminal slide helix which are both crucial elements in the transitional twisting between closed-state and open-state conformations (Zubcevic *et al.*, 2014) (see Fig. 4.1H). Bavro *et al.* (2012) showed that single site mutations in the inner helix (TM6) can lead to an open channel conformation by removing the mechanical closing via bundle crossing of the TM6-helices. Additionally, activation mutations in the cytoplasmic domain of KirBac3.1 were identified, hinting at a connection between channel conductivity and interplay between the cytoplasmic domains. The findings by Bavro *et al.* (2012) hint at two distinct gates in KirBac3.1, a g-helix formed gate as well as a cytoplasmic gate. In order to address the gate formed by crossing of helices, we chose to subdivide the inner pore helix at position 123 into a transmembrane part (TM6, residue 107 to 123) and a so called g-helix (residue 124 to 136). Comparing open- and closed-state conformations of KirBac3.1 should reveal significant differences at the g-helix due to mechanical closing of the channel.

Kir2.2

Kir channel proteins are involved in the regulation of resting membrane potential. These eukaryotic channels share high similarities to their prokaryotic counterparts of the KirBac family. As such, Kir2.2 allows us to directly compare structural dependency patterns between eukaryotic and prokaryotic

channels. Here, we use the closed-state Kir2.2, which is structurally closed at the activation gate" or g-helix (Tao *et al.*, 2009). This closure at the g-helix differs from the bundle crossing in the closed KcsA (Tao *et al.*, 2009). In Kir2.2 we find a glycine – similarly positioned to the ones found in KirBac1.1 and KirBac3.1 – at position 169 in the crystal structure (see Fig. 4.1F). While no kink is observable in the structure, we assume that residues 170 to 183 belong to the g-helix and are capable to close the channel mechanically. Hence, we consider these residues as structural homolog to the g-helix forming residues found in both prokaryotic Kir channels.

GIRK2

G-protein-gated inward rectifying (GIRK) channels form the channel pore with two transmembrane domains and a Kir channel common cytoplasmic domain (CTD) (see Fig. 4.1E). As the name implies, GIRK channels require G proteins in complex with PIP₂ to activate. GIRK2 – also known as Kir3.2 – shares high similarities to Kir2.2 with the exception of highly structured turret regions and the structural orientation of the g-helix and the CTD (Whorton and MacKinnon, 2011). The GIRK2 structure has been hypothesized to possess two gates: one formed by the g-helices and another by parts of the CTD. As GIRK2 is shut at both gates (Whorton and MacKinnon, 2011), we will be able to compare the configurations of other structures to this channel closing at the g-helices. Here, the g-helix begins with serine directly following a glycine at position 180. Serine and glycine have been reported to distort α -helices (Levitt and Greer, 1977; Ballesteros *et al.*, 2000).

KcsA

Being the first potassium channel crystallized by Doyle *et al.* (1998), KcsA has been used as prime example for analysis of potassium channel function and structural comparison. While KcsA stems from the prokaryotic *Streptomyces lividans*, its amino acid sequence shows great resemblance to sequences of eukaryotic six TMD channels such as the Kv channels, especially in the pore region (see Fig. 4.1B). Whereas the opening and closing of Kv channels is mediated via changes in the membrane potential, the inward rectifying KcsA is mediated by pH. In acidic pH, this minimalistic potassium channel favors the open-state configuration (Zimmer *et al.*, 2006; Thompson *et al.*, 2008). Kuo *et al.* hypothesized that the truncated KcsA structure of Doyle *et al.* (1998) is in fact an open channel with a closed gating section based upon the orientation of the pore helix. Even though the structure has been discussed widely, KcsA has been considered the prime example of a physically shut potassium channel through the so called bundle crossing. Hence, we assume that the g-helix is started with a glycine in position 83. While the here used structure of KcsA is without a visible slide helix (s-helix), we can assume that residues 1 through 8 could function in a similar matter through the distortion of the helix entropy at the glycine in position 9 and a subsequent kink prior to this residue. The extended structure was derived from KcsA of Doyle *et al.* (1998) by S. Tayefeh.

MthK

The MthK from *Methanobacterium thermoautotrophicum* is a Ca²⁺-dependent gated potassium channel. Gating through the straightening of the inner helix (TM6) and subsequent bundle crossing had been suggested (Jiang *et al.*, 2002). In contrast, Posson *et al.* (2013) showed that MthK possesses a voltage-

dependent gate close to the selectivity filter. Here, we will use the blocked, open-state MthK structure to investigate structural and fold similarities of the MthK pore module with other, more complex channels (see Fig. 4.1C). In order to assess the physical gating of the channel at the C-terminal end of the inner helix, we subdivide this α -helix in a TM6 segment. Residues following a glycine at position 83 are assumed to belong to the g-helix.

NaK

While the aforementioned potassium channels show a high potassium specificity, other channels such as NaK possess reduced ion selectivity. This non-selective cation channel from *Bacillus cereus* can conduct both sodium and potassium ions across the membrane (Shi *et al.*, 2006). Whereas selectivity filter in potassium channels consists of the amino acid sequence TVGFG, in NaK's selectivity filter the aromatic phenylalanine is replaced by charged aspartic acid. Due to this substitution the architecture of the pore is changed, allowing sodium as well as potassium ions to flow through the pore (Vora *et al.*, 2008) (see Fig. 4.1D for the NaK pore). Due to this, NaK presents with the opportunity to investigate the influence of altered pore architecture on structural relationships. Additionally, Alam and Jiang (2009) reported the crystal structure of an open NaK channel. This structure was derived by truncating the N-terminal s-helix. Here, we will compare structural connections of both the open-state (Alam and Jiang, 2009) and the closed-state channels (Shi *et al.*, 2006). For both configurations we assume that residues with the index ranging from 90 to 105 form the g-helix. Again, the g-helical part of the inner TMD is preceded by a Glycine at position 89.

TrkH

The TrkH potassium ion transporter from *Vibrio parahaemolyticus* has presumably evolved from simplistic potassium channels (Cao *et al.*, 2011). Whereas the channels presented here function as homotetramer, TrkH shows a monomeric organization of four low homology domains. Each domain consists of channel-like regions similar to an inner, outer and a pore helix. This structural organization rudimentary resembles homologous segments in ion channels. Contrary to potassium channels, TrkH selectivity filter regions are highly degraded when compared to the conserved filter of potassium channels (such as Kcv, KirBac1.1 or Kv channels). While the asymmetric unit is build as dimer of TrkH, we will elucidate the structural dependencies between the TrkH regions (see Table 4.1) based on a monomeric structure. Regional blocks were appointed according to the annotation in the Protein Data Bank (Berman *et al.*, 2000) and structural resemblance to regions of other channels.

4.1.2 TM6 Ion Channels

Channels with six transmembrane domains are the other major group of channel proteins we will focus on. From a structural point of view, six TMD channels share most features with their smaller 2TMD counterparts but inherit structural extension in form of additional four N-terminal segments (S1 to S4). In voltage-gated 6TMD channels, such as the Kv-channel family, the S4 helix contains positive charged residues. These residues play an essential role in the channels reaction to changes in the membrane potential by moving this voltage sensing domain (VSD) through the membrane (Yang *et al.*, 1996; Mannuzzu *et al.*, 1996; Yusaf *et al.*, 1996; Baker *et al.*, 1998; Long *et al.*, 2005; Lefoulon

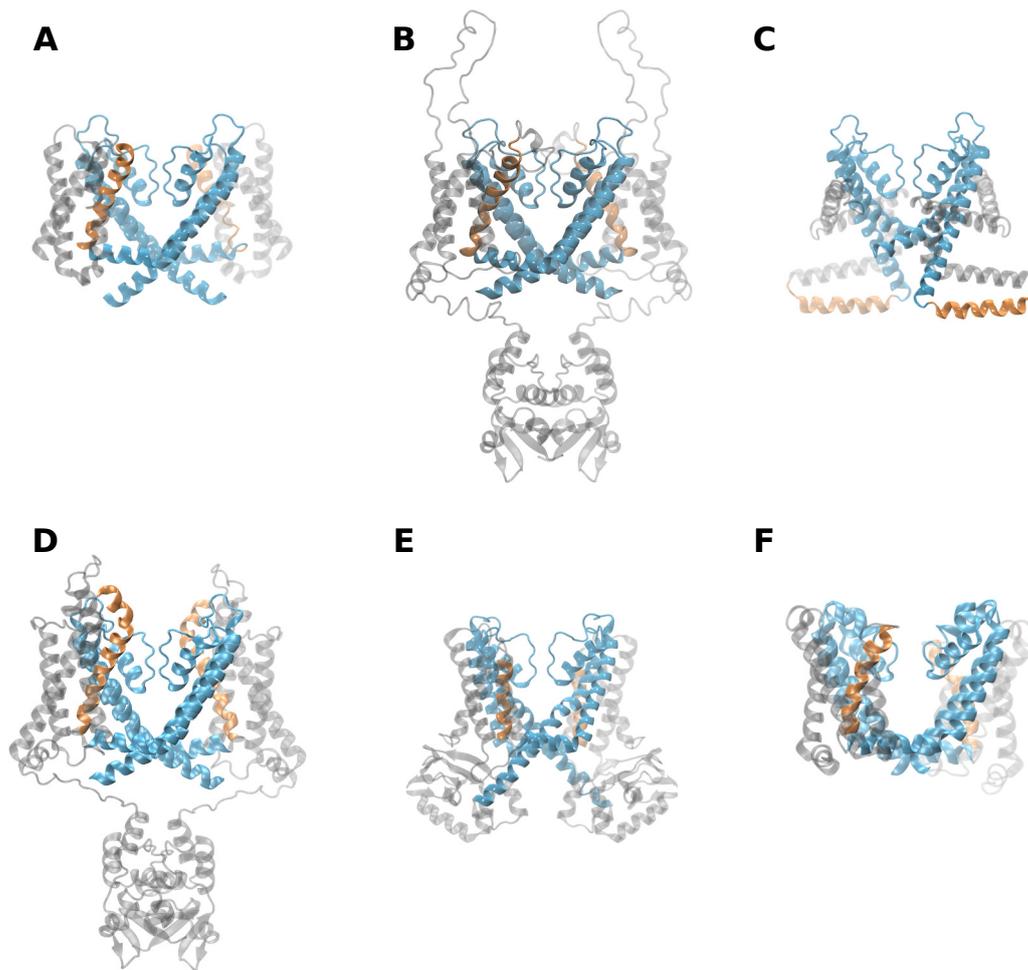


Figure 4.2.: Overview of the six transmembrane domain channels presented in this study. The central pore region is colored in blue while the VSD (the S4 segment) appears in orange. Non-pore regions are colored in gray. Here, the channels are: **(A)** KAT1 (closed-state), **(B)** Kv1.2 (open-state), **(C)** KvAP (open-state), **(D)** Kv1.2-2.1 paddle chimera (open-state), **(E)** MloK1 (open-state) and **(F)** NavAB (closed-state). Only two opposing monomers of the tetramers are shown.

et al., 2014). This movement of the VSD has been reported to produce a cork-screw-like movement of the VSD when the channel moves from closed- to open-state (Campos *et al.*, 2007; Pathak *et al.*, 2007; Jensen *et al.*, 2012; Li *et al.*, 2014). To compare the structural regions of six TMD channels with their smaller two TMD homologs, we label the S4-S5 linker helix as s-helix in order to compare differences in interaction of the linker helix with the slide helix predominantly found in the two TMD channels.

KAT1

The inward-rectifying potassium channel of *Arabidopsis thaliana* (KAT1) has been suggested to play a pivotal role in the opening of stomata in plants (Kwak *et al.*, 2001) with KAT1 inhibition leading to stomata closure (Pandey *et al.*, 2007). As a voltage-dependent eukaryotic channel, KAT1 last two of its six TMD form the channel pore (see Fig. 4.2A). The VSD in S4 is linked through a helical S4-S5 linker to the actual ion conducting part of the channel. Movement of this voltage sensor has been studied heavily (Lefoulon *et al.*, 2014) revealing a highly volatile VSD organization.

Through this, open- and closed-state KAT1 conformations were achieved through single site mutations at a central phenylalanine within S2 and S3 helices. Here, we will use the open- and closed-state structures obtained from molecular dynamics simulations (Lefoulon *et al.*, 2014) to gain insight on configurational reorientations occurring upon opening of voltage-dependent channels. Observably, KAT1 possesses a kink in the inner helix of the pore domain introduced by the occurrence of glycine in position 293. Hence, we are able to subdivide the S6 segment into a transmembrane part (TM6, position 273 to 293) and the gate helix (g-helix, position 294 to 307).

Kv1.2

Kv1.2 is a voltage-gated, inward-rectifying, delayed rectifier channel mediating potassium ions across membranes of excitable cells, found in, for example, the brain and cardiovascular system (see Fig. 4.2B). Here, Kv1.2 opens in response to the depolarization of the membrane and spontaneously closes slowly. Through movement of the the S4 segment, voltage induced opening and closing of the channel is mediated. Hereby, S4 moves at least partially through the membrane environment with subsequent conformational changes propagating fold change information through the S4-S5-linker to the actual pore (Long *et al.*, 2005; Tao *et al.*, 2010). Channel gating of Kv channels has been shown to be dependent upon a highly conserved triplet sequence of proline-X-proline, where X can be any amino acid. For the purpose of comparing structural properties of channels in general, we chose to divide the inner helix into two structural regions. Starting with the proline at position 406 (the second proline of the aforementioned triplet) we consider all following residue as members of the g-helix region.

KvAP

KvAP is a slow inactivating potassium channel from the archaeobacteria *Aeropyrum Pernix* with similarities to the eukaryotic Kv1.2 (Ruta *et al.*, 2003). Despite the similarities in the amino acid sequences, Kv1.2 and KvAP display great structural differences (see Fig. 4.2C). Whereas the VSD of Kv1.2 is in an almost vertical position within the membrane, the VSD of KvAP has been reported to be positioned horizontally along the membrane (Jiang *et al.*, 2003). While others have suggested that more conventional corkscrew VSD models can be developed for KvAP (Shrivastava *et al.*, 2004; Lee *et al.*, 2005) we will use the originally proposed model (Jiang *et al.*, 2003) for our analysis of the interactions of pore segment. The subdivision of the inner pore helix into a TM6 part and a g-helical part is inspired by a kink at position 225. The here found threonine is known to induce bending in α -helices (Ballesteros *et al.*, 2000). As the S4-S5-linker between the VSD and the outer pore helix is missing while the actual outer helix is elongated, we adopted the subdivision of the outer pore helix as in Jiang *et al.* (2003).

Kv1.2-Kv2.1 Paddle Chimera

The so-called paddle chimera of Kv1.2 with a modified voltage-sensor paddle from Kv2.1 can be considered as a prime example for the modularity of channels. Here, exchanging the VSD of one channel with a conceptually similar but sequentially different VSD of another channel produced a fully functional ion channel (Long *et al.*, 2007) (see Fig. 4.2D). Comparing interactions between the VSD and all other regions from the original Kv1.2 structure (Long *et al.*, 2005) with its derivative,

Kv1.2-Kv2.1 paddle chimera can reveal necessary interactions for voltage-dependent gating. Here, we use a paddle chimera with charybdotoxin (CTX) bound to the extracellular pore entry. CTX blocks the entry of potassium ions into the filter region formed by the four monomers of the channel without altering the conformation of the pore (Banerjee *et al.*, 2013). Therefore, the here presented paddle chimera represents a blocked but open channel, as crystallization generally occurs at 0 mV (Long *et al.*, 2005; Banerjee *et al.*, 2013). To achieve a closed structure through voltage-dependent gating, Long *et al.* (2007) hypothesized a straightening of this inner helix as well as vertical movement of the S4-S5-linker (s-helix) induced by movement of the VSD. In order to address this proposed bundle crossing upon closing of the channel, we subdivide the sixth TMD into two parts: a TM6 part (residue 381 to 401) and the g-helix (residue 402 to 417) as we can observe a kink in the inner helix following a proline at position 401.

MloK1

Cyclic-nucleotide regulated channels possess six TMD and a C-terminal, cytoplasmic cyclic-nucleotide binding domain (CNBD). Members of this class of channel proteins are – amongst others – the hyperpolarization-activated HCN channels and the prokaryotic, homologous MloK1. Originating from the gram negative bacteria *Mesorhizobium loti*, MloK1 shares many features of voltage-gated channels. It presents an arrangement of six TMDs of which the last two (S5 and S6) form a typical conformation of an ion conducting pore as a homotetramer (see Fig. 4.2E). Structurally, S1 to S4 are found in a conformation similar to the first four TMDs of the voltage-gated Kv1.2 (Clayton *et al.*, 2008). Nonetheless, as the VSD of MloK1 is populated by fewer positively charged residues than, for example, KAT1 and Kv1.2, voltage-sensing properties of this regions have yet to be established (Kowal *et al.*, 2014). We will investigate conformational changes between the cAMP bound and unbound states of the open, full length MloK1 (Kowal *et al.*, 2014). Furthermore, we will take a closer look into configurational changes between these open-state structures and the closed-state structure of Clayton *et al.* (2008), a MloK1 structure without the CNBD. While no voltage-sensing characteristic in the S4 of MloK1 have been reported, we will analyze the influence of this missing property on interactions of other regions with the VSD. Hence, we consider the residues 98 to 111 for all three MloK1 structures as part of the VSD. Clayton *et al.* (2008) suggested that the S4-S5-linker helix (here called s-helix) influences the gating of the channel through interaction with the C-terminal part of the inner helix (i.e. the g-helix). We consider the g-helix to start after the glycine at position 208 in all three structures.

NavAB

The voltage-gated sodium channel NavAB from *Acrobacter butzleri* belongs to the bacterial NaChBac channel family, a structure-function model for the more complex eukaryotic sodium channels with up to 2000 amino acids. Whereas voltage-gated potassium channels terminate the action potential, the voltage-gated sodium counterpart initiates action potential. Despite this opposed mode of operation in eukaryotic cells, the model sodium channel NavAB shares high structural similarities with Kv channels (Payandeh *et al.*, 2011) (see Fig. 4.2F). We will compare structural dependencies of the closed-state NavAB structure to such in voltage-gated potassium channels. To investigate the mechanical gating at the N-terminal part of the S6 segment, we subdivide the inner helix at position 1206. Here, we find a threonine which could form a kink in the inner helix of the pore, similarly to the threonine found in

the same region in Kcv and KvAP. Hence, TM6 region ranges from residues 1194 to 1206, while the g-helix ranges from positions 1207 to 1219.

4.2 Background

Anisotropic Network Models (ANM) were developed by Atilgan *et al.* (2001) as an extension of the Gaussian Network Models (GNM) (Bahar *et al.*, 1997). Both are coarse grained approaches to characterize protein dynamics through the generalization of complex intra-molecular interaction potentials conceptually close to the single parameter Normal Mode Analysis (Tirion, 1996). Generally, in Elastic Network Models (ENM, such as GNM and ANM) amino acids are treated as single beads at the coordinates of the C α atoms, connected by Hooke's springs to one another. Whereas in GNM residue fluctuation around the equilibrium structure is assumed to behave isotropic and Gaussian distributed, ANM separate the spatial displacement of the beads into X-, Y- and Z directions. Hence, the harmonic potential V of the displacement around the equilibrium structure can be expressed as:

$$V = \frac{1}{2} \sum_{i,j|i \neq j} \gamma_{ij} (\mathbf{r}_{ij} - \mathbf{r}_{ij}^0)^2 \quad (4.1)$$

Here, γ_{ij} describes the interaction strength, through a scalar spring constant, between the i th and j th residues. \mathbf{r}_{ij}^0 and \mathbf{r}_{ij} symbolize the displacement vector between the coordinates of the i th and j th amino acids in the equilibrium and a perturbed structure, respectively. In the original work from Atilgan *et al.* (2001), a uniform γ_{ij} was chosen if the distance $|\mathbf{r}_{ij}^0|$ was below a defined threshold. Dehouck and Mikhailov (2013) proposed a distance and interaction type specific parameterization of γ_{ij} in form of a $\gamma(|\mathbf{r}_{ij}^0|, \alpha_i, \alpha_j)$. Here, α_i and α_j represent interacting amino acids i and j with $|\mathbf{r}_{ij}^0|$ is the distance between these. While in the model of Dehouck and Mikhailov the beads definition remains the same, the physico-chemical interaction properties between the residues are reflected within $\gamma(|\mathbf{r}_{ij}^0|, \alpha_i, \alpha_j)$. The distance and amino acid interaction specific coupling parameter are derived from fluctuations of over 1500 NMR structures.

With γ_{ij} or the γ -function being independent of spatial components of \mathbf{r} , the Hessian Matrix in the ANM can be expressed as the second term of a Taylor expansion of the potential (see Eqn. 4.1) with respect to the X, Y and Z directions.

$$\mathbf{H}_{ij} = \begin{bmatrix} \partial^2 v_{ij} / \partial X_i \partial X_j & \partial^2 v_{ij} / \partial X_i \partial Y_j & \partial^2 v_{ij} / \partial X_i \partial Z_j \\ \partial^2 v_{ij} / \partial Y_i \partial X_j & \partial^2 v_{ij} / \partial Y_i \partial Y_j & \partial^2 v_{ij} / \partial Y_i \partial Z_j \\ \partial^2 v_{ij} / \partial Z_i \partial X_j & \partial^2 v_{ij} / \partial Z_i \partial Y_j & \partial^2 v_{ij} / \partial Z_i \partial Z_j \end{bmatrix} \quad (4.2)$$

The Hessian Matrix describes the change in directional force upon small perturbation of the equilibrium structure and thus gives insights on (co-)dependencies of the fluctuations. This follows directly from

truncating the Taylor series expansion of the potential at the equilibrium after the second term. Each $3 \times 3 \mathbf{H}_{ij}$ can be arranged to form the $3N \times 3N$ Hessian Matrix \mathcal{H} for a protein with N residues:

$$\mathcal{H} = \begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & \cdots & \mathbf{H}_{1N} \\ \mathbf{H}_{21} & \ddots & & \vdots \\ \vdots & & & \\ \mathbf{H}_{N1} & \cdots & & \mathbf{H}_{NN} \end{bmatrix} \quad (4.3)$$

Whereas off-diagonal Hessian elements are computed as in Eqn. 4.1, the diagonal elements of \mathcal{H} represent the summation over all off-diagonal 3×3 Hessian submatrices as $\mathbf{H}_{ii} = -\sum_{j|j \neq i}^N \mathbf{H}_{ij}$. Through inversion of the molecular Hamiltonian \mathcal{H} we can obtain the covariance matrix \mathcal{C} . Due to rotational and translation symmetries the covariance matrix cannot be derived directly. Here, spectral decomposition of \mathcal{H} yields six zero-eigenvalues representing aforementioned symmetries and from the remaining sorted $3N - 6$ eigenvalues the mechanical covariance matrix \mathbf{H}^+ can be constructed. Hence, the Moore-Penrose pseudoinverse (Moore, 1920; Penrose and Todd, 1955) of the Hessian can be formulated as $\mathcal{C} = \mathbf{H}^+ = \sum_{i=1}^{N-6} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$.

Within \mathbf{H}^+ covariation of spatial displacement upon perturbation can be obtained. From the Moore-Penrose pseudoinverse we can clearly see that eigenvectors \mathbf{u}_i with small λ_i contribute largely to \mathbf{H}^+ . These *slow* modes represent functional fluctuation of the protein, whereas *fast* modes define structural (compensatory) fluctuation due to their small overall influence on the covariance matrix (Bahar *et al.*, 1998).

4.2.1 Partition Function

In addition to investigating general dynamics based upon the modes of the ENM, we can expand the analysis to investigate the free energy landscape of proteins. For this we need to define the partition function through the configurational integral over all possible microstates of the protein. From the partition function $Z_{\mathcal{H}}$ many thermodynamic properties can be derived, such as the Helmholtz free energy and the Internal Energy (Dill and Bromberg, 2010). For a protein with N residues the partition function reads

$$Z_{\mathcal{H}} = \int d^{3N} \mathbf{r} e^{-\frac{\beta}{2} \Delta \mathbf{r} \mathcal{H} \Delta \mathbf{r}^T}. \quad (4.4)$$

Here, $\Delta \mathbf{r}$ describes the spontaneous spatial displacement \mathbf{r} of the protein in comparison to the equilibrium configuration $\Delta \mathbf{r}$ and, reads $\Delta \mathbf{r} = \mathbf{r} - \mathbf{r}^0$. For our means, this integral can be solved by diagonalization of the Hessian in the molecular Hamiltonian as per $\mathcal{H} = \mathbf{P} \mathbf{\Lambda} \mathbf{P}^T$. With this $Z_{\mathcal{H}}$ can be written as:

$$Z_{\mathcal{H}} = \int d^{3N} \mathbf{r} e^{-\frac{\beta}{2} \Delta \mathbf{r} \mathbf{P} \mathbf{\Lambda} \mathbf{P}^T \Delta \mathbf{r}^T} \quad (4.5)$$

$$= \int d^{3N} \mathbf{u} J(\mathbf{u}) e^{-\frac{\beta}{2} \mathbf{u} \mathbf{\Lambda} \mathbf{u}^T} \quad (4.6)$$

The interested reader will note the transformation of spatial coordinates $\Delta \mathbf{r}$ to the internal coordinates \mathbf{u} introduces the Jacobian $J(\mathbf{u})$. Here, the Jacobian $J(\mathbf{u})$ can be expressed through its determinant $\frac{\partial(\mathbf{r}_1, \dots, \mathbf{r}_{3N})}{\partial(\mathbf{u}_1, \dots, \mathbf{u}_{3N})}$. Since the coordinate transformation only involves rotation with \mathbf{P} , the determinant of the Jacobian equals 1 and can be omitted. Hence, we obtain – after applying basic calculus – the following expression for the partition function:

$$Z_{\mathcal{H}} = \left(\frac{2\pi}{\beta} \right)^{\frac{3N-6}{2}} \left[\prod_{k=1}^{3N-6} \lambda_k^{-1} \right]^{1/2} \quad (4.7)$$

Translational and rotational symmetries require that the smallest six eigenvalues are omitted as the Hessian matrix is singular. In fact, the product $\prod_{k=1}^{3N-6} 1/\lambda_k$ can be understood as the inverse of the pseudo-determinant of \mathcal{H} . Based on this formulation of the partition function we can easily compute the (overall) Helmholtz free energy F as

$$F = -RT \ln Z_{\mathcal{H}} = -\frac{RT}{2} \left[(3N-6) \ln \left(\frac{2\pi}{\beta} \right) - \sum_{k=1}^{3N-6} \ln \lambda_k \right] \quad (4.8)$$

We can clearly see that the Helmholtz free energy depends mainly on the $3N-6$ non-zero eigenvalues λ_k of the Hessian matrix \mathcal{H} . The first term within the brackets in Eqn. 4.8 is a constant, linearly depending on the number of residues of observed system.

4.2.2 Free Energy perturbation

Based on Eqn. 4.8, we can now take a look at perturbation analysis on specific contacts. Through the investigation of free energy perturbation (Zwanzig, 1954; Hamacher, 2011), we can easily discriminate between structurally important and unimportant residue-residue interactions. Naively, the differences in free energy ΔF for any configurations \mathcal{H}_1 and \mathcal{H}_2 can be expressed as:

$$\Delta F = F_2 - F_1 = \frac{RT}{2} \ln \frac{|\mathcal{H}_2|}{|\mathcal{H}_1|} \quad (4.9)$$

$$= \frac{RT}{2} \left[\sum_{j=1}^{3N-6} \ln \lambda_k^{(2)} - \sum_{k=1}^{3N-6} \ln \lambda_k^{(1)} \right] \quad (4.10)$$

In cases where \mathcal{H}_2 can be expressed as $\mathcal{H}_1 + \mathbf{P}$ with \mathbf{P} being a sparse matrix with very few non-zero elements, Eqn. 4.10 can be further simplified by using the matrix determinant lemma (Press *et al.*, 1988) and spectral decomposition (Hamacher, 2011):

$$\Delta F = \frac{RT}{2} \ln \frac{|\mathcal{H}_1 + \mathbf{P}|}{|\mathcal{H}_1|} = \frac{RT}{2} \ln \frac{|\mathcal{H}_1 + \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T|}{|\mathcal{H}_1|} \quad (4.11)$$

$$= \frac{RT}{2} \ln \frac{|\mathcal{H}_1| |\mathbf{\Lambda}| |\mathbf{\Lambda}^{-1} + \mathbf{V}^T \mathcal{H}_1^{-1} \mathbf{U}|}{|\mathcal{H}_1|} \quad (4.12)$$

$$= \frac{RT}{2} \ln(|\Lambda||\Lambda^{-1} + \mathbf{V}^T \mathcal{H}_1^{-1} \mathbf{U}|) \quad (4.13)$$

One can easily see that the computational complexity of ΔF in Eqn. 4.13 is highly dependent on solving $\mathbf{P} = \mathbf{U}\Lambda\mathbf{V}^T$. For single point perturbations, this results in effectively diagonalization of a 6×6 matrix. Consequential, it is possible to assess the influence of individual contacts upon the folding dynamics within a protein structure.

4.2.3 Domain Interaction Perturbation

The comparison of free energy is a straight forward approach for two configurations of the same protein. Similar to the ansatz in the Perturbation Theory removal of a specific interaction within the ANM's Hamiltonian can give us insight into the sensitivity of the system for changes in one very specific interaction. Nonetheless, it is not possible to compare the results from these single contact switch-offs of two different proteins as

- (i.) the free energy F linearly depends on the number of particles in the system,
- (ii.) the total number of contacts can be different, and/or
- (iii.) no correct mapping of contacts is possible as multiple weak could represent few stronger contacts when comparing different structures.

Hence, we propose a domain interaction switch-off that eliminates all contacts between two protein structure elements – e.g. between the S5 and S6 α -helices of channel proteins – and measures the changes in a subspace of the Hessian, the Schur complement (Haynsworth, 1968; Cottle, 1974; Haruna and de Oliveira, 2007).

Let us consider that the Hessian matrix \mathcal{H} is composed of three different block matrices, \mathbf{H}_{ss} , \mathbf{H}_{se} and \mathbf{H}_{ee} . \mathbf{H}_{ss} describes the relation of a *subsystem* with itself and \mathbf{H}_{se} the interaction of said subsystem with the rest of the system (namely, the *environment*). Then, \mathbf{H}_{ee} contains the relations of the environment with itself. Resorting of \mathcal{H} yields:

$$\mathcal{H} = \begin{bmatrix} \mathbf{H}_{ss} & \mathbf{H}_{se} \\ \mathbf{H}_{se}^T & \mathbf{H}_{ee} \end{bmatrix} \quad (4.14)$$

In equilibrium, we can obtain Schur complement of the block \mathbf{H}_{ee} of the matrix \mathcal{H} ($\mathcal{H}/\mathbf{H}_{ee}$) (Haynsworth, 1968; Cottle, 1974; Haruna and de Oliveira, 2007; Eom *et al.*, 2007; Lezon *et al.*, 2009; Ghysels *et al.*, 2010). The second moment of the potential for the subsystem in respect to the dynamics of the environment can be calculated as:

$$\mathcal{H}/\mathbf{H}_{ee} = \mathbf{H}_{ss} - \mathbf{H}_{se}\mathbf{H}_{ee}^{-1}\mathbf{H}_{se}^T \quad (4.15)$$

Essentially, $\mathcal{H}/\mathbf{H}_{ee}$ describes the behavior of the subsystem while still being influenced by the environment (\mathbf{H}_{ee}). Whereas $\mathcal{H}/\mathbf{H}_{ee}$ is a $3N \times 3N$ matrix, $\mathcal{H}/\mathbf{H}_{ee}$ has the same dimensions as the $3M \times 3M$ matrix \mathbf{H}_{ss} , with $M \ll N$.

4.3 Methods

We derived the coarse grained ANM as described by Atilgan *et al.* (2001) (illustrated in Fig. 4.3.A). For this we assumed that the crystallized structures and structures from structure prediction are in equilibrium. No additional molecular dynamics simulations were conducted. To enhance the predictability of the ANM we used distance and amino acid specific force constants proposed by Dehouck and Mikhailov (2013). Conceptually, C α atoms with a distance greater than 16.5 Å are treated as not interacting and receive a $\gamma_{ij} = 0$. The distance and amino acid specific $\gamma(|\mathbf{r}_{ij}^0|, \alpha_i, \alpha_j)$ -function is taken from `bi3d` (Skjærven *et al.*, 2014). All analyses were performed on basis of the so obtained Hessian matrices.

Comparison of Conformational Flexibility

In order to compare the flexibility properties of conformationally different channels, we chose to derive theoretical B-factors from the ANM's Hessian matrix for both conformations (Eyal *et al.*, 2006). Here, we turned our focus upon structures with known open- and closed-state conformation. We derived $\Delta\mathbf{B}_i^{\text{theo}}$ as

$$\Delta\mathbf{B}_i^{\text{theo}} = \mathbf{B}_i^{\text{open}} - \mathbf{B}_i^{\text{closed}} \quad (4.16)$$

$$\text{with } \mathbf{B}_i = \frac{8\pi^2 k_B T}{3} \text{tr}(\mathbf{H}_{ii}^+). \quad (4.17)$$

Here, low B-factors are associated to structural regions with a high degree of order, whereas flexible residues possess a large theoretical B-factor. $\Delta\mathbf{B}_i^{\text{theo}}$ represents the absolute change in flexibility, with negative $\Delta\mathbf{B}_i^{\text{theo}}$ pointing towards a higher flexibility in the closed conformation. A more comprehensive picture of dynamical differences can be achieved by comparing the absolute $\Delta\mathbf{B}_i^{\text{theo}}$ to the relative change in respect to the open structure ($\Delta\mathbf{B}_i^{\text{theo}}/\mathbf{B}_i^{\text{open}}$). As the biological assemblies of all here investigated channels consist of four identical subunits, we used Eq. 4.15 to reduce the dimensions of the Hessian without losing information of interactions between and within the other subunits. In this context, we considered three of the four channel subunits *environmental* and projected their information upon the first subunit.

Interaction Shut-off between Regions

Whereas different conformations of one structure can be compared directly, for example by comparing the B-factors, a fundamentally different approach is necessary for the comparison of two differing structures. While differences in the amino acid sequence are possible, two channel proteins can also vary in their arrangement of secondary structures as well. To compensate this and still be able to investigate the importance of interactions between different structural elements, we developed a perturbation based gedankenexperiment: we "switch off" all interactions between two structural blocks and analyze changes in folding dynamics via the overall free energy differences. Our analysis follows these three steps for each possible regional interaction:

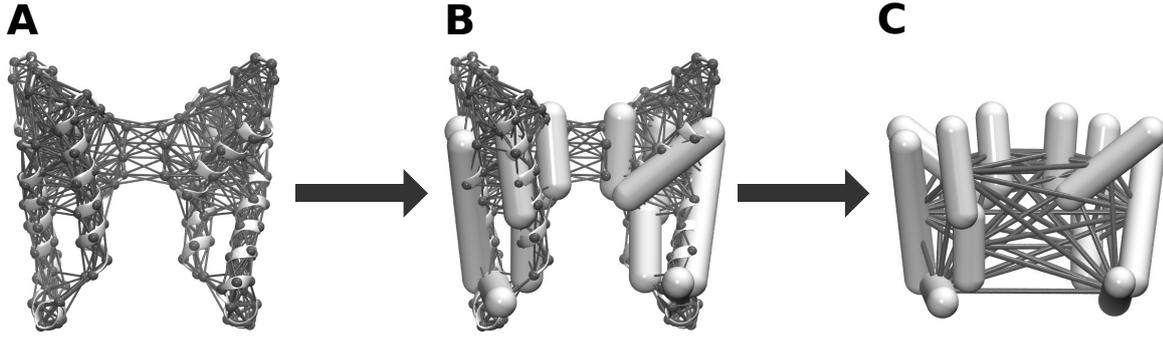


Figure 4.3.: Schematic workflow for the vibrational subsystem analysis of the Schur complement for different ion channels. (A) shows the opposing monomers of the open-state Kcv channel structure with all residue-residue interactions within 10 Å plotted as solid lines. Residues are represented as beads at the C_α coordinates. (B) illustrates the regions found in the pore beginning at the N-terminal s-helix, continuing with the outer helix (TM5) and the p-helix, followed by the filter sequence and the inner helix (TM6) and the C-terminal end (the g-helix). (C) shows the interaction of the six pore regions for two opposing monomers. Note, that, for example, no direct interaction between residues in opposing s-helices exist at 10 Å or 16.5 Å distance. Nevertheless, their interaction in (C) in the Schur complement originates from indirect coupling of the regions.

- (i) Introduction of perturbation in the Hessian matrix \mathcal{H}_1 to obtain \mathcal{H}_2
- (ii) Computation of the Schur complement $\mathcal{H}_2/\mathbf{H}_{ee}^{(2)}$
- (iii) Comparison of the free energy of the unperturbed $\mathcal{H}_1/\mathbf{H}_{ee}^{(1)}$ and the perturbed $\mathcal{H}_2/\mathbf{H}_{ee}^{(2)}$

In (i), we introduce the perturbation in the original coarse-grained model – i.e. in the Hessian matrix \mathcal{H}_1 – by setting all force constants between all C_α between two different regions to zero (see Fig. 4.3.B). Thus, we effectively removed all non-covalent interactions between residues in these blocks. To ensure no backbone breaks are introduced by this procedure, covalent interactions between beads in adjacent blocks remain unaffected. The compensation for differing chain lengths of the analyzed channel proteins is achieved by subdivision of the Hessian matrix as seen in Eq. 4.14 and subsequent Schur complementation (see (ii) and Eq. 4.15). Here, each rigid block is represented by one central residue and all other residues are considered to be environmental residues. These central residues are defined as the residue closest to the center of mass of its structural block. For example, a tetrameric structure with M residues possesses a 72×72 Schur complement $\mathcal{H}/\mathbf{H}_{ee}$ for having six different structural blocks per chain. The dimensions of the Schur complement are only dependent on the number of rigid blocks and independent of residues. Yet, due to the properties of the Schur complement, the global motions are coupled into the 24 particle subsystems (Hafner and Zheng, 2009).

Based on this gedankenexperiment we are able to analyze perturbations of interactions between rigid blocks of amino acids in full sized proteins (exemplarily shown in Fig. 4.3). In the third step, we analyze free energy changes between the Schur complements of the original $\mathcal{H}_1/\mathbf{H}_{ee}^{(1)}$ and the perturbed $\mathcal{H}_2/\mathbf{H}_{ee}^{(2)}$ as per Eq. 4.10. Additionally, we replaced the channel specific amino acid sequence with

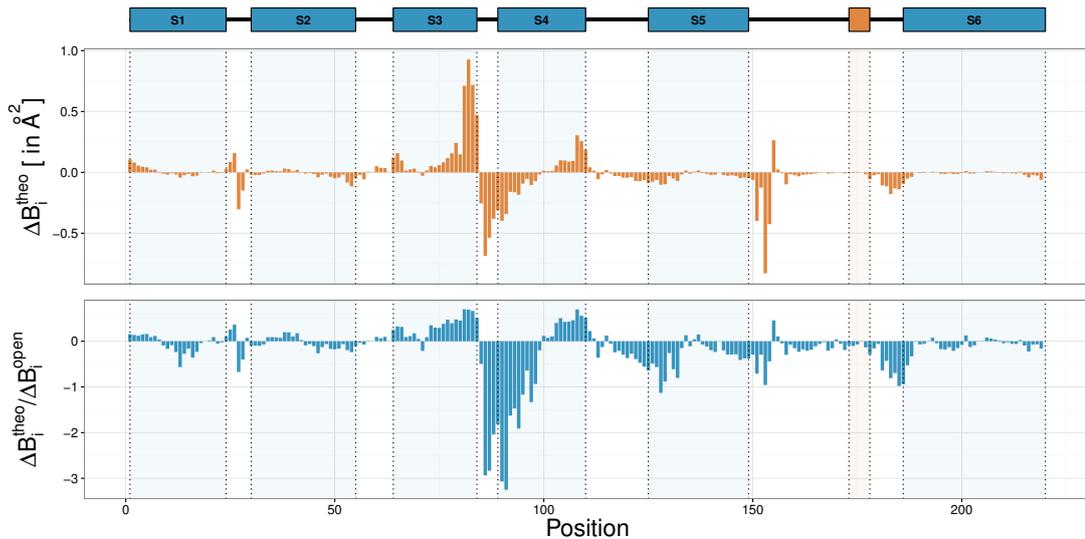


Figure 4.4.: Comparison of theoretical B-factors for the open and closed structures of KAT1. The **upper** plot shows the absolute ΔB_i^{theo} whereas the **lower** plot depicts the relative change in respect to the theoretical B-factors of the open structure derived from elastic network models. Transmembrane helices and the filter sequence are highlighted in blue and orange, respectively, above the plots. Interestingly, upon opening of the channel, the C-terminal ends of S3 and S4 exhibit higher absolute and relative flexibility. In contrast, large portions of S4 as well as S5 experience a strong reduction in flexibility.

a sequence composed of Alanine to investigate the influence of the amino acid sequence on the free energy differences. The "all-Alanine" ANM resembles a single force-constant, distance-dependent parametrization of the ANM.

4.4 Results

4.4.1 Conformational Changes in Channels upon Opening and Closing

For the comparison of dynamics in open and closed structures we considered three different channels: KAT1, MloK1 and Kcv. With these three differently sized potassium channels we can focus on three distinct questions in terms of channel dynamics.

First, we investigated the transition between open and closed structures of KAT1 which were generated through extensive molecular dynamics simulations by Lefoulon *et al.* (2014). In this study, structural modeling was based upon the structures of Kv1.2. Through single site mutations in the second TMD a closed state conformation of the channel was obtained. By comparing both structures we can investigate the consequences of conformational changes in the S1 to S4 segment on the architecture of the central pore (segments S5 and S6).

The second test case focuses on the structural changes in MloK1 upon binding of cAMP and on changes occurring upon opening of the channel pore. While the cAMP in fact does not induce opening of this voltage independent channel, it mediates a higher open propensity of this particular channel (Kowal *et al.*, 2014). Hence, we can examine here how conformational changes in the CNBD affect

distant regions and can increase the propensity for channel opening. Additionally, MloK1 allows us to compare these configurational changes to a closed-state conformation of the channel segments S1 to S6.

In the third case, we investigate the structural changes upon introduction of a protonation/deprotonation of a specific residue in the structure of Kcv (Tayefeh *et al.*, 2009). Here, a homology model based on KirBac1.1 with a deprotonated Lysine at position 29 serves as open-state channel, while the model with a protonated Lysine 29 is considered the closed-state. Both channels underwent exhaustive molecular dynamics simulation to confirm stability of the structures. In Kcv, the short S6 helices prohibit the mechanical closing of the pore through bundle crossing (Tayefeh *et al.*, 2009). From the minimalistic Kcv channel and the comparisons with the other two open/close models we aim at deriving a common functional behavior, independent of the channel size and function.

Flexibility Changes in KAT1 upon Channel Closing

While protein molecular dynamics simulations are not as precise as their experimental counterpart (NMR crystallography), they can give accurate descriptions of dynamical properties of proteins (Lindorff-Larsen *et al.*, 2012). This allows us to investigate changes in structural flexibility of KAT1 based on the structures obtained from short MD simulations (Lefoulon *et al.*, 2014). Contrary to Lefoulon *et al.* (2014), we focus our attention solely on simplified potentials and interactions of the coarse-grained ANM and derive theoretical B-factors from this approximation.

In Fig. 4.4 both the absolute and relative changes of the theoretical B-factors are shown. Here, we find rather large changes in the flexibility of the S3 and S4 as well as portions of S5 segment. Interestingly, no large scale conformational changes can be observed in the second TMD, the region in which the introduced single site mutation converted the open-state channel into closed-state (Lefoulon *et al.*, 2014). In contrast, the filter region, essential for conducting ions through the pore, is unaffected by the opening of the channel. We observe only small changes in flexibility at the C-terminal end of the S6 helix (the g-helix). While this does not rule out the possibility of mechanical gating through bundle crossing, changes in flexibility in such an event appear to be minor for KAT1.

On the one hand, we observe a largely increased flexibility of the C-terminal part of S3 and – to lesser extend – of S4 upon opening of the channel. On the other hand, the N-terminal part of the S4 displays a significant reduction in flexibility. Similarly, the S5 – a central part of the channel pore architecture – shows increased inflexibility upon channel opening depicted by negative $\Delta B_i^{\text{theo}}/B_i^{\text{open}}$ values. These findings suggest movement of the S4 (Lefoulon *et al.*, 2014) as well as drastic changes of the interface residues between S4 and S5, especially when considering that no changes in flexibility can be detected for the rest of the pore module. Interestingly, we can observe reduced flexibility of the S4-S5 linker at opening of the channel suggesting a role of this region in the opening of the channel pore.

Binding of cAMP leads to flexibility reduction in MloK1

Fig. 4.5 shows the changes in structural flexibility upon binding of cAMP in MloK1. Here, we see a predictable reduction in relative flexibility at the C-terminal region of the protein, i.e. the cyclic nucleotide binding domain (CNBD, position ~ 230 to ~ 340). The binding of cAMP into the structure

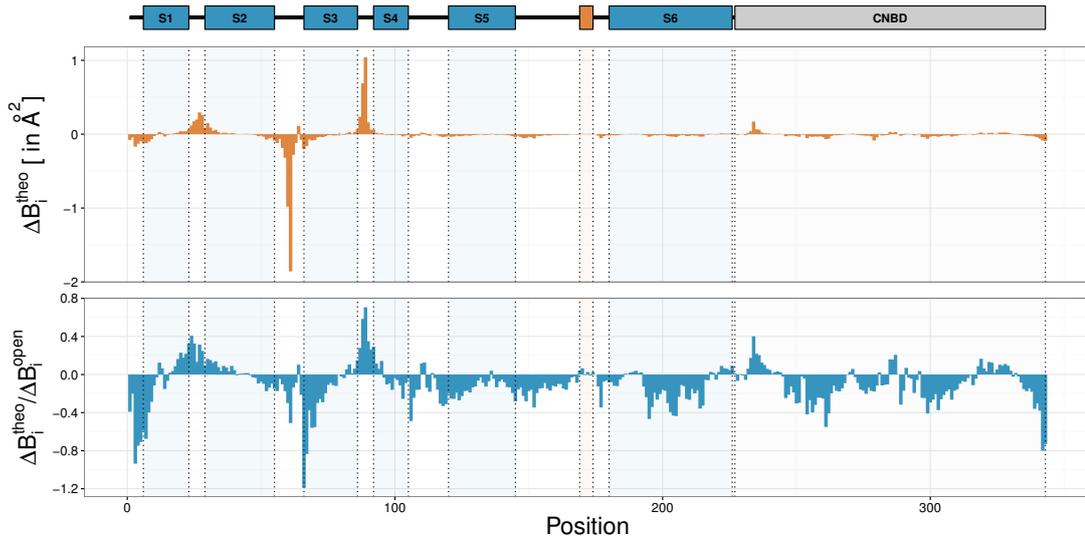


Figure 4.5.: Comparison of theoretical B-factors for the cAMP-bound and unbound structures of MloK1. The **upper** plot shows the absolute ΔB_i^{theo} whereas the **lower** plot depicts the relative change in respect to the theoretical B-factors of the cAMP-bound structure. Transmembrane helices and the filter sequence are highlighted in blue and orange, respectively, above and within the plots. Noteworthy is here that upon cAMP binding to the C-terminal cAMP binding domain the relative flexibility in this area is reduced. A similar flexibility reduction can be observed for the N-terminal half of the S3 segment and a general reduction in flexibility for the entire channel can be observed.

of the CNBD leads to additional conformational constraints through extra non-covalent interactions between cAMP and the protein. Nevertheless, in our coarse-grained model, ions and non-protein components of the structures are omitted. Hence, we can attribute the resulting changes in the flexibility to conformational shifts of the residues in this domain.

Whereas in the rigid, cytoplasmic CNBD only small absolute ΔB^{theo} predominate, we observe comparably large changes in absolute flexibility in the loops of the membrane-bound part of the channel. Here, the loop connecting S2 with S3 exhibits reduction of absolute flexibility, while the loop connecting S1 with S2 and the S3-S4 loop show increased B^{theo} upon binding of cAMP to the CNBD. The transmembrane helices connected by these loops display similar tendencies in relative flexibility, with successively increasing flexibility changes towards the ends of the TMDs. Hence, the binding of cAMP not only affects flexibility and placement of loops but also the structural orientation of the transmembrane helices S1 to S4. Interestingly, upon binding of cAMP to the CNBD the α -helices of the pore region (S5 and S6) also experience reduced flexibility at their C- and N-terminal ends, while the rigidity of the filter region remains almost unchanged. Overall, we can conclude that binding of cAMP to the CNBD affects the flexibility of the entire channel. Almost all structural regions exhibit a relative strong reduction in flexibility, indicating closer packing of the entire protein.

When comparing the closed-state structure of MloK1 to the two open-state configurations, we can observe major increase in flexibility of the open-state in the first two TMDs (see Fig. 4.6). In comparison to the cAMP-bound state of MloK1, we find that the closed configuration experiences higher flexibility at the N-terminal part of the fourth TMD (the voltage-sensor in Kv channels). Contrary to this, the cAMP-bound state shows almost no change in flexibility in this region when compared to the closed structure. Nonetheless, we can clearly see increased flexibility for both open-states in S1,

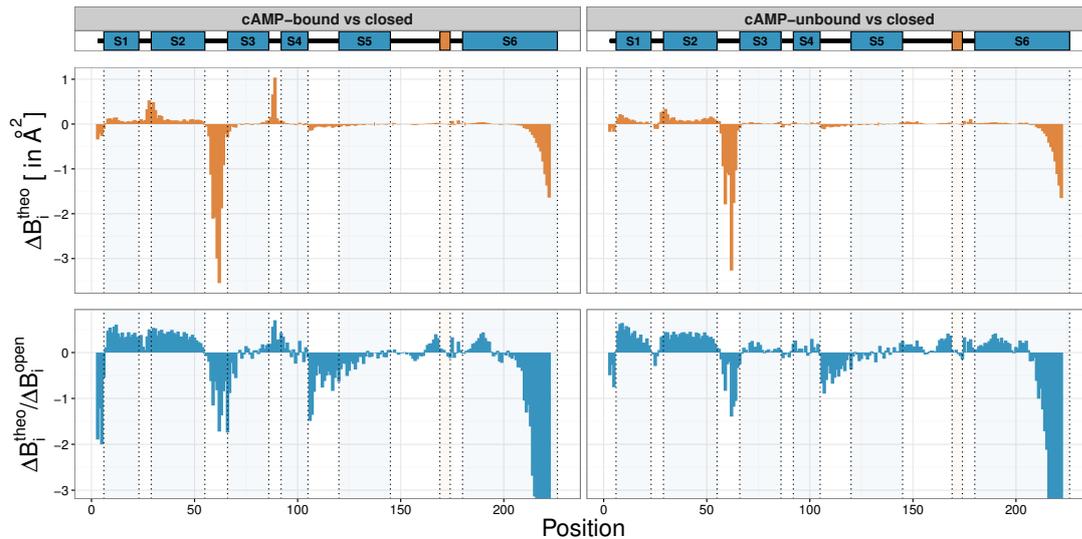


Figure 4.6.: Comparison of theoretical B-factors of the cAMP bound and unbound structures with the B-factors obtained from the closed MloK1 structure. The **upper** row of plot shows the absolute ΔB_i^{theo} whereas the **lower** row shows the relative change in respect to the theoretical B-factors of the open structure. The left column of plots illustrates the comparison of the cAMP-bound-state with a closed-state MloK1 configuration, while the right column of plots shows the comparison of the cAMP-unbound-state with the closed-state. Transmembrane helices and the filter sequence are highlighted in blue and orange, respectively, above the plots. Note the subtle change in the flexibility at the C-terminal end of the S6 segment between the bound and unbound state. Overall, binding of cAMP in the CNBD yields only minor changes in comparison to the closed-state structure.

S2 as well as for residues of inner pore helices (S6). The outer helix of the pore (S5) experiences a reduction in flexibility, a finding similar to the results of KAT1 above. The rather large differences in flexibility at the C-terminus are resulted by the missing CNBD in the closed structure. Due to this, terminal residues which would experience configurational constrictions in presence of the CNBD show higher flexibility. Subsequently, the binding of cAMP can bring residues responsible for the eventual opening of the channel in close proximity. Even though MloK1 is missing a voltage sensing domain in the S4 segment, we could assume that a similar mechanism results in structural changes near to the voltage sensor in, for example, HCN.

KCV Pore Opening Results in Reduction of Flexibility

While the structure of Kcv has yet to be confirmed via crystallography, the structures proposed by Tayefeh *et al.* (2009) based on a homology model of KirBac1.1 revealed signatures of open- and closed-state channel pores. Both structures share the same amino acid sequence and only the protonation state of a single residue is changed. From an ANM perspective, both structures possess identical sequence information and only differ in their spatial orientation.

Based on these structures we can see strong changes in the flexibility of the Kcv residues upon channel opening (see Fig. 4.7). Similarly to the patterns observed in KAT1 (see Fig. 4.4), we find a reduction in flexibility at the C-terminal part of the S5. Furthermore, Kcv experiences rigidification of both turret regions in the open-state when compared to the closed conformation. Interestingly, the largest

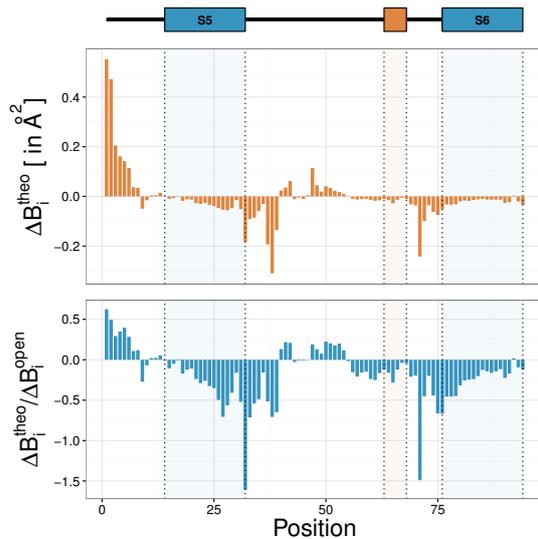


Figure 4.7.: Comparison of theoretical B-factors for the protonated and deprotonated structures of Kcv. The **upper** plot shows the absolute ΔB_i^{theo} whereas the **lower** plot depicts the relative change in respect to the theoretical B-factors of the open structure derived from elastic network models. Transmembrane helices and the filter sequence are highlighted in blue and orange, respectively, above the plots. Interestingly, the highest differences in flexibility can be observed for the N-terminus of channel (the so called slide helix). In general, opening of the channel leads to a wide ranging reduction in flexibility for many residues.

conformational changes in terms of flexibility occur in the N-terminal s-helix. Here, the open-state conformation shows a much higher flexibility for the first eight residues. Afterwards, the residues in the s-helix exhibit reduced flexibility indicating more interaction with the rest of the structure. The importance of these residues for the channel gating has already been shown by Hoffgaard *et al.* (2015) with a different ANM approach. While the S6 segment appears to be too short for bundle-crossing induced gating, we observe reduced flexibility of this segment upon opening of the channel, indicating conformational changes in this region as well.

4.4.2 General Dynamics of the Channel Pore Module

We investigated the similarities of different size channel proteins by a perturbation based approach and subsequent vibrational subsystem analysis. This very specific ansatz allows us to analyze the influence of interactions between entire regions or blocks of a protein while making the results comparable for proteins with corresponding structural order. Nevertheless, we subdivided the available proteins in two classes:

- (a) small, tetrameric channel proteins with only two transmembrane helices per monomer. These proteins possess no voltage sensing domain in form of an S4 and are considered "small" even though they might possess large cytoplasmic domains, and
- (b) large, tetrameric channel proteins with six transmembrane helices. These proteins can possess a voltage sensor in the S4, but are not required to be voltage-dependent channels.

Similarities in all Channels

The relative free energy differences $\Delta F/F_1$ between the perturbed and the unperturbed system are shown on the left hand side of Fig. 4.8. $\Delta F/F_1$ describes the free energy change upon removal of a specific set of interactions in relation to the free energy of the unperturbed configuration. On the right

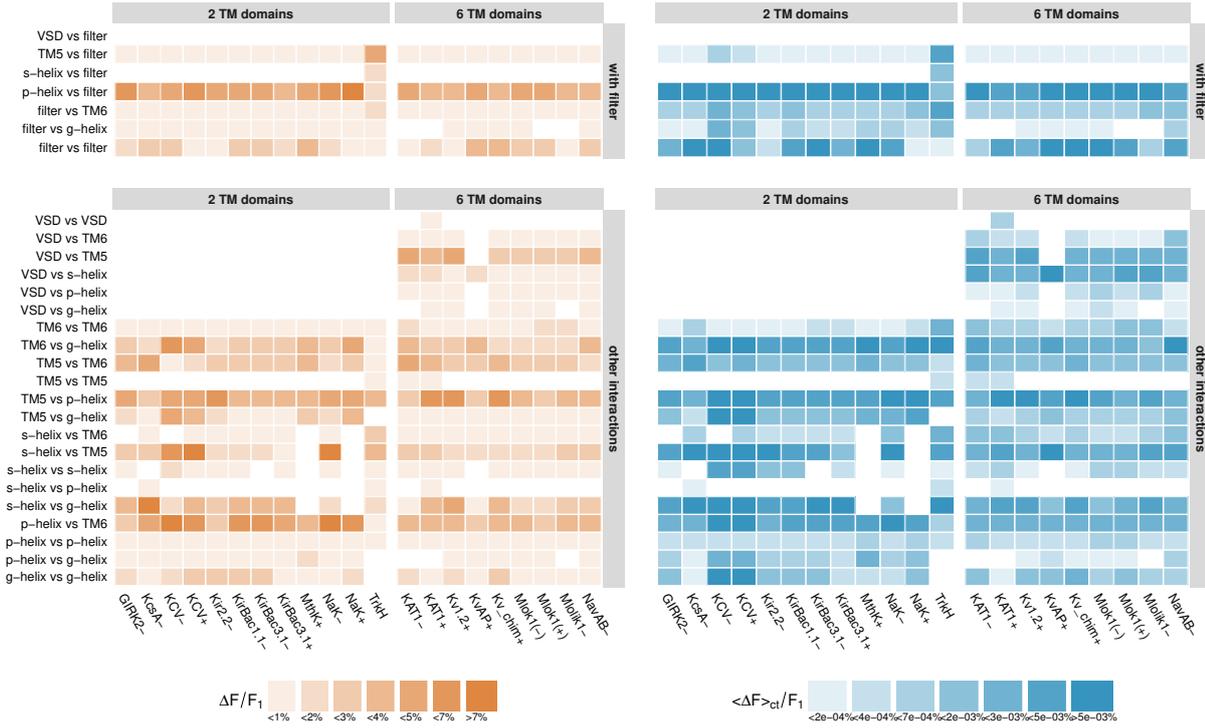


Figure 4.8.: Influence of interaction shut-off between the channel regions for two and six transmembrane domain channels. The results for the sequence specific influence of the shut-off are shown. **(Left)** $\Delta F/F_1$ describes the relative change in free energy upon perturbation of the interaction between the blocks on the left hand side. The filter region is mostly decoupled from the rest of the protein for all tested proteins. **(Right)** $\langle \Delta F \rangle_{ct}/F_1$ describes the average free energy difference between the perturbed and the unperturbed configuration relative to the latter. Again, the filter shares significantly lower average couplings to other structural regions than to itself and the proximal p-helix.

hand side of Fig. 4.8 we show the average, relative free energy difference $\langle \Delta F \rangle / F_1$ per amino acid interaction perturbed (see Eqn. 4.18).

$$\langle \Delta F \rangle_{ct} / F_1 = \frac{1}{K} \cdot \Delta F / F_1 \quad (4.18)$$

$\langle \Delta F \rangle_{ct} / F_1$ provides insight into the composition of single amino acid interactions within the regional interaction as the mean contribution to the free energy difference of one contact per K contacts between the two regions. The combination of $\Delta F/F_1$ and $\langle \Delta F \rangle_{ct} / F_1$ enables us to discriminate between four different interaction groups:

- (i) high $\Delta F/F_1$ ($\geq 4\%$) and high $\langle \Delta F \rangle_{ct} / F_1$ ($\geq 0.002\%$):
Strong interactions between regions coupled with strong average residual interactions between these implies essential interactions for the fold of the protein.
- (ii) high $\Delta F/F_1$ ($\geq 4\%$) and low $\langle \Delta F \rangle_{ct} / F_1$ ($< 0.002\%$):
This combination indicates important structural coupling of two regions through a high number of minor amino acid interactions.

(iii) low $\Delta F/F_1$ ($< 4\%$) and high $\langle \Delta F \rangle_{ct}/F_1$ ($\geq 0.002\%$):

Minor change in free energy upon perturbation combined with (on average) strong inter-residual coupling suggests few strong interactions between the two regions whose perturbation affects the fold marginally.

(iv) low $\Delta F/F_1$ ($< 4\%$) and low $\langle \Delta F \rangle_{ct}/F_1$ ($< 0.002\%$):

While in relative proximity, interactions between the two regions with both low $\Delta F/F_1$ and low $\langle \Delta F \rangle_{ct}/F_1$ do not influence the fold and consist on average of weak residue-residue interactions.

Additionally, we subdivided the results into switch-off experiments involving the filter regions and into experiments involving all other regions.

While the filter sequence acts as the central element for ion channels, Fig. 4.8 suggests that only interactions between the filter and the p-helix have fundamental influence on the folding dynamics of channel proteins and belong to group (i) type interactions. With the exception of the filter to filter interaction (group (iii)), interactions between the filter and all other regions belong to group (iv). This leads us to assume that the amino acids forming the filter are uncoupled from the rest of the protein, as perturbations of interactions with the filter result in only minor changes in the relative free energy.

For all other interactions between structural domains we find two predominant group (i) type interactions. In all channels, regardless of their conformation (i.e. open- or closed-state), interaction between the TM5 and the pore helix as well as interaction between TM6 and the pore helix contribute largely to the overall free energy (high $\Delta F/F_1$) with strong contributing individual interactions (high $\langle \Delta F \rangle_{ct}/F_1$). This indicates a central role of the p-helix, coupling the inner and outer membrane helix, conceptually holding the protein together. The interaction between TM5 and TM6 falls in group (iii) as perturbation experiments switching off all interactions between the inner and outer helices reveal only minor influence on the fold.

Open-state Transitions with in G-helix Reorientation

Transiting from a closed-state Kcv to its open-state we observe changes in the interaction of the slide helix (s-helix) with other regions, most prominent of these changes being at the interaction between the s-helix and the g-helix (transition from group (iii) to group (i)). Here, switch-off experiments in the open-state Kcv draw more severe consequences in terms of free energy. Our results show that opening of Kcv appears to be linked with changes in orientation of the s-helix which eventually influence the interaction of the s-helix. These findings are in conformity to earlier published results (Hoffgaard *et al.*, 2015).

The results for KirBac3.1 conformations are conform to earlier findings (Bavro *et al.*, 2012; Zubcevic *et al.*, 2014) as we can observe reduced interactions between the g-helices in the open-state conformation in comparison to the closed-state for both $\Delta F/F_1$ and $\langle \Delta F \rangle_{ct}/F_1$. We find a similar effect for the comparison of the open-/closed-state conformations of KAT1. On top of the reduced g-helix/g-helix interactions we find, similar to the results of Kcv, an increased influence of the interaction between s- and g-helices on the overall free energy. Interestingly, the s-helix serves different roles in Kcv and KAT1. In KAT1 the s-helix functions as a linker between the VSD and the TM5. Here, opening of the

channel is connected to voltage-sensing, and subsequent movement and reorientation of the VSD. Kcv, on the other hand, is a two TMD channel without a VSD.

Noteworthy is an observably conflicting behavior of the g-helix/g-helix interaction when comparing open- and closed-state conformations. Whereas in Kcv we can see increased importance for the fold of the aforementioned interaction upon opening of the channel, KAT1 and KirBac3.1 exhibit opposing tendencies. These findings indicate difference in the open-/closed-state transitions of these channels with Kcv being unlikely to perform bundle-crossing at the g-helix (Tayefeh *et al.*, 2009; Abenavoli *et al.*, 2009) and both KAT1 and KirBac3.1 experiencing opening via bundle-opening.

Furthermore, the open structures of KAT1, K_v1.2 and the K_v1.2-K_v2.1 chimera present increased influence of the TM5/p-helix interaction, while no such definite trend can be observed for two TMD channels. Both Kcv conformers as well as the closed conformation of the sodium channel NaK present large $\Delta F/F_1$ for the switch-off of all interaction between the s-helix and the TM5.

Structural coupling of the VSD to the Pore through TM5

For all six TMD channels, we find that perturbing the TM5/VSD interaction results in rather large changes in $\Delta F/F_1$ combined with strong residue-residue interactions (group (i)). Perturbing all other possible interactions between the VSD and any other region yields minor implications for the fold of the protein (group (iv)) with the sole exception being interactions between the VSD and the s-helix (group (iii)). The VSD/s-helix interaction presents strong mean residue-residue coupling strengths with marginal effect on the fold. Interestingly, a strong relation through the S4-S5 linker (the s-helix in six transmembrane domain channels) between voltage sensing and the opening and closing of channels has been reported earlier (Chen *et al.*, 2001).

Whereas all other six transmembrane domain channels share the above described feature, KvAP with its unique conformation for voltage-gated channels, behaves differently. Here, the proposed orientation of the voltage sensing domain is horizontal to rather than vertical within the membrane. Contrary to all other channels with a VSD and due to this unique orientation, our results reveal no interaction between the VSD and the TM5 within 16.5 Å. Due to this, we observe different region switch-off patterns in KvAP.

The Functionally Different TrkH pore

Interestingly, TrkH shows very little similarities to all other analyzed potassium and sodium channels. While Cao *et al.* (2011) suggested that the structure of the TrkH transporter protein originated from potassium channel gen duplication and/or fusion, our results point to very different structural dependencies between the pore regions. Whereas all other channels display low relative free-energy changes between TM5 and the filter sequence, TrkH exhibits extraordinary strong dependencies between these regional blocks. Nonetheless, we find for this potassium ion transporter a feature found in all channel proteins (two and six TMD), fold influencing interactions between TM and the pore helix.

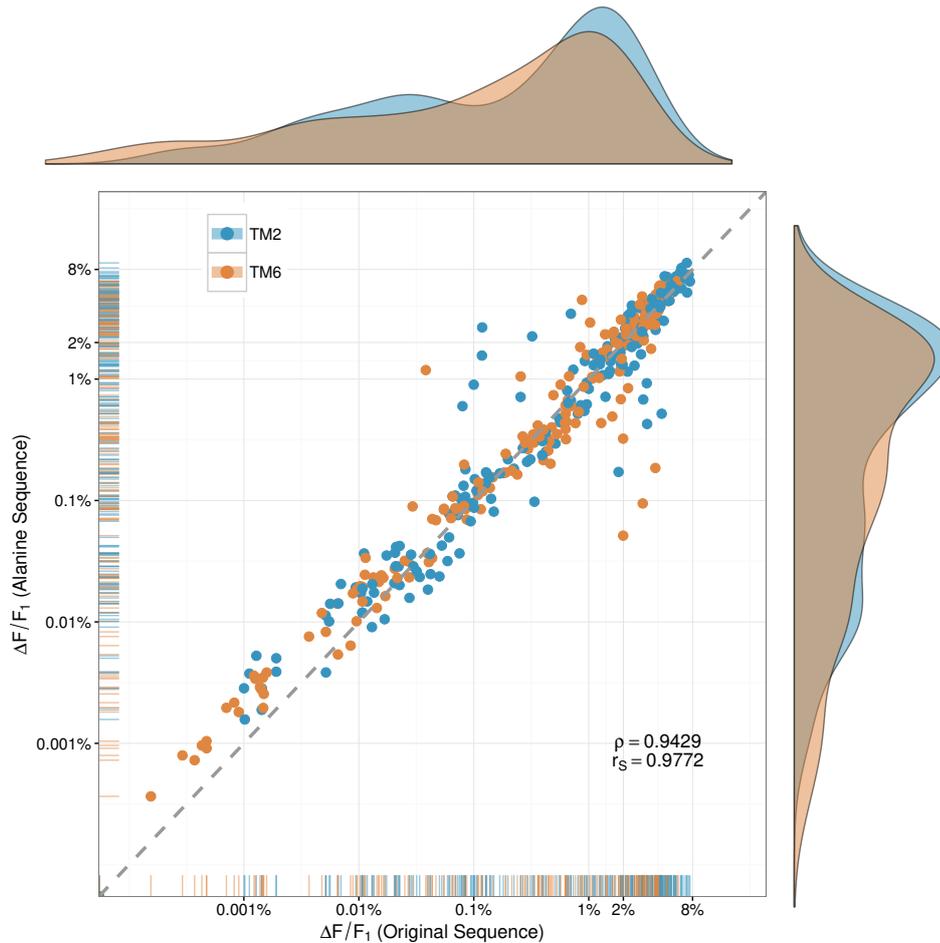


Figure 4.9.: Comparison of the relative changes in free Energy induced through perturbation between the protein specific parameterization and the all Alanine parameterization. In blue, correlations between small, two transmembrane channels are shown and for large six transmembrane domain channels in orange. Based on the scatter and density plots, sequence based parameterization yields similar results to only distance dependent parameterization. The p-values for both correlation coefficients are close to zero at $< 2.2 \cdot 10^{-16}$ and represent a close to zero probability to observe these or higher correlations by chance from the given data.

Channel Behavior Inherited from Spatial Organization

Since our ANM ansatz contained sequence and distance dependent interaction constants, it is conceptually different from the single force constant approach developed by Atilgan *et al.* (2001). By removing all sequence information by setting all residues to Alanine, we can compare the influence of aforementioned sequence information on the thermostability of our model. Fig. 4.9 shows the scatter plot of the region switch-off experiment for two different parameterizations of the ANM (sequence specific vs sequence unspecific parameterization). For both channel type (two TMD and six TMD) we find similar distribution of the results along the diagonal axis representing perfect correlation. Overall, the free energy from the parameterizations show high correlations. Hence, we are lead to assume that our findings are independent of sequence specificity, at least in our coarse-grained ansatz.

4.5 Discussion

We presented a coarse-grained approach to compare dynamics of different proteins with similar structural organization. In order to achieve comparability we followed a two step coarse-graining procedure by constructing an Anisotropic Network Model (Atilgan *et al.*, 2001; Dehouck and Mikhailov, 2013) and projected the interactions on a subspace with a fixed number of particles (Haynsworth, 1968; Cottle, 1974; Haruna and de Oliveira, 2007; Eom *et al.*, 2007; Lezon *et al.*, 2009; Ghysels *et al.*, 2010). The dynamics of the subspace of rigid regions is mostly independent of the amino acid sequence of the protein and appear to be dependent solely on the structure of the protein. While this observation could be attributed to the coarse-graining procedures and the approximation of amino acid interactions through Hooke's spring potentials, our findings could also indicate that the observed (general) regional dependencies are attributed to the fold of the protein and not to its sequence.

When focusing on the differences in flexibility of open- and closed-state structures of the eukaryotic KAT1, the prokaryotic MloK1 and the viral Kcv channels, we find similar tendencies. In all three channels, opening leads to an overall reduction in flexibility of the central pore forming helices (S5 and S6). Additionally, the selectivity filters – presenting only minor flexibility changes – appear to be unaffected by the structural changes occurring in the rest of the pore. This indicates a decoupling of the filter architecture from the rest of the channel. Our findings for the switch-off of regions Gedankenexperiment support the hypothesis of a decoupled selectivity filter. Here, only the interaction between the selectivity filter and the p-helix appears to be important. Removal of other interactions influenced the fold only marginally, depicted by small free energy changes.

Generally, we find that the interactions between the p-helix and the TM5 (a major part of S5) and the TM6 (the major part of segment S6) are predominantly important for all channels analyzed here. Interestingly, the p-helix forms the central part of the channel as switch-off experiments involving p-helix interaction result in major changes in the free energy. One noteworthy exception to this rule is the interaction between the s-helix and the g-helix (an N-terminal segment of S6). We find stronger energetic coupling of the s-helix/g-helix interaction in six TMD channels upon opening of the channel, indicating a pivotal role of the interaction of both regions in the mechanical gating through bundle crossing in six TMD channels. Earlier, Jensen *et al.* (2012) observed upon activation of the VSD of six TMD channels differences in the packing density of the S4-S5-linker. Similarly, we find increased structural importance between the g-helix – mostly a part of the inner helix – and the s-helix upon opening of the channels. These strong dependencies between the s-helix and the g-helix have also been reported elsewhere (Chen *et al.*, 2001; Ding and Horn, 2002, 2003; Nishizawa and Nishizawa, 2009).

Nonetheless, Lörinczi *et al.* (2015) showed a voltage dependent gating without the presence of the s-helix in KCNH, indicating that the S4-S5 linker appears to be unnecessary for transducing information from the VSD to the channel pore, at least for some channels. We show here that the interaction between the VSD and the TM5 region could be responsible for this observation. While we cannot observe a strong direct regional dependency between the VSD and any other structural region in the switch-off experiments, movement of the VSD changes its interaction interface with other parts of the protein. Even though we neglect the possible interaction of the VSD to the segments S1 to S3, our results suggest that a stimulus through VSD movement can be transduced to the central pore region through interaction with the TM5 region. Further transduction to other regions can then be achieved

through the observably strong interaction of TM5 with the p-helix, forming a pathway via the central element of channel proteins.

The gating through bundle crossing has been reported for different channels (Jiang *et al.*, 2002; Thompson *et al.*, 2008; Tao *et al.*, 2009; Payandeh *et al.*, 2011; Bavro *et al.*, 2012). Contrary to this, Tayefeh *et al.* (2009) suggested that the small viral Kcv is not capable to gate through this feature as the channel presents a very short S6 segment. Our results suggest that the C-terminal part of the S6 segment (g-helix) plays a role in the opening/closing of the channel as this region experiences structural changes upon said channel behavior. Both the strength of g-helix/g-helix and g-helix/s-helix interaction (in terms of contribution to the overall free energy) increase upon opening of Kcv, similarly KAT1. Nevertheless, the pattern for both channels appears to be different from that shown by KirBac3.1, a channel reported to experience mechanical gating through bundle crossing (Bavro *et al.*, 2012).

Altogether, we were able to show that – at least on a conceptual level – the process of opening and closing is for the here investigated channels relatively similar. Furthermore, when considering large scale dynamics and folding effects, we showed a marginal influence of the parameterization of elastic network models on free energy based ENM perturbation methodologies.



Conclusion

Throughout this thesis we presented novel approaches to comprehend evolution of proteins within the sequence and structure realm. Since multiple sequence alignments (MSAs) are the foundation for many computational biology procedures, we investigated the influence of substitution matrix choice on homologous sequence search and MSA construction in our contributing work. Eventually, we proposed a corrected form of the popular and often used BLOSUM substitution matrix (Hess *et al.*, 2016) and subsequently derived the structural-alignment based PFASUM matrix (Keul *et al.*, 2017). This matrix consistently outperformed the widely used BLOSUM, VTML or PAM matrices for multiple test cases. Furthermore, the PFASUM algorithm allowed us to develop novel, protein family-specific substitution matrices, which are capable of capturing unique properties of the underlying protein family and in turn increase MSA quality when used in such manner (see Chapter 2).

Based on the information captured in MSAs, we introduced an algorithm to interactively explore mutational patterns and were able to determine evolutionary bottlenecks through this method (Lenz *et al.*, 2014). Furthermore, we expanded the understanding of statistical complexity in relation to sensible background distributions to identify evolutionary active as well as epistatic hot-spots (Keul and Hamacher, 2017). From this, we developed the co-evolutionary complexity using interdependency information from mutual information to gain valuable insights in the complex evolutionary mechanisms in proteins (see Chapter 3).

Within this thesis, we used information theoretical methods in conjunction with elastic network models to characterize evolutionary and structural properties of ion channels. These ion channels are present in all cells and organisms and play essential roles in multiple diseases. Due to their general structural similarity we investigated sequence-structure-function correlates and closely examined evolutionary mechanisms as well as general structural interdependencies. While not all channels possess a voltage sensing domain (VSD), we are able to show that the interaction between the VSD and the S5 helix influences channels profoundly on an evolutionary and structural level. When compared to channels without VSD, we find less intricate evolutionary processes in the S5 segment while we observe a strong functional coupling between VSD and S5. Hence, we can assume that the structural proximity between VSD and S5, combined with the ability of the VSD to move upon changes in voltage, leads to reduced evolutionary promiscuity in S5, conserving specific physico-chemical relationships in this segment.

Channel opening, exemplified by comparison of open and closed structures of Kcv and KAT1, leads to an apparent reduction of flexibility in S5 as well as in the inner helix S6. We find that channel opening reduces the flexibility of S5 and S6, with increased interaction between these two structural elements in neighboring monomers. Similarly, we find increased interaction between the slide helix (s-helix) and the C-terminal part of the inner pore helix (g-helix) upon opening of channels. Hence, we can assume that mechanical gating occurs at these two segments (see Chapter 4).

While the structural interaction between g- and s-helix is prominent in larger (six TMD) and smaller (two TMD) channels, we find unchanged evolutionary dynamics between these two segments when

comparing prokaryotic and eukaryotic channel sequences. Nevertheless, we are able to reveal the existence of more complex evolutionary dependencies within the g-helix in prokaryotic potassium channels. Contrarily, we find higher co-evolutionary complexity near the extracellular portions of the S5 and S6 segments in eukaryotic channels possibly stemming from complex interactions with ligand binding domains (see Chapter 3).

Interestingly, we find that the transition between the g-helix and its structural neighbor, the TM6 segment, is often indicated by a prominent kink, which is often introduced by the presence of glycines. In small potassium channel sequences glycine appears to possess a more pronounced structural and/or functional role as we can observe strong tendencies favoring conservation of glycine in these channels (see Chapter 2). When considering the evolutionary substitution events of potassium channels (from families with six TMD and two TMD) we generally find rather large differences between generalized substitution matrices (such as BLOSUM) and family-specific evolution models (such as PFASUM). Family-specific PFASUM matrices show great promise to improve multiple sequence alignments for homologous sequences. Within the here analyzed PFASUM matrices we find a strong separation of amino acid based on their physico-chemical properties indicating family specific evolution patterns (especially for potassium channels).

Overall, we show that general dynamics of channel proteins within the scope of elastic network models are dominated by their structural organization. The sequence independence of the large scale dynamics is further underlined by the global dissimilarity of the amino acid sequences forming the channel structures. Whereas the amino acid composition can show significant variance in large portions of the channel sequence, the highly hydrophobic filter region is the most conserved region of a channel. We are able to show that perturbations of interaction between the filter region and other parts of the protein results in only minor changes in fold energy. Thus, we are lead to assume that the filter is mostly decoupled from structural changes in other regions. The filter sequence has a unique set of properties: It is anchored in the sequence space through conservation of sequence and function, while being structurally decoupled from the rest of the channel.

Bibliography

- Abenavoli, A., DiFrancesco, M. L., Schroeder, I., Epimashko, S., Gazzarrini, S., Hansen, U. P., Thiel, G. and Moroni, A., 2009, *Fast and slow gating are inherent properties of the pore module of the K⁺ channel Kcv*, *The Journal of General Physiology*, 134(3): 219–229.
- Aittoniemi, J., Fotinou, C., Craig, T. J., de Wet, H., Proks, P. and Ashcroft, F. M., 2009, *SUR1: a unique ATP-binding cassette protein that functions as an ion channel regulator*, *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1514): 257–267.
- Al-Lazikani, B., Jung, J., Xiang, Z. and Honig, B., 2001, *Protein structure prediction*, *Current Opinion in Chemical Biology*, 5(1): 51–56.
- Alam, A. and Jiang, Y., 2009, *High-resolution structure of the open NaK channel*, *Nature structural & molecular biology*, 16(1): 30–34.
- Aldrich, R., Corey, D. and Stevens, C., 1983, *A reinterpretation of mammalian sodium channel gating based on single channel recording*, *Nature*, 306(5942): 436–441.
- Altschul, S. F., 1991, *Amino acid substitution matrices from an information theoretic perspective*, *Journal of Molecular Biology*, 219(3): 555–565.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J., 1990, *Basic local alignment search tool*, *Journal of Molecular Biology*, 215(3): 403–410.
- Altschul, S. F., Wootton, J. C., Gertz, E. M., Agarwala, R., Morgulis, A., Schäffer, A. A. and Yu, Y.-K., 2005, *Protein database searches using compositionally adjusted substitution matrices*, *FEBS Journal*, 272(20): 5101–5109.
- Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S. E., Hubbard, T. J., Chothia, C. and Murzin, A. G., 2008, *Data growth and its impact on the SCOP database: new developments*, *Nucleic Acids Research*, 36(suppl 1): D419–D425.
- Assmann, S. M., 1993, *Signal transduction in guard cells*, *Annual Review of Cell Biology*, 9: 345–375.
- Asti, L., Uguzzoni, G., Marcatili, P. and Pagnani, A., 2016, *Maximum-Entropy Models of Sequenced Immune Repertoires Predict Antigen-Antibody Affinity*, *PLoS Computational Biology*, 12: e1004870.
- Atilgan, A., Durell, S., Jernigan, R., Demirel, M., Keskin, O. and Bahar, I., 2001, *Anisotropy of fluctuation dynamics of proteins with an elastic network model*, *Biophysical Journal*, 80(1): 505–515.
- Bahar, I., Atilgan, A. R. and Erman, B., 1997, *Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential*, *Folding and Design*, 2(3): 173–181.
- Bahar, I., Atilgan, A. R., Demirel, M. C. and Erman, B., 1998, *Vibrational dynamics of folded proteins: significance of slow and fast motions in relation to function and stability*, *Physical Review Letters*, 80(12): 2733.

-
- Bairoch, A., 2004, *Swiss-Prot: Juggling between evolution and stability*, Briefings in Bioinformatics, 5(1): 39–55.
- Baker, O., Larsson, H. P., Mannuzzu, L. and Isacoff, E., 1998, *Three transmembrane conformations and sequence-dependent displacement of the S4 domain in shaker K⁺ channel gating*, Neuron, 20(6): 1283–1294.
- Baldassi, C., Zamparo, M., Feinauer, C., Procaccini, A., Zecchina, R., Weigt, M. and Pagnani, A., 2014, *Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners*, PLoS ONE, 9(3): e92721.
- Ballesteros, J. A., Deupi, X., Olivella, M., Haaksma, E. E. and Pardo, L., 2000, *Serine and threonine residues bend α -helices in the χ 1= g- conformation*, Biophysical Journal, 79(5): 2754–2760.
- Banerjee, A., Lee, A., Campbell, E. and MacKinnon, R., 2013, *Structure of a pore-blocking toxin in complex with a eukaryotic voltage-dependent K⁺ channel*, eLIFE, 2: e00594.
- Barbato, G., Ikura, M., Kay, L. E., Pastor, R. W. and Bax, A., 1992, *Backbone dynamics of calmodulin studied by nitrogen-15 relaxation using inverse detected two-dimensional NMR spectroscopy: the central helix is flexible*, Biochemistry, 31(23): 5269–5278.
- Battiti, R., 1994, *Using mutual information for selecting features in supervised neural net learning*, IEEE Transactions on Neural Networks, 5(4): 537–550.
- Bavro, V. N., De Zorzi, R., Schmidt, M. R., Muniz, J. R., Zubcevic, L., Sansom, M. S., Vénien-Bryan, C. and Tucker, S. J., 2012, *Structure of a KirBac potassium channel with an open bundle crossing indicates a mechanism of channel gating*, Nature Structural & Molecular Biology, 19(2): 158–163.
- Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E., 2000, *The Protein Data Bank*, Nucleic Acids Research, 28(1): 235–242.
- Blackshields, G., Wallace, I. M., Larkin, M. and Higgins, D. G., 2006, *Analysis and comparison of benchmarks for multiple sequence alignment*, In silico biology, 6(4): 321–339.
- Bleicher, L., Lemke, N. and Garratt, R. C., 2011, *Using Amino Acid Correlation and Community Detection Algorithms to Identify Functional Determinants in Protein Families*, PloS ONE, 6(12): e27786.
- Boba, P., Weil, P., Hoffgaard, F. and Hamacher, K., 2010, *Intra- and Inter-Molecular Coevolution: The Case of HIV1 Protease and Reverse Transcriptase*, in *International Joint Conference on Biomedical Engineering Systems and Technologies*, 356–366, Springer.
- Borst, A. and Theunissen, F. E., 1999, *Information theory and neural coding*, Nature Neuroscience, 2(11): 947–957.
- Brenner, S. E., Chothia, C. and Hubbard, T. J., 1998, *Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships*, Proceedings of the National Academy of Sciences, 95(11): 6073–6078.
- Browne, D. L., Gancher, S. T., Nutt, J. G., Brunt, E. R., Smith, E. A., Kramer, P. and Litt, M., 1994, *Episodic ataxia/myokymia syndrome is associated with point mutations in the human potassium channel gene, KCNA1*, Nature Genetics, 8(2): 136–140.

-
- Cahill, N. D., 2010, *Normalized Measures of Mutual Information with General Definitions of Entropy for Multimodal Image Registration*, in *Biomedical Image Registration*, 258–268, Springer Science + Business Media.
- Cai, C., Han, L., Ji, Z. L., Chen, X. and Chen, Y. Z., 2003, *SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence*, *Nucleic acids research*, 31(13): 3692–3697.
- Campos, F. V., Chanda, B., Roux, B. and Bezanilla, F., 2007, *Two atomic constraints unambiguously position the S4 segment relative to S1 and S2 segments in the closed state of Shaker K channel*, *Proceedings of the National Academy of Sciences*, 104(19): 7904–7909.
- Cao, Y., Jin, X., Huang, H., Derebe, M. G., Levin, E. J., Kabaleeswaran, V., Pan, Y., Punta, M., Love, J., Weng, J. *et al.*, 2011, *Crystal structure of a potassium ion transporter, TrkH*, *Nature*, 471(7338): 336–340.
- Chang, M. W., Giffin, M. J., Muller, R., Savage, J., Lin, Y. C., Hong, S., Jin, W., Whitby, L. R., Elder, J. H., Boger, D. L. *et al.*, 2010, *Identification of Broad-Based HIV-1 Protease Inhibitors From Combinatorial Libraries*, *Biochemical Journal*, 429(3): 527–532.
- Chen, C., Natale, D. A., Finn, R. D., Huang, H., Zhang, J., Wu, C. H. and Mazumder, R., 2011, *Representative Proteomes: A Stable, Scalable and Unbiased Proteome Set for Sequence Analysis and Functional Annotation*, *PLoS ONE*, 6(4): e18910.
- Chen, J., Mitcheson, J. S., Tristani-Firouzi, M., Lin, M. and Sanguinetti, M. C., 2001, *The S4–S5 linker couples voltage sensing and activation of pacemaker channels*, *Proceedings of the National Academy of Sciences*, 98(20): 11277–11282.
- Chen, X., Wang, Q., Ni, F. and Ma, J., 2010, *Structure of the full-length Shaker potassium channel Kv1.2 by normal-mode-based X-ray crystallographic refinement*, *Proceedings of the National Academy of Sciences*, 107(25): 11352–11357.
- Cheng, W. W., Enkvetchakul, D. and Nichols, C. G., 2009, *KirBac1.1: It's an Inward Rectifying Potassium Channel*, *The Journal of general physiology*, 133(3): 295–305.
- Claire, L., Aurélien, B., Christine, C., Florence, T., François, T., Pascal, S.-P. and Patricia, T., 2011, *A novel substitution matrix fitted to the compositional bias in Mollicutes improves the prediction of homologous relationships*, *BMC Bioinformatics*, 12(1): 457.
- Clarke, O. B., Caputo, A. T., Hill, A. P., Vandenberg, J. I., Smith, B. J. and Gulbis, J. M., 2010, *Domain Reorientation and Rotation of an Intracellular Assembly Regulate Conduction in Kir Potassium Channels*, *Cell*, 141(6): 1018–1029.
- Clayton, G. M., Altieri, S., Heginbotham, L., Unger, V. M. and Morais-Cabral, J. H., 2008, *Structure of the transmembrane regions of a bacterial cyclic nucleotide-regulated channel*, *Proceedings of the National Academy of Sciences*, 105(5): 1511–1515.
- Cline, M., Hughey, R. and Karplus, K., 2002, *Predicting reliable regions in protein sequence alignments*, *Bioinformatics*, 18(2): 306–314.
- Cottle, R. W., 1974, *Manifestations of the Schur complement*, *Linear Algebra and its Applications*, 8(3): 189–211.

-
- Dayhoff, M. O. and Schwartz, R. M., 1978, *A Model of Evolutionary Change in Proteins*, in *Atlas of Protein Sequence and Structure*, Citeseer.
- Dehouck, Y. and Mikhailov, A. S., 2013, *Effective Harmonic Potentials: Insights into the Internal Cooperativity and Sequence-Specificity of Protein Dynamics*, PLoS Computational Biology.
- Dill, K. and Bromberg, S., 2010, *Molecular Driving Forces: Statistical Thermodynamics in Biology, Chemistry, Physics, and Nanoscience*, Garland Science.
- Ding, S. and Horn, R., 2002, *Tail End of the S6 Segment Role in Permeation in Shaker Potassium Channels*, The Journal of General Physiology, 120(1): 87–97.
- Ding, S. and Horn, R., 2003, *Effect of S6 Tail Mutations on Charge Movement in Shaker Potassium Channels*, Biophysical Journal, 84(1): 295–305.
- Doyle, D. A., Cabral, J. M., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L., Chait, B. T. and MacKinnon, R., 1998, *The Structure of the Potassium Channel: Molecular Basis of K⁺ Conduction and Selectivity*, Science, 280(5360): 69–77.
- Dudík, M., Phillips, S. J. and Schapire, R. E., 2005, *Correcting sample selection bias in maximum entropy density estimation*, in *Advances in Neural Information Processing Systems*, 323–330.
- Dunn, S., Wahl, L. and Gloor, G., 2007, *Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction*, Bioinformatics, 24(3): 333–340.
- Durbin, R., Eddy, S. R., Krogh, A. and Mitchison, G., 1998, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge university press.
- Eddy, S., 1995, *Multiple alignment using hidden Markov models.*, Proceedings International Conference on Intelligent Systems for Molecular Biology, 3: 114–120.
- Eddy, S. R., 1998, *Profile hidden Markov models.*, Bioinformatics, 14(9): 755–763.
- Edgar, R. C., 2004, *MUSCLE: multiple sequence alignment with high accuracy and high throughput*, Nucleic Acids Research, 32(5): 1792–1797.
- Edgar, R. C., 2009, *Optimizing substitution matrix choice and gap parameters for sequence alignment*, BMC Bioinformatics, 10(1): 1.
- Eigen, M., McCaskill, J. and Schuster, P., 1988, *Molecular quasi-species*, The Journal of Physical Chemistry, 92(24): 6881–6891.
- Ekeberg, M., Hartonen, T. and Aurell, E., 2014, *Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences*, Journal of Computational Physics, 276: 341–356.
- Endres, D. M. and Schindelin, J. E., 2003, *A new metric for probability distributions*, IEEE Transactions on Information Theory.
- Eom, K., Baek, S.-C., Ahn, J.-H. and Na, S., 2007, *Coarse-graining of protein structures for the normal mode studies*, Journal of Computational Chemistry, 28(8): 1400–1410.
- Eyal, E., Yang, L.-W. and Bahar, I., 2006, *Anisotropic network model: systematic evaluation and a new web interface*, Bioinformatics, 22(21): 2619–2627.

-
- Feldman, D. P. and Crutchfield, J. P., 1998, *Measures of statistical complexity: Why?*, Physics Letters A, 238(4): 244–252.
- Fernandes, A. D. and Gloor, G. B., 2010, *Mutual information is critically dependent on prior assumptions: would the correct estimate of mutual information please identify itself?*, Bioinformatics, 26(9): 1135–1139.
- Finn, R. D., Tate, J., Mistry, J., Coghill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. L. and Bateman, A., 2008, *The Pfam protein families database*, Nucleic Acids Research, 36(Database issue): D281–D288.
- Finn, R. D., Coghill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A. *et al.*, 2016, *The Pfam protein families database: towards a more sustainable future*, Nucleic Acids Research, 44(D1): D279–D285.
- Frank, H. Y. and Catterall, W. A., 2003, *Overview of the voltage-gated sodium channel family*, Genome Biology, 4(3): 1.
- Frank, H. Y., Yarov-Yarovoy, V., Gutman, G. A. and Catterall, W. A., 2005, *Overview of Molecular Relationships in the Voltage-Gated Ion Channel Superfamily*, Pharmacological Reviews, 57(4): 387–395.
- Fruchterman, T. M. J. and Reingold, E. M., 1991, *Graph Drawing by Force-directed Placement*, Software.
- Georges, A. and Yedidia, J. S., 1991, *How to expand around mean-field theory using high-temperature expansions*, Journal of Physics A: Mathematical and General, 24(9): 2173.
- Ghysels, A., Van Speybroeck, V., Pauwels, E., Catak, S., Brooks, B. R., Van Neck, D. and Waroquier, M., 2010, *Comparative study of various normal mode analysis techniques based on partial Hessians*, Journal of Computational Chemistry, 31(5): 994–1007.
- Gloor, G. B., Martin, L. C., Wahl, L. M. and Dunn, S. D., 2005, *Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions*, Biochemistry, 44(19): 7156–7165.
- Gomes, M., Hamer, R., Reinert, G. and Deane, C. M., 2012, *Mutual information and variants for protein domain-domain contact prediction*, BMC Research Notes, 5(1): 472.
- Gondro, C. and Kinghorn, B., 2007, *A simple genetic algorithm for multiple sequence alignment*, Genetics and Molecular Research, 6(4): 964–982.
- Gouveia-Oliveira, R. and Pedersen, A. G., 2007, *Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation*, Algorithms for Molecular Biology, 2(1): 12.
- Grassberger, P., 1988, *Finite sample corrections to entropy and dimension estimates*, Physics Letters A, 128(6-7): 369–373.
- Green, R. E. and Brenner, S. E., 2002, *Bootstrapping and normalization for enhanced evaluations of pairwise sequence comparison*, Proceedings of the IEEE, 90(12): 1834–1847.

-
- Hafner, J. and Zheng, W., 2009, *Approximate normal mode analysis based on vibrational subsystem analysis with high accuracy and efficiency*, *The Journal of Chemical Physics*, 130(19): 194111.
- Hamacher, K., 2011, *Free energy of contact formation in proteins: Efficient computation in the elastic network approximation*, *Physical Review E*, 84(1): 016703.
- Handel, A., Regoes, R. R. and Antia, R., 2006, *The Role of Compensatory Mutations in the Emergence of Drug Resistance*, *PLoS Computational Biology*, 2(10): e137.
- Hartigan, J. A. and Wong, M. A., 1979, *Algorithm AS 136: A k-means clustering algorithm*, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1): 100–108.
- Hartley, R. V. L., 1928, *Transmission of information*, *The Bell System Technical Journal*, 7(3): 535–563.
- Haruna, L. F. and de Oliveira, M. C., 2007, *Physical properties of the Schur complement of local covariance matrices*, *Journal of Physics A: Mathematical and Theoretical*, 40(47): 14195.
- Haswell, E. S., Phillips, R. and Rees, D. C., 2011, *Mechanosensitive Channels: What Can They Do and How Do They Do It?*, *Structure*, 19(10): 1356–1369.
- Haynsworth, E., 1968, *On the Schur complement*, *Basel Mathematical Notes*.
- Henikoff, S. and Henikoff, J. G., 1991, *Automated assembly of protein blocks for database searching*, *Nucleic Acids Research*, 19(23): 6565–6572.
- Henikoff, S. and Henikoff, J. G., 1992, *Amino acid substitution matrices from protein blocks*, *Proceedings of the National Academy of Sciences*, 89(22): 10915–10919.
- Hensen, C., Hermann, J. C., Nam, K., Ma, S., Gao, J. and Höltje, H.-D., 2004, *A Combined QM/MM Approach to Protein–Ligand Interactions: Polarization Effects of the HIV-1 Protease on Selected High Affinity Inhibitors*, *Journal of Medicinal Chemistry*, 47(27): 6673–6680.
- Hess, M., Keul, F., Goesele, M. and Hamacher, K., 2016, *Addressing inaccuracies in BLOSUM computation improves homology search performance*, *BMC Bioinformatics*, 17(1): 1.
- Hille, B., 2001, *Ion channels of excitable membranes*, *Sinauer*.
- Hoffgaard, F., Weil, P. and Hamacher, K., 2010, *BioPhysConnectoR: Connecting sequence information and biophysical models*, *BMC Bioinformatics*, 11(1): 1.
- Hoffgaard, F., Kast, S., Moroni, A., Thiel, G. and Hamacher, K., 2015, *Tectonics of a K⁺ channel: The importance of the N-terminus for channel gating*, *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1848(12): 3197–3204.
- Holsinger, L. J., Nichani, D., Pinto, L. H. and Lamb, R. A., 1994, *Influenza A virus M2 ion channel protein: a structure-function analysis*, *Journal of Virology*, 68(3): 1551–1563.
- Ishima, R. and Torchia, D. A., 2000, *Protein dynamics from NMR*, *Nature Structural & Molecular Biology*, 7(9): 740–743.
- Jarymowycz, V. A. and Stone, M. J., 2006, *Fast Time Scale Dynamics of Protein Backbones: NMR Relaxation Methods, Applications, and Functional Consequences*, *Chemical Reviews*, 106(5): 1624–1671.

-
- Jensen, J. L. W. V., 1906, *Sur les fonctions convexes et les inégalités entre les valeurs moyennes*, Acta Mathematica, 30(1): 175–193.
- Jensen, M. Ø., Jogini, V., Borhani, D. W., Leffler, A. E., Dror, R. O. and Shaw, D. E., 2012, *Mechanism of voltage gating in potassium channels*, Science, 336(6078): 229–233.
- Jiang, Y., Lee, A., Chen, J., Cadene, M., Chait, B. T. and MacKinnon, R., 2002, *The open pore conformation of potassium channels*, Nature, 417(6888): 523–526.
- Jiang, Y., Lee, A., Chen, J., Ruta, V., Cadene, M., Chait, B. T. and MacKinnon, R., 2003, *X-ray structure of a voltage-dependent K⁺ channel*, Nature, 423(6935): 33–41.
- Jogini, V. and Roux, B., 2005, *Electrostatics of the intracellular vestibule of K⁺ channels*, Journal of Molecular Biology, 354(2): 272–288.
- Jones, D. T., Buchan, D. W., Cozzetto, D. and Pontil, M., 2012, *PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments*, Bioinformatics, 28(2): 184–190.
- Jones, D. T., Singh, T., Kosciolk, T. and Tetchner, S., 2015, *MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins*, Bioinformatics, 31(7): 999–1006.
- Kann, M., Qian, B. and Goldstein, R. A., 2000, *Optimization of a new score function for the detection of remote homologs*, Proteins: Structure, Function, and Bioinformatics, 41(4): 498–503.
- Kantz, H., Kurths, J. and Mayer-Kress, G. (Editors), 1998, *Nonlinear Analysis of Physiological Data*, Springer.
- Katoh, K., Misawa, K., Kuma, K.-i. and Miyata, T., 2002, *MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform*, Nucleic Acids Research, 30(14): 3059–3066.
- Keul, F. and Hamacher, K., 2017, *Consistent Quantification of Complex Dynamics via a Novel Statistical Complexity Measure*, Physical Review E (in submission).
- Keul, F., Hess, M., Goesele, M. and Hamacher, K., 2017, *PFASUM: A substitution matrix from Pfam structural alignments*, Bioinformatics (in review).
- Kowal, J., Chami, M., Baumgartner, P., Arbeit, M., Chiu, P.-L., Rangl, M., Scheuring, S., Schröder, G. F., Nimigean, C. M. and Stahlberg, H., 2014, *Ligand-induced structural changes in the cyclic nucleotide-modulated potassium channel MloK1*, Nature Communications, 5.
- Kullback, S. and Leibler, R. A., 1951, *On Information and Sufficiency*, The Annals of Mathematical Statistics, 22(1): 79–86.
- Kunze, J., Todoroff, N., Schneider, P., Rodrigues, T., Geppert, T., Reisen, F., Schreuder, H., Saas, J., Hessler, G., Baringhaus, K.-H. et al., 2014, *Targeting Dynamic Pockets of HIV-1 Protease by Structure-Based Computational Screening for Allosteric Inhibitors*, Journal of Chemical Information and Modeling, 54(3): 987–991.
- Kuo, A., Gulbis, J. M., Antcliff, J. F., Rahman, T., Lowe, E. D., Zimmer, J., Cuthbertson, J., Ashcroft, F. M., Ezaki, T. and Doyle, D. A., 2003, *Crystal structure of the potassium channel KirBac1.1 in the closed state*, Science, 300(5627): 1922–1926.

-
- Kwak, J. M., Murata, Y., Baizabal-Aguirre, V. M., Merrill, J., Wang, M., Kemper, A., Hawke, S. D., Tallman, G. and Schroeder, J. I., 2001, *Dominant negative guard cell K^+ channel mutants reduce inward-rectifying K^+ currents and light-induced stomatal opening in Arabidopsis*, *Plant Physiology*, 127(2): 473–485.
- Lance, G. N. and Williams, W. T., 1967, *A General Theory of Classificatory Sorting Strategies: I. Hierarchical Systems*, *The Computer Journal*, 9(4): 373–380.
- Lee, S.-Y., Lee, A., Chen, J. and MacKinnon, R., 2005, *Structure of the KvAP voltage-dependent K^+ channel and its dependence on the lipid membrane*, *Proceedings of the National Academy of Sciences*, 102(43): 15441–15446.
- Leelananda, S. P., Kloczkowski, A. and Jernigan, R. L., 2016, *Fold-specific sequence scoring improves protein sequence matching*, *BMC Bioinformatics*, 17(1): 328.
- Lefoulon, C., Karnik, R., Honsbein, A., Gutla, P. V., Grefen, C., Riedelsberger, J., Poblete, T., Dreyer, I., Gonzalez, W. and Blatt, M. R., 2014, *Voltage-sensor transitions of the inward-rectifying K^+ channel KAT1 indicate a latching mechanism biased by hydration within the voltage sensor*, *Plant Physiology*, 166(2): 960–975.
- Lemaitre, C., Barré, A., Citti, C., Tardy, F., Thiaucourt, F., Sirand-Pugnet, P. and Thébaud, P., 2011, *A novel substitution matrix fitted to the compositional bias in Mollicutes improves the prediction of homologous relationships*, *BMC Bioinformatics*, 12(1): 1.
- Lenz, O., Keul, F., Bremm, S., Hamacher, K. and von Landesberger, T., 2014, *Visual analysis of patterns in multiple amino acid mutation graphs*, in *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, 93–102, IEEE.
- Levitt, M. and Greer, J., 1977, *Automatic identification of secondary structure in globular proteins*, *Journal of Molecular Biology*, 114(2): 181–239.
- Lezon, T. R., Srivastava, I., Zheng, Y. and Bahar, I., 2009, *Elastic network models for biomolecular dynamics: theory and application to membrane proteins and viruses*, *Handbook on Biological Networks*, 129–58.
- Li, Q., Wanderling, S., Paduch, M., Medovoy, D., Singharoy, A., McGreevy, R., Villalba-Galea, C. A., Hulse, R. E., Roux, B., Schulten, K. *et al.*, 2014, *Structural mechanism of voltage-dependent gating in an isolated voltage-sensing domain*, *Nature Structural & Molecular Biology*, 21(3): 244–252.
- Li, W., McWilliam, H., Goujon, M., Cowley, A., Lopez, R. and Pearson, W. R., 2012, *PSI-Search: iterative HOE-reduced profile SSEARCH searching*, *Bioinformatics*, 28(12): 1650–1651.
- Liao, L. and Noble, W. S., 2003, *Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships*, *Journal of Computational Biology*, 10(6): 857–868.
- Lin, J., 1991, *Divergence measures based on the Shannon entropy*, *IEEE Transactions on Information Theory*, 37(1): 145–151.
- Lindahl, E. R., 2008, *Molecular dynamics simulations*, *Molecular Modeling of Proteins*, 3–23.

-
- Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M. P., Dror, R. O. and Shaw, D. E., 2012, *Systematic Validation of Protein Force Fields against Experimental Data*, PloS ONE, 7(2): e32131.
- Liu, H., Elstner, M., Kaxiras, E., Frauenheim, T., Hermans, J. and Yang, W., 2001, *Quantum mechanics simulation of protein dynamics on long timescale*, Proteins: Structure, Function, and Bioinformatics, 44(4): 484–489.
- Liu, H., Sun, J., Liu, L. and Zhang, H., 2009, *Feature selection with dynamic mutual information*, Pattern Recognition, 42(7): 1330–1339.
- Lockless, S. W. and Ranganathan, R., 1999, *Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families*, Science, 286(5438): 295–299.
- Long, S. B., Campbell, E. B. and MacKinnon, R., 2005, *Voltage Sensor of Kv1.2: Structural Basis of Electromechanical Coupling*, Science, 309(5736): 903–908.
- Long, S. B., Tao, X., Campbell, E. B. and MacKinnon, R., 2007, *Atomic structure of a voltage-dependent K^+ channel in a lipid membrane-like environment*, Nature, 450(7168): 376–382.
- Lopez-Ruiz, R., Mancini, H. and Calbet, X., 1995, *A Statistical Measure of Complexity*, Physics Letters A.
- Lörinczi, É., Gómez-Posada, J. C., de La Peña, P., Tomczak, A. P., Fernández-Trillo, J., Leipscher, U., Stühmer, W., Barros, F. and Pardo, L. A., 2015, *Voltage-dependent gating of KCNH potassium channels lacking a covalent link between voltage-sensing and pore domains*, Nature Communications, 6.
- MacQueen, J. *et al.*, 1967, *Some methods for classification and analysis of multivariate observations*, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 281–297, Oakland, CA, USA.
- Maes, F., Collignon, A., Vandermeulen, D., Marchal, G. and Suetens, P., 1997, *Multimodality Image Registration by Maximization of Mutual Information*, IEEE Transactions on Medical Imaging, 16(2): 187–198.
- Mannuzzu, L. M., Moronne, M. M. and Isacoff, E. Y., 1996, *Direct Physical Measure of Conformational Rearrangement Underlying Potassium Channel Gating*, Science, 271(5246): 213.
- Mao, W., Kaya, C., Dutta, A., Horovitz, A. and Bahar, I., 2015, *Comparative study of the effectiveness and limitations of current methods for detecting sequence coevolution*, Bioinformatics, 31(12): 1929–1937.
- Martell, M., Esteban, J. I., Quer, J., Genesca, J., Weiner, A., Esteban, R., Guardia, J. and Gomez, J., 1992, *Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution.*, Journal of Virology, 66(5): 3225–3229.
- Martin, M., Plastino, A. and Rosso, O., 2006, *Generalized statistical complexity measures: Geometrical and analytical properties*, Physica A: Statistical Mechanics and its Applications, 369(2): 439–462.
- Matthews, E. and Hanna, M. G., 2014, *Cav1.1 Channel and Hypokalemic Periodic Paralysis*, in *Pathologies of Calcium Channels*, 135–149, Springer.

-
- Matthews, E., Labrum, R., Sweeney, M., Sud, R., Haworth, A., Chinnery, P., Meola, G., Schorge, S., Kullmann, D., Davis, M. and Hanna, M., 2009, *Voltage sensor charge loss accounts for most cases of hypokalemic periodic paralysis*, *Neurology*, 72(18): 1544–1547.
- Metzner, K. J., Giulieri, S. G., Knoepfel, S. A., Rauch, P., Burgisser, P., Yerly, S., Gunthard, H. F. and Cavassini, M., 2009, *Minority quasispecies of drug-resistant HIV-1 that lead to early therapy failure in treatment-naïve and-adherent patients*, *Clinical Infectious Diseases*, 48(2): 239–247.
- Miller, C., 2000, *An overview of the potassium channel family*, *Genome Biology*, 1(4): 1.
- Mohabatkar, H., Beigi, M. M. and Esmaeili, A., 2011, *Prediction of GABA A receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine*, *Journal of Theoretical Biology*, 281(1): 18–23.
- Moore, E. H., 1920, *On the reciprocal of the general algebraic matrix*, *Bulletin of the American Mathematical Society*, 26(9): 385–397.
- Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T. and Weigt, M., 2011, *Direct-coupling analysis of residue coevolution captures native contacts across many protein families*, *Proceedings of the National Academy of Sciences*, 108(49): E1293–E1301.
- Mould, J. A., Drury, J. E., Frings, S. M., Kaupp, U. B., Pekosz, A., Lamb, R. A. and Pinto, L. H., 2000, *Permeation and activation of the M2 ion channel of influenza A virus*, *Journal of Biological Chemistry*, 275(40): 31038–31050.
- Müller, T. and Vingron, M., 2000, *Modeling amino acid replacement*, *Journal of Computational Biology*, 7(6): 761–776.
- Nagel, G., Szellas, T., Huhn, W., Kateriya, S., Adeishvili, N., Berthold, P., Ollig, D., Hegemann, P. and Bamberg, E., 2003, *Channelrhodopsin-2, a directly light-gated cation-selective membrane channel*, *Proceedings of the National Academy of Sciences*, 100: 13940–13945.
- Neupärtl, M., Meyer, C., Woll, I., Frohns, F., Kang, M., Van Etten, J. L., Kramer, D., Hertel, B., Moroni, A. and Thiel, G., 2008, *Chlorella viruses evoke a rapid release of K⁺ from host cells during the early phase of infection*, *Virology*, 372(2): 340–348.
- Ng, P. C., Henikoff, J. G. and Henikoff, S., 2000, *PHAT: a transmembrane-specific substitution matrix*, *Bioinformatics*, 16(9): 760–766.
- Nishizawa, M. and Nishizawa, K., 2009, *Coupling of S4 Helix Translocation and S6 Gating Analyzed by Molecular-Dynamics Simulations of Mutated Kv Channels*, *Biophysical Journal*, 97(1): 90–100.
- Notredame, C. and Higgins, D. G., 1996, *SAGA: sequence alignment by genetic algorithm*, *Nucleic Acids Research*, 24(8): 1515–1524.
- Notredame, C., Higgins, D. G. and Heringa, J., 2000, *T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment*, *Journal of Molecular Biology*, 302(1): 205–217.
- O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufu, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D. *et al.*, 2015, *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation*, *Nucleic Acids Research*, gkv1189.

-
- O'Sullivan, B. P. and Freedman, S. D., 2009, *Cystic fibrosis*, *The Lancet*, 373: 1891–1904.
- Pace, C. N. and Scholtz, J. M., 1998, *A Helix Propensity Scale Based on Experimental Studies of Peptides and Proteins*, *Biophysical Journal*, 75(1): 422–427.
- Pais, F. S.-M., de Cássia Ruy, P., Oliveira, G. and Coimbra, R. S., 2014, *Assessing the efficiency of multiple sequence alignment programs*, *Algorithms for Molecular Biology*, 9(1): 4.
- Pandey, S., Zhang, W. and Assmann, S. M., 2007, *Roles of ion channels and transporters in guard cell signal transduction*, *FEBS Letters*, 581(12): 2325–2336.
- Pathak, M. M., Yarov-Yarovoy, V., Agarwal, G., Roux, B., Barth, P., Kohout, S., Tombola, F. and Isacoff, E. Y., 2007, *Closing in on the resting state of the Shaker K⁺ channel*, *Neuron*, 56(1): 124–140.
- Payandeh, J., Scheuer, T., Zheng, N. and Catterall, W. A., 2011, *The crystal structure of a voltage-gated sodium channel*, *Nature*, 475(7356): 353–358.
- Pazos, F. and Valencia, A., 2001, *Similarity of phylogenetic trees as indicator of protein–protein interaction*, *Protein Engineering*, 14(9): 609–614.
- Pazos, F., Helmer-Citterich, M., Ausiello, G. and Valencia, A., 1997, *Correlated mutations contain information about protein-protein interaction*, *Journal of Molecular Biology*, 271(4): 511–523.
- Pearson, W. R., 1991, *Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms*, *Genomics*, 11(3): 635–650.
- Pearson, W. R. and Lipman, D. J., 1988, *Improved tools for biological sequence comparison*, *Proceedings of the National Academy of Sciences*, 85(8): 2444–2448.
- Penrose, R. and Todd, J. A., 1955, *A generalized inverse for matrices*, *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(03): 406.
- Pierri, C. L., Parisi, G. and Porcelli, V., 2010, *Computational approaches for protein function prediction: a combined strategy from multiple sequence alignment to molecular docking-based virtual screening*, *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, 1804(9): 1695–1712.
- Plugge, B., Gazzarrini, S., Nelson, M., Cerana, R., Van, J. L., Derst, C., DiFrancesco, D., Moroni, A., Thiel, G. *et al.*, 2000, *A Potassium Channel Protein Encoded by Chlorella Virus PBCV-1*, *Science*, 287(5458): 1641–1644.
- Posson, D. J., McCoy, J. G. and Nimigeon, C. M., 2013, *The voltage-dependent gate in MthK potassium channels is located at the selectivity filter*, *Nature Structural & Molecular Biology*, 20(2): 159–166.
- Press, W. H., Teukolsky, S., Vetterling, W. and Flannery, B., 1988, *Numerical Recipes in C*, Cambridge University Press, 1: 3.
- Price, G. A., Crooks, G. E., Green, R. E. and Brenner, S. E., 2005, *Statistical evaluation of pairwise protein sequence comparison with the Bayesian bootstrap*, *Bioinformatics*, 21(20): 3824–3831.
- R Core Team, 2015, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

-
- Rhee, S.-Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J. and Shafer, R. W., 2003, *Human immunodeficiency virus reverse transcriptase and protease sequence database*, *Nucleic Acids Research*, 31(1): 298–303.
- Rueckert, D., Hayes, C., Studholme, C., Summers, P., Leach, M. and Hawkes, D. J., 1998, *Non-rigid registration of breast MR images using mutual information*, in *Medical Image Computing and Computer-Assisted Intervention — MICCAI'98*, 1144–1152, Springer Nature.
- Ruta, V., Jiang, Y., Lee, A., Chen, J. and MacKinnon, R., 2003, *Functional analysis of an archaeobacterial voltage-dependent K^+ channel*, *Nature*, 422(6928): 180–185.
- Sanchez, J. A., Dani, J. A., Siemen, D. and Hille, B., 1986, *Slow permeation of organic cations in acetylcholine receptor channels.*, *The Journal of General Physiology*, 87(6): 985–1001.
- Schrödinger, LLC, 2015, *The PyMOL Molecular Graphics System, Version 1.8*.
- Schubert, U., Ferrer-Montiel, A. V., Oblatt-Montal, M., Henklein, P., Strebel, K. and Montal, M., 1996, *Identification of an ion channel activity of the Vpu transmembrane domain and its involvement in the regulation of virus release from HIV-1-infected cells*, *FEBS Letters*, 398(1): 12–18.
- Shackelford, G. and Karplus, K., 2007, *Contact prediction using mutual information and neural nets*, *Proteins*, 69(S8): 159–164.
- Shafer, R. W., Dupnik, K., Winters, M. A. and Eshleman, S. H., 2001, *A Guide to HIV-1 Reverse Transcriptase and Protease Sequencing for Drug Resistance Studies*, *HIV Sequence Compendium*, 2001: 1–51.
- Shannon, C. E., 1948, *A Mathematical Theory of Communication*, *The Bell System Technical Journal*, 27.
- Shannon, C. E., 1949, *Communication Theory of Secrecy Systems*, *Bell Labs Technical Journal*, 28(4): 656–715.
- Shi, N., Ye, S., Alam, A., Chen, L. and Jiang, Y., 2006, *Atomic structure of a Na^+ - and K^+ -conducting channel*, *Nature*, 440(7083): 570–574.
- Shrivastava, I. H., Durell, S. R. and Guy, H. R., 2004, *A Model of Voltage Gating Developed Using the KvAP Channel Crystal Structure*, *Biophysical Journal*, 87(4): 2255–2270.
- Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J. et al., 2011, *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega*, *Molecular Systems Biology*, 7(1): 539.
- Skjærven, L., Yao, X.-Q., Scarabelli, G. and Grant, B. J., 2014, *Integrating protein structural dynamics and evolutionary analysis with Bio3D*, *BMC Bioinformatics*, 15(1): 399.
- Smith, T. F. and Waterman, M. S., 1981, *Identification of common molecular subsequences*, *Journal of Molecular Biology*, 147(1): 195–197.
- Song, D., Chen, J., Chen, G., Li, N., Li, J., Fan, J., Bu, D. and Li, S. C., 2015, *Parameterized blosum matrices for protein alignment*, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(3): 686–694.

-
- Sonnhammer, E. L., Eddy, S. R., Durbin, R. *et al.*, 1997, *Pfam: A comprehensive database of protein domain families based on seed alignments*, *Proteins: Structure Function and Genetics*, 28(3): 405–420.
- Steinbrecher, T. and Elstner, M., 2013, *QM and QM/MM simulations of proteins*, *Biomolecular Simulations: Methods and Protocols*, 91–124.
- Stinson, D., 2005, *Cryptography*, Taylor & Francis Ltd.
- Strehl, A. and Ghosh, J., 2002, *Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions*, *Journal of Machine Learning Research*, 3(Dec): 583–617.
- Studholme, C., Hill, D. and Hawkes, D., 1999, *An overlap invariant entropy measure of 3D medical image alignment*, *Pattern Recognition*, 32(1): 71–86.
- Styczynski, M. P., Jensen, K. L., Rigoutsos, I. and Stephanopoulos, G., 2008, *BLOSUM62 miscalculations improve search performance*, *Nature Biotechnology*, 26(3): 274–275.
- Suri, S. and Reinartz, P., 2010, *Mutual-Information-Based Registration of TerraSAR-X and Ikonos Imagery in Urban Areas*, *IEEE Transactions on Geoscience and Remote Sensing*, 48(2): 939–949.
- Szczepanski, J., Arnold, M., Wajnryb, E., Amigó, J. M. and Sanchez-Vives, M. V., 2011, *Mutual information and redundancy in spontaneous communication between cortical neurons*, *Biological Cybernetics*, 104(3): 161–174.
- Tao, X., Avalos, J. L., Chen, J. and MacKinnon, R., 2009, *Crystal structure of the eukaryotic strong inward-rectifier K^+ channel Kir2.2 at 3.1 Å resolution*, *Science*, 326(5960): 1668–1674.
- Tao, X., Lee, A., Limapichat, W., Dougherty, D. A. and MacKinnon, R., 2010, *A Gating Charge Transfer Center in Voltage Sensors*, *Science*, 328(5974): 67–73.
- Tayefeh, S., Kloss, T., Kreim, M., Gebhardt, M., Baumeister, D., Hertel, B., Richter, C., Schwalbe, H., Moroni, A., Thiel, G. *et al.*, 2009, *Model development for the viral Kcv potassium channel*, *Biophysical Journal*, 96(2): 485–498.
- Thompson, A. N., Posson, D. J., Parsa, P. V. and Nimigeon, C. M., 2008, *Molecular mechanism of pH sensing in KcsA potassium channels*, *Proceedings of the National Academy of Sciences*, 105(19): 6900–6905.
- Thompson, J. D., Higgins, D. G. and Gibson, T. J., 1994, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice*, *Nucleic Acids Research*, 22(22): 4673–4680.
- Thompson, J. D., Koehl, P., Ripp, R. and Poch, O., 2005, *BALiBASE 3.0: latest developments of the multiple sequence alignment benchmark*, *Proteins: Structure, Function, and Bioinformatics*, 61(1): 127–136.
- Thompson, J. D., Linard, B., Lecompte, O. and Poch, O., 2011, *A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives*, *PloS ONE*, 6(3): e18093.
- Tirion, M. M., 1996, *Large Amplitude Elastic Motions in Proteins from a Single-Parameter Atomic Analysis*, *Physical Review Letters*, 77(9): 1905.

-
- Vilim, R., Cunningham, R., Lu, B., Kheradpour, P. and Stevens, F. J., 2004, *Fold-specific substitution matrices for protein classification*, *Bioinformatics*, 20(6): 847–853.
- Vinh, N. X., Epps, J. and Bailey, J., 2010, *Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance*, *Journal of Machine Learning Research*.
- Volkov, A. G., Adesina, T., Markin, V. S. and Jovanov, E., 2008, *Kinetics and Mechanism of *Dionaea muscipula* Trap Closing*, *Plant Physiology*, 146: 694–702.
- Vora, T., Bisset, D. and Chung, S.-H., 2008, *Conduction of Na^+ and K^+ through the NaK Channel: Molecular and Brownian Dynamics Studies*, *Biophysical Journal*, 95(4): 1600–1611.
- Wall, M. E., Rechtsteiner, A. and Rocha, L. M., , *Singular Value Decomposition and Principal Component Analysis*, *A Practical Approach to Microarray Data Analysis*, 91–109.
- Wang, L. and Jiang, T., 1994, *On the Complexity of Multiple Sequence Alignment*, *Journal of Computational Biology*, 1(4): 337–348.
- Weil, P., Hoffgaard, F. and Hamacher, K., 2009, *Estimating sufficient statistics in co-evolutionary analysis by mutual information*, *Computational Biology and Chemistry*, 33(6): 440–444.
- Whorton, M. R. and MacKinnon, R., 2011, *Crystal structure of the mammalian GIRK2 K^+ channel and gating regulation by G proteins, PIP2, and sodium*, *Cell*, 147(1): 199–208.
- Yang, L.-W., Liu, X., Jursa, C. J., Holliman, M., Rader, A., Karimi, H. A. and Bahar, I., 2005, *iGNM: a database of protein functional motions based on Gaussian Network Model*, *Bioinformatics*, 21(13): 2978–2987.
- Yang, N., George, A. L. and Horn, R., 1996, *Molecular Basis of Charge Movement in Voltage-Gated Sodium Channels*, *Neuron*, 16(1): 113–122.
- Yusaf, S. P., Wray, D. and Sivaprasadarao, A., 1996, *Measurement of the movement of the S4 segment during the activation of a voltage-gated potassium channel*, *Pflügers Archiv*, 433(1-2): 91–97.
- Zaki, M. J., 2015, *Data Mining and Analysis*, Cambridge University Press.
- Zhang, Y. and Skolnick, J., 2005, *TM-align: a protein structure alignment algorithm based on the TM-score*, *Nucleic Acids Research*, 33: 2302–2309.
- Zimmer, J., Doyle, D. A. and Grossmann, J. G., 2006, *Structural characterization and pH-induced conformational transition of full-length KcsA*, *Biophysical Journal*, 90(5): 1752–1766.
- Ziv, J. and Lempel, A., 1977, *A Universal Algorithm for Sequential Data Compression*, *IEEE Transactions on Information Theory*, 23(3): 337–343.
- Zubcevic, L., Bavro, V. N., Muniz, J. R., Schmidt, M. R., Wang, S., De Zorzi, R., Venien-Bryan, C., Sansom, M. S., Nichols, C. G. and Tucker, S. J., 2014, *Control of KirBac3.1 Potassium Channel Gating at the Interface between Cytoplasmic Domains*, *Journal of Biological Chemistry*, 289(1): 143–151.
- Zwanzig, R. W., 1954, *High-temperature equation of state by a perturbation method. I. nonpolar gases*, *The Journal of Chemical Physics*, 22(8): 1420–1426.

Appendices



A Acronyms

List of abbreviations

ANM	anisotropic network model
APC	average product corrected mutual information
BLOSUM	substitution matrix based on the BLOCKS database
cAMP	cyclic adenosine monophosphate
CNBD	cyclic-nucleotide binding domain
CTD	cytoplasmic domain
ΔC	statistical complexity difference
ΔC_{MI}	co-evolutionary complexity difference
DCA	direct coupling analysis
D_{JS}	Jensen-Shannon divergence
D_{KL}	Kullback-Leibler divergence
DNA	deoxyribonucleic acid
ENM	elastic network model
GNM	Gaussian network models
g-helix	C-terminal gate helix
HIV	human immunodeficiency virus
HIVP	protease of the human immunodeficiency virus
HMM	hidden Markov model
MD	molecular dynamics
MI	mutual information
MI_{2P}	joint entropy normalized mutual information
MI_{arith}	arithmetical mean normalized mutual information
MI_{min}	minimum entropy normalized mutual information
MSA	multiple sequence alignment
NMI	joint entropy normalize mutual information
NMR	nuclear magnetic resonance spectrography
p-helix	pore helix
PAM	(substitution matrix based on) point accepted mutations
PF00520	ion transporter protein family
PF07885	ion channel family
Pfam	protein family database

PFASUM	substitution matrix based on Pfam
Q-score	PREFAB quality score
QM/MM	combined quantum mechanics and molecular dynamics simulations
RCW	row-column weight normalized mutual information
RP	representative proteome
RPG	representative proteome group
s-helix	slide helix
S5	the outer channel pore forming helix
S6	the inner channel pore forming helix
SVD	singular value decomposition
TC score	BAlIbASE total column score
TMD	transmembrane domain
TM5	C-terminal portion of the outer channel pore forming helix
TM6	N-terminal portion of the inner channel pore forming helix
VSD	voltage sensing domain
Z _{APC}	Z-score statistics of average product corrected mutual information
Z _{RCW}	Z-score statistics of row-column weight normalized mutual information

List of amino acid abbreviations

A	alanine
B	asparagine or aspartic acid
C	cysteine
D	aspartic acid
E	glutamic acid
F	phenylalanine
G	glycine
H	histidine
I	isoleucine
J	leucine or isoleucine
K	lysine
L	leucine
M	methionine
N	asparagine
P	proline
R	arginine
S	serine
T	threonine
V	valine
W	tryptophan
X	all
Y	tyrosine
Z	glutamic acid or glutamine



B Channel specific substitution matrices

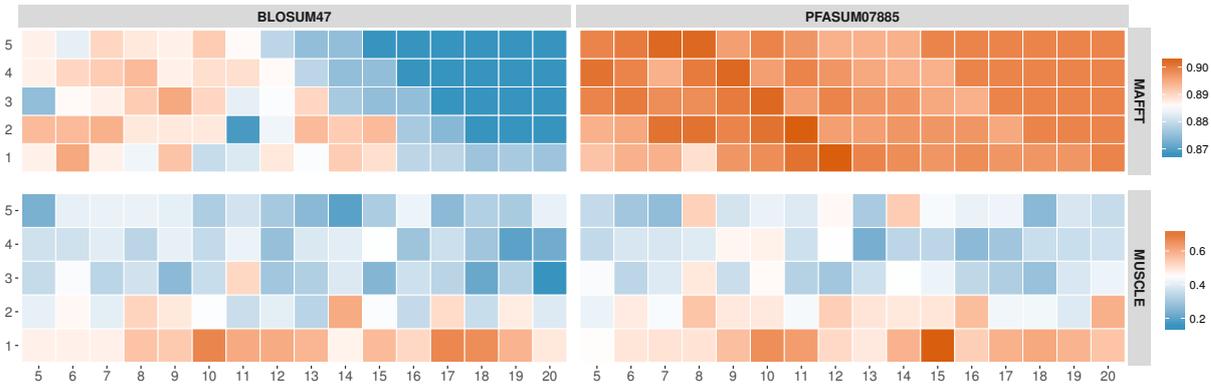


Figure B.1.: Performance comparison through the Cline shift score of the BLOSUM47 and the family-specific PFASUM07885 substitution matrices on the RP15 dataset of small ion channel proteins (PF07885).

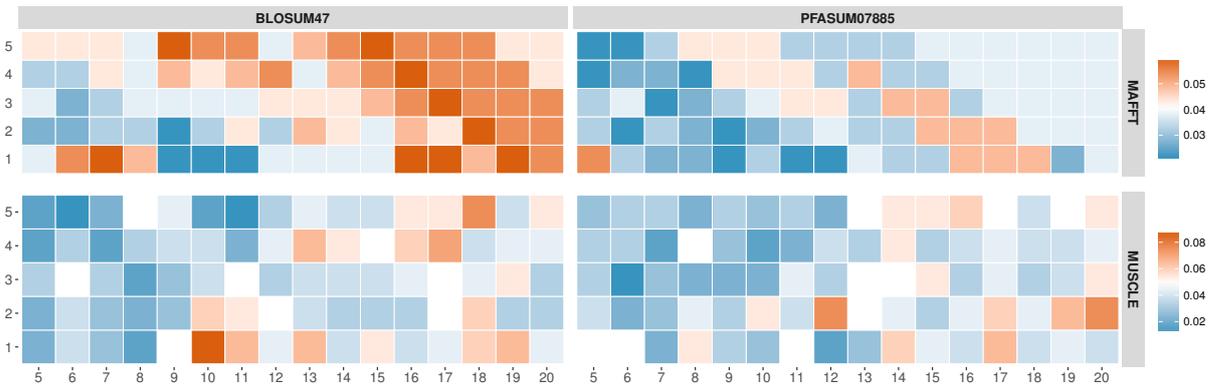


Figure B.2.: Performance comparison through the total column score of the BLOSUM47 and the family-specific PFASUM07885 substitution matrices on the RP15 dataset of small ion channel proteins (PF07885).

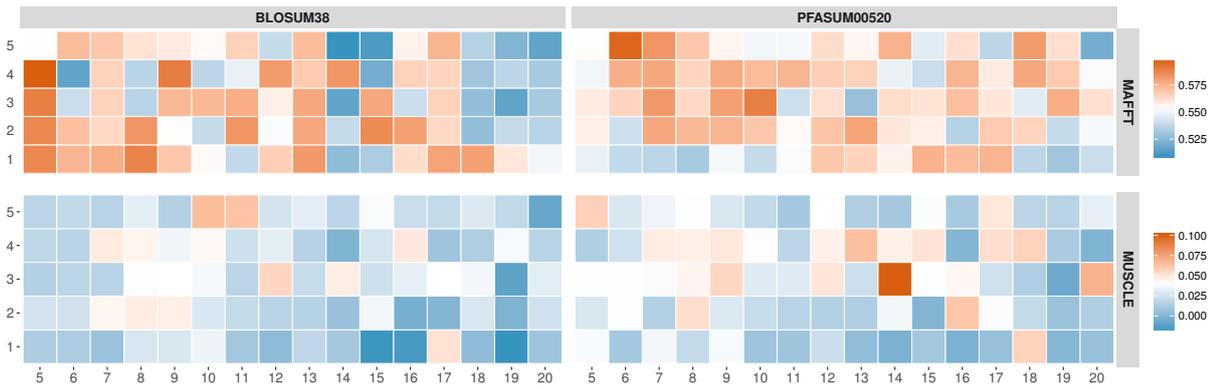


Figure B.3.: Performance comparison through the Cline shift score of the BLOSUM38 and the family-specific PFASUM00520 substitution matrices on the RP15 dataset of small ion channel proteins (PF00520).

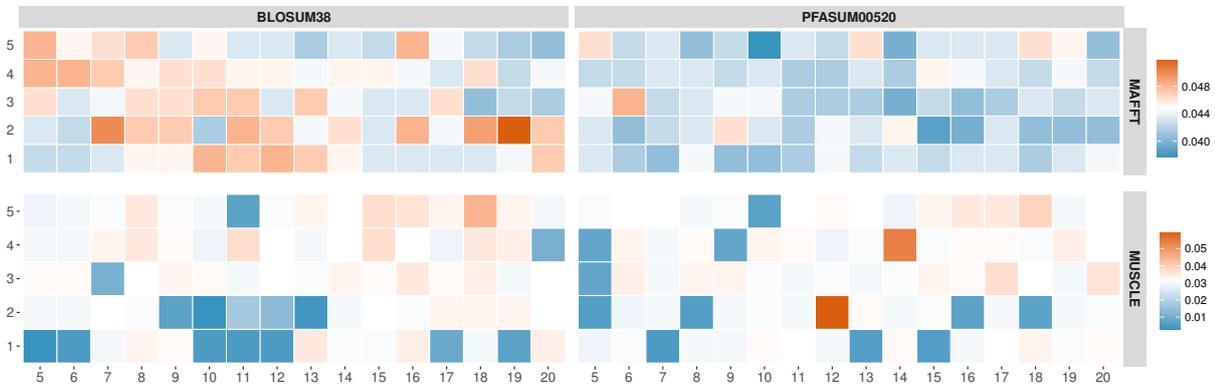


Figure B.4.: Performance comparison through the total column score of the BLOSUM38 and the family-specific PFASUM00520 substitution matrices on the RP15 dataset of small ion channel proteins (PF00520).

C Evolutionary Information in Sequence Data

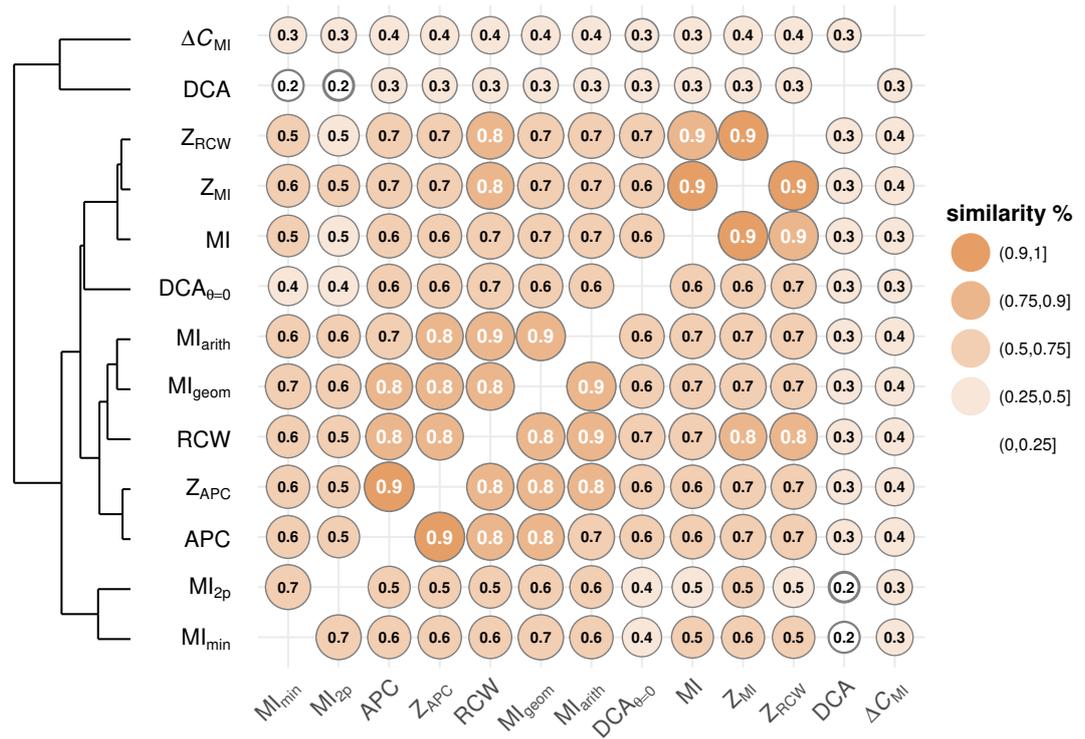


Figure C.1.: Similarity matrix of the here tested information theoretic measures based on the top 10% of values for all HIVP sequences of treated patients. Noteworthy is that DCA and ΔC_{MI} appear to be significantly different from all other methods, as these share a maximum of 40% similarity with any other method.



Acknowledgement

Here, I would like to mention all those who have supported me during my time in the *Computational Biology and Simulation* working group and well beyond. Without these people, this thesis would have not been possible.

I thank **Prof. Dr. Kay Hamacher** and **Prof. Dr. Gerhard Thiel** for sponsoring me, asking challenging questions and always having an open door. I thank you for motivating, inspiring and supporting my work as well as encouraging creative approaches to solving problems.

I thank **Patrick, Sabine** and **Steffi** for insightful discussions, creative diversions and an overall pleasant working environment. You guys helped me focus on important things and were right there during (c/b)runch time.

I thank **Sven, Christine, Michael** and **Sebastian** for working and bearing with me on various projects as well as creating a supportive environment. Your unique perspective on things made me look at them differently.

I thank **Martin** for his productive persistence, patience and fun times working on our projects. I already miss the uncountable discussion on multiple sequence alignments, correct phylogenetic relations and the meaning of gaps.

I thank **Philipp** for all the great advice, hour-long conversations and baby photo sharing sessions.

I thank all students I supervised during practice courses for their enthusiasm to learn. Foremost I thank **Julius** and **Basti** for listening to my advice, the diverse academic discussions and their friendship beyond academia.

I thank **Gisela, Caroline** and our Admins **Martin** and **Andreas** for keeping the machine running. Without you doing all the necessary small and big administrative tasks, academic life would have been much harder.

I thank **my family, my wife** and **my son, my parents** and **parents-in-law** for continuous support, great enthusiasm and patients. Without your backing this thesis would not have been possible.





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.

Balingen, 12. April 2017

Frank Keul