Systems Biomedicine

ISSN: 2162-8130 (Print) 2162-8149 (Online) Journal homepage: www.tandfonline.com/journals/kssb20

Learning diagnostic signatures from microarray data using L1-regularized logistic regression

Preetam Nandy, Michael Unger, Christoph Zechner, Kushal K Dey & Heinz Koeppl

To cite this article: Preetam Nandy, Michael Unger, Christoph Zechner, Kushal K Dey & Heinz Koeppl (2013) Learning diagnostic signatures from microarray data using L1-regularized logistic regression, Systems Biomedicine, 1:4, 240-246, DOI: 10.4161/sysb.25271

To link to this article: https://doi.org/10.4161/sysb.25271

RESEARCH PAPER

Systems Biomedicine 1:4, 240–246; October/November/December 2013; © 2013 Landes Bioscience

Learning diagnostic signatures from microarray data using L1-regularized logistic regression

Preetam Nandy,[1] Michael Unger,[1] Christoph Zechner,[1] Kushal K Dey,[2] Heinz Koeppl[1,*]

[1]BISON Group; Automatic Control Laboratory; ETH Zurich; Zurich, Switzerland; [2]Indian Statistical Institute; Kolkata, India

Keywords: classification, gene expression, L1-regularization, LASSO, logistic regression, microarray data, RMA normalization, Wilcoxon rank sum test

Abbreviations: AC, adenocarcinoma; AUPR, area under Precision-Recall curve; AUPR_Avg, average of the area under Precision-Recall curve across the classes; BCM, belief confusion metric; CCEM, Correct Class Enrichment Metric; COPD, chronic obstructive pulmonary disease; IMPROVER, Industrial Methodology for Process Verification in Research; LASSO, least absolute shrinkage and selection operator; LC, lung cancer; MSD, multiple sclerosis diagnostic; RMA, robust multi-array average; SCC, squamous cell carcinoma

Making reliable diagnoses and predictions based on high-throughput transcriptional data has attracted immense attention in the past few years. While experimental gene profiling techniques—such as microarray platforms—are advancing rapidly, there is an increasing demand of computational methods being able to efficiently handle such data.

In this work we propose a computational workflow for extracting diagnostic gene signatures from high-throughput transcriptional profiling data. In particular, our research was performed within the scope of the first IMPROVER challenge. The goal of that challenge was to extract and verify diagnostic signatures based on microarray gene expression data in four different disease areas: psoriasis, multiple sclerosis, chronic obstructive pulmonary disease and lung cancer. Each of the different disease areas is handled using the same three-stage algorithm. First, the data are normalized based on a multi-array average (RMA) normalization procedure to account for variability among different samples and data sets. Due to the vast dimensionality of the profiling data, we subsequently perform a feature pre-selection using a Wilcoxon's rank sum statistic. The remaining features are then used to train an L1-regularized logistic regression model which acts as our primary classifier. Using the four different data sets, we analyze the proposed method and demonstrate its use in extracting diagnostic signatures from microarray gene expression data.

Introduction

The effective treatment of diseases often relies on making early and accurate diagnoses. However, this can be highly challenging, especially for diseases with complex genetic causes. Microarray techniques are able to capture the expression levels of thousands of genes, opening up a huge source of information about the genetic profiles of patients. While the potential of microarray technologies for medical purposes was repeatedly demonstrated,[1-3] challenges arise in the computational handling of such data sets. Typically, approaches from statistics and machine learning[4,5] are employed to extract disease-relevant information and to predict diagnostic features such as a patient's disease state. Most of these approaches are *supervised*, meaning that they rely on the availability of labeled training data.[6] Common techniques include linear discriminant analysis, nearest-neighbor classifiers, classification trees, bagging, and boosting,[7] support-vector machines,[8,9] neural networks,[10] hierarchical Bayesian models[11] and regularized regressions.[12,13]

Typically, the number of case and control samples is just a fraction of the number of probes on a single microarray chip, posing one of the main difficulties in handling such data. Mathematically, the corresponding inverse problems are said to be *ill-posed* or *underdetermined* and their solution requires specialized algorithms.

The same situation applies for the data from the first IMPROVER (Industrial Methodology for Process Verification in Research) challenge,[14,15] the Diagnostic Signature Challenge. Only a few hundred training samples were provided for each of the four disease data sets, psoriasis, multiple sclerosis diagnostic (MSD), chronic obstructive pulmonary disease (COPD) and lung cancer (LC), in order to train the classifiers. Based on those, the goal was to predict the disease-probabilities of additional samples from an unlabeled test data set.

In this work we lay out a computational workflow, which accounts for the complex nature of the high-dimensional microarray data sets. The validity of the approach is benchmarked using four independent data sets within the scope of the

*Correspondence to: Heinz Koeppl; Email: koepplh@ethz.ch
Submitted: 01/10/2013; Accepted: 06/03/2013; Published Online: 08/30/2013
http://dx.doi.org/10.4161/sysb.25271

Table1. Number of genes selected by the pre-selection algorithm that correspond to each of the sub-challenges

Sub-challenge	Psoriasis	MSD	COPD	LC (2 classes)	LC (AC stage)	LC (SCC stage)
# genes selected	15502	9591	2000* (1152)	3260	2000* (3)	2000* (1012)

MSD, multiple sclerosis diagnostic; COPD, chronic obstructive pulmonary disease; LC, lung cancer. For the psoriasis and MSD sub-challenges, a large number of genes with significant *P*-value scores were selected. For COPD, LC (AC) and LC (SCC), because the number of selected genes was low they were replaced by 2000 of the most significant genes in terms of their *P*-values. The numbers in brackets are the number of variables (the genes) with *P*-value less than 0.1/(total number of genes) in the Wilcoxon rank sum test.

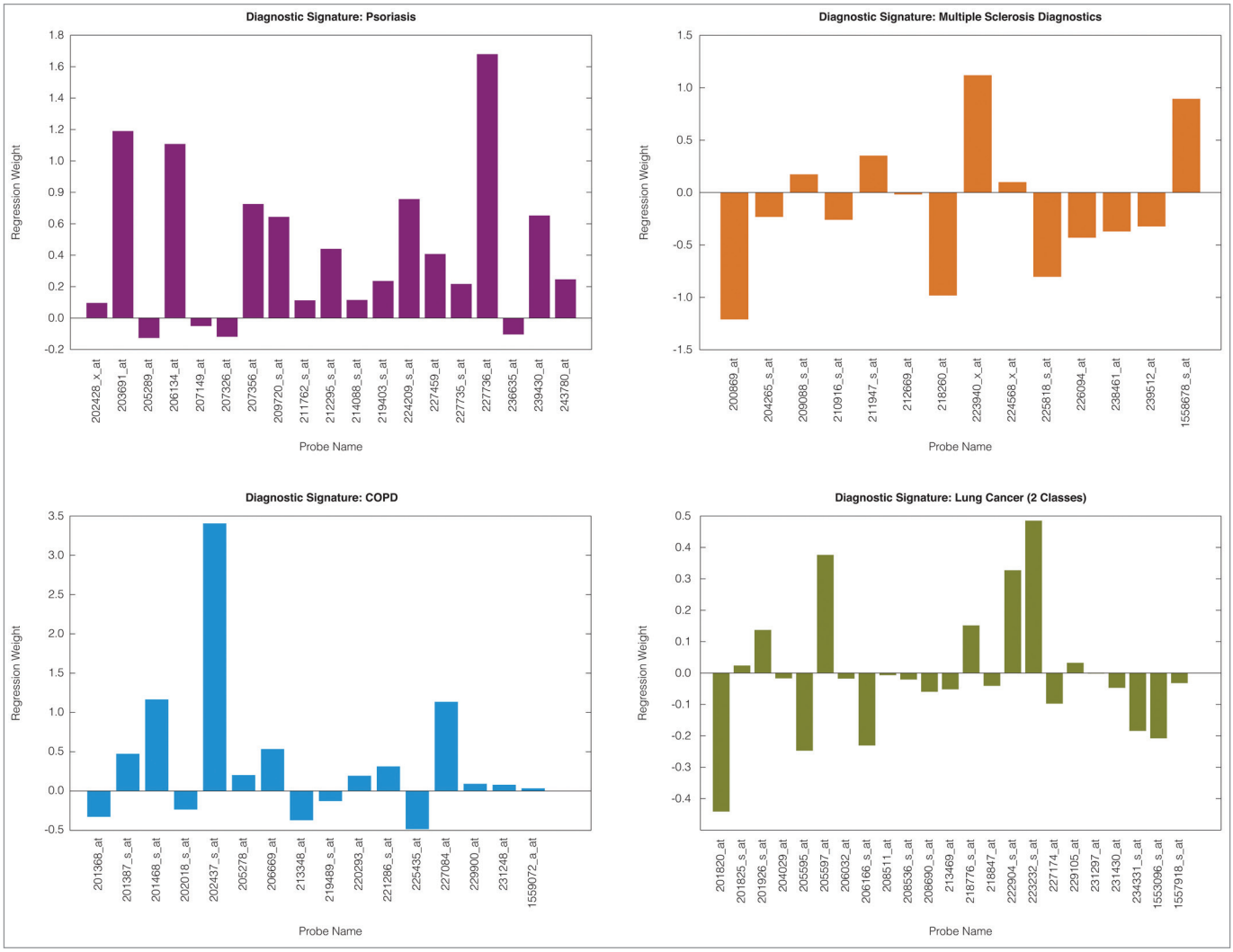

Figure 1. Diagnostic signatures for each of the four sub-challenges. Bar heights indicate how each probe is weighted in the final regressor.

IMPROVER challenge. In particular, we show that the method is able to extract disease-relevant gene profiles and demonstrate its potential in making diagnostic predictions.

Results

The Diagnostic Signature Challenge encompassed four independent classification tasks (sub-challenges), each task corresponded to a particular disease and data set. Three of the four sub-challenges were designated to distinguish between the disease/non-disease (i.e., binary classification) states. The goal of the fourth task, the lung cancer sub-challenge, was to predict four disease states corresponding to two different cancer types (adenocarcinoma (AC) and squamous cell carcinoma (SCC)) and their respective stages (stages I and II). The performance of each classifier was assessed by estimating its prediction success probability.

Computational workflow

Although the L1-regularized logistic regression provides a natural mechanism for feature selection and prevention of

overfitting (see Materials and Methods), it would require massive amounts of computational resources when directly applied to the high-dimensional data set. Thus, further preprocessing and data reduction had to be performed. More specifically, we followed a workflow that consisted of three main steps. In step one, we normalized the pooled data (comprising both the training and test data sets) for each of the sub-challenges using a standard robust multi-array average (RMA) normalization procedure.[16] In step two, we significantly reduced the dimensionality of the feature space using a nonparametric method based on the Wilcoxon rank sum test statistic.[17,18] In step three, the remaining features were used to train an L1-regularized logistic regression model. As indicated above, this approach allows to further reduce the number of features used in the final model.[4,5] The overall predictor for each disease is a monotonic function of the pre-processed and weighted feature intensities corresponding to the diagnostic signatures. Detailed descriptions of the three individual building blocks can be found in Materials and Methods.

Experimental results

The numbers of significantly expressed genes revealed by the feature pre-selection algorithm at a 10% level of significance are shown in **Table 1**. To some extent, the number of pre-selected genes reflects the richness of the disease signature in the expression profiles. Although a large number of pre-selected genes may improve the predictability of the disease state, the complexity of the subsequent classification task increases: the dimensionality becomes large compared with the sample size and standard approaches will inherently suffer from overfitting. Appropriate regularization strategies, such as provided by the L1-regularized logistic regression, can handle such problems to produce more reliable predictions.

The selected probe names and their corresponding weights for all four sub-challenges are shown in **Figure 1**. For each of those, the pre-selection algorithm was able to substantially reduce the number of features and hence, the dimensionality of the resulting data set. Because all the variables were standardized before training, the absolute weights represent the significance of the corresponding regressor.

Performance measures of our predictions were based on the score values of three IMPROVER standard quality metrics; namely, the belief confusion metric (BCM), the correct class enrichment metric (CCEM), and the average of the area under the precision recall curve (AUPR) across the classes (AUPR_Avg). **Table 2** shows the performance of our predictions according to those score values and the corresponding rank obtained for each of the sub challenges. Psoriasis was predicted well, while the other diseases were not. This might be partially explained by differences in the amount of available training data (**Fig. 2**). The graphic shows that most training samples where available for the psoriasis data set, which ranked best in our study. In contrast, the worst performance was achieved for the MS diagnostic data set, associated with a particularly small sample size.

However, a variety of other causes might have contributed to the variability in the performance. The tissue used to perform the microarray experiments did not always originate from a location primarily affected by the disease. This might cause strong qualitative differences between the training and test data sets, which might in turn have significant impact on the classification performance.

Table 2. The quality score values for the three standard quality metrics for each of the sub-challenges

Quality score	(BCM)	(CCEM)	(AUPR_Avg)	Rank obtained
Psoriasis	0.99	0.99	1.00	2
MSD	0.54	0.52	0.62	12*
COPD	0.66	0.68	0.66	4
LC (2 classes)**	0.82	0.84	0.94	N/A
LC (4 classes)	0.43	0.48	0.50	5

BCM, belief confusion metric; CCEM, correct class enrichment metric; AUPR_Avg, average of the area under the precision recall curve (AUPR) across the classes; MSD, multiple sclerosis diagnostic; COPD, chronic obstructive pulmonary disease; LC, lung cancer. *The original rank was 37. The training data set that we used for the MSD sub-challenge reported in this paper is different (basically a subset of the one used in the challenge) from that was used in the IMPROVER challenge. **LC (2classes) was not part of the IMPROVER challenge.

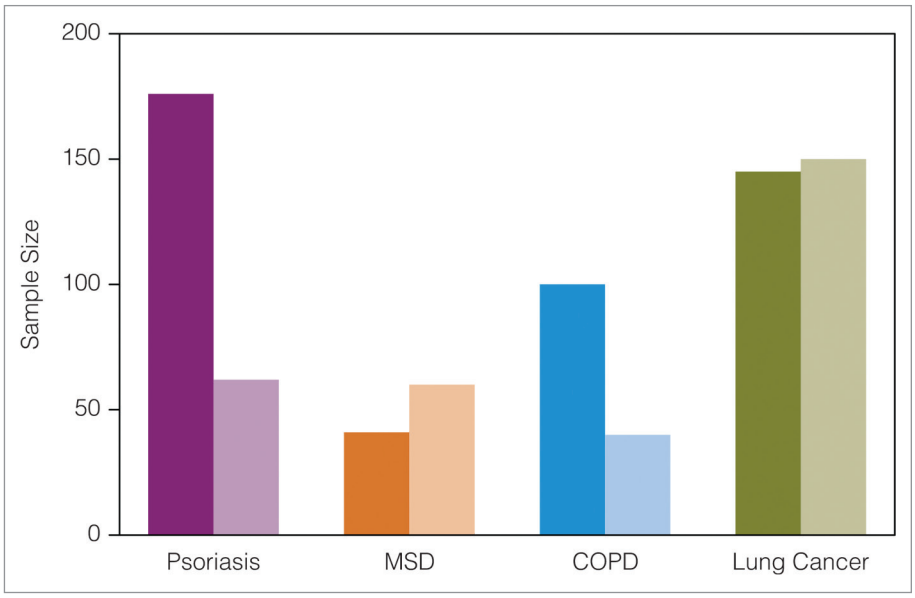

Figure 2. Sizes of the data sets that were available for each of the sub-challenges. Dark bars correspond to training data sets, light bars to test data sets.

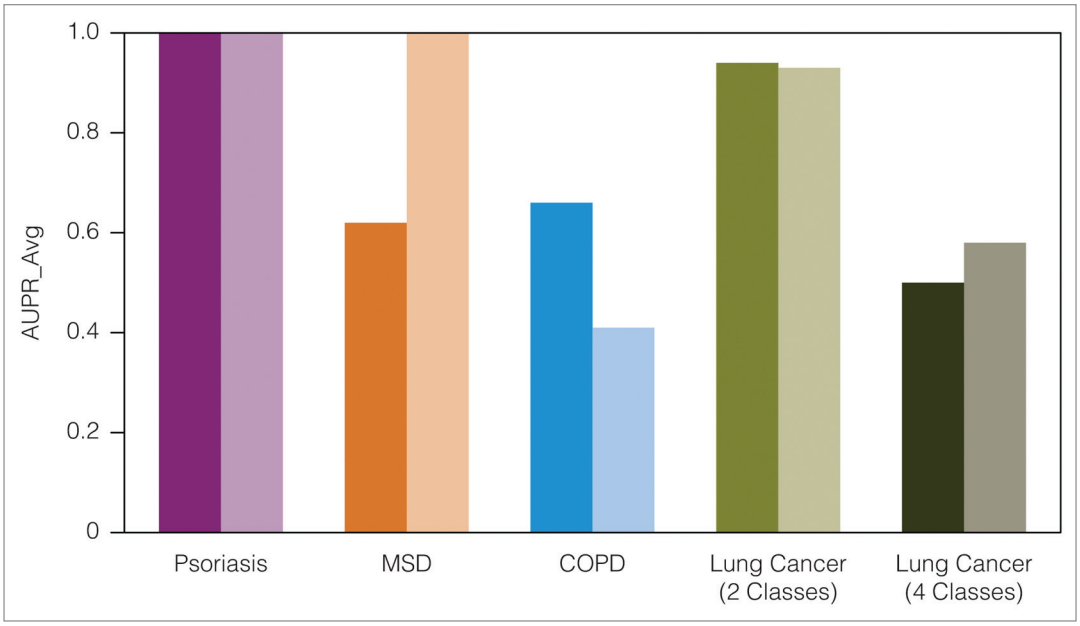

Figure 3. Performance of the Classifier based on the AUPR_Avgmetric scores. Dark bars correspond to training data sets, light bars to test data sets.

Table 3. The quality score values for the three standard quality metrics for the leave-one-out cross validation study

Quality score	(BCM)	(CCEM)	(AUPR_Avg)
Psoriasis	0.99	0.98	1.00
MSD	0.995	0.998	1
COPD	0.47	0.48	0.41
LC (2 classes)	0.77	0.80	0.93
LC (4 classes)	0.48	0.54	0.58

The test data corresponding to each of the sub challenges was used. BCM, belief confusion metric; CCEM, correct class enrichment metric; AUPR_Avg, average of the area under the precision recall curve (AUPR) across the classes; MSD, multiple sclerosis diagnostic; COPD, chronic obstructive pulmonary disease; LC, lung cancer. [a]This check was possible only after the gold standard labels of the test samples were published online.

In order to test for such differences, we evaluated our classifier against the test sets of the respective sub-challenges using a leave-one-out cross-validation.[a] Those results were compared with the original predictions obtained from the training data sets by means of the AUPR_Avg metric (**Fig. 3**). The remaining performance scores are listed in **Table 3**. In case of the psoriasis data, we observed only minor differences in the performance, even though the classifier was obtained from significantly fewer samples.

In accordance with our hypothesis, a considerable improvement was obtained for the MSD data set, indicating strong differences between the training and test data set. Furthermore, when using the latter, the number of available training samples (i.e., N = 59) was higher than the original sample size of the training data set (i.e., N = 41).

For the COPD sub-challenge, the performance of the classifier trained solely on the test data set was no better than a coin flip (i.e., the success probability was around 0.5). This result suggests that either the available data set does not contain enough disease-relevant information or the proposed approach is unable to unravel the complexity of the underlying expression patterns. Although COPD is manifested in small airways, the goal was to identify a COPD signature valid for large airways (such as, in this case, the test data set) for which sample collection is less complex. In case of the training data set, consisting of samples from both large and small airways, it seems that the classifier was indeed able to extract predictive gene signatures for large airways data.

For the LC sub-challenge, the size of the training set (N = 145) and the size of the test set (N = 150) were roughly the same. However, when LC was considered as a binary classification problem (i.e., classes AC and SCC irrespective of their stage), we found that the classifier performed well in both cases, while for the initial four-class problem (i.e., discriminating between their corresponding stages) the performance was only moderate.

Discussion

In this work we proposed a three-stage computation workflow for extracting diagnostic gene signatures from microarray gene expression data. In order to account for technical and biological variations between individual samples, we first preprocessed the data using a robust multi-array normalization scheme. In order to reduce the dimensionality of the data sets, we applied a feature pre-selection algorithm using a Wilcoxon's rank sum statistic.

The primary classification algorithm is based on an L1-regularized logistic regression model, which on the one hand is able to prevent overfitting and on the other hand, provides a simple strategy to identify predictive gene signatures. More

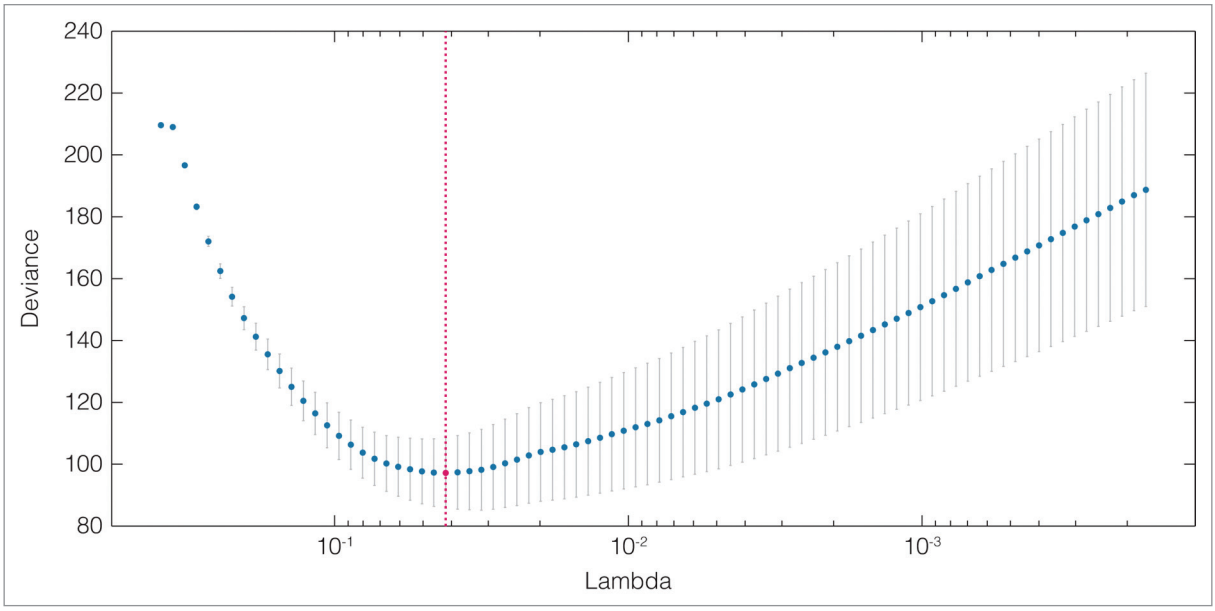

Figure 4. Cross-validated estimates of the deviance and confidence bound of the LASSO fit.

specifically, the regression weights of the model directly indicate the significance of each gene and thus, allow a straightforward interpretation of the obtained results.

We demonstrated the usefulness of the approach using microarray data sets from four different disease areas, i.e., psoriasis, multiple sclerosis, chronic obstructive pulmonary disease and lung cancer. For most of the prediction tasks, the classification algorithm performed reasonably well. In particular, the psoriasis data sets were handled surprisingly well. In cases where weak scores were achieved, we performed additional analyses to pinpoint the factors that may have led to a decreased performance. For instance, in case of the MSD data set, our results from the leave-one-out cross-validation study indicate significant qualitative differences between the training and test data sets.

Our results demonstrate that statistical methods in conjunction with modern microarray gene expression technology provide powerful and important means to accurately diagnose complex diseases.

Materials and Methods

Data normalization

For all sub-challenges, only training data stemming from Affymetrix® GeneChip Human Genome U133 Plus 2.0 microarrays were used, since all test data sets were generated on this platform. This was done to avoid any bias to our model that could have been introduced by including data from other chips in the training phase. Thus, normalization between different types of microarray chips was not needed, but normalization to remove batch effects between different experiments was still essential to make the data sets comparable. We normalized the pooled data sets (comprising both the training and test data sets for each of the individual sub-challenges) using a standard RMA normalization procedure.[16]

Feature pre-selection

Before we used the data sets to train the classifier, the dimensionality of the feature space was reduced substantially by applying a feature pre-selection method. The aim was to select only those features that were significantly up or downregulated between case and control groups. We applied a nonparametric method based on the Wilcoxon ranksum test statistic.[17,18]

For each feature, we tested the null hypothesis that the distributions of its expression value over the case and control probes in the microarray data sets are equal, against the alternative that one distribution is stochastically larger than the other. This test is equivalent to the Wilcoxon two-sample test (also known as the Mann-Whitney U test). For each gene *g*, we obtain,

$$Score(g) = \sum_{i \in N_0} \sum_{j \in N_1} 1_{\{x_j^{(g)} - x_i^{(g)} \le 0\}},$$

where $x_j^{(g)}$ is the expression value of gene *g* for an individual *i* and N_m represents the set of indices having a response in $m \in \{0, 1\}$. The score function counts the number of instances where an expression value corresponding to a response *1* is smaller than an expression value corresponding to a response *0*. Therefore, the score would be close to the maximum score $|N_0|\,|N_1|$ for any gene that tends to be under-expressed in response *1* and close to *0* for a gene that tends to be overexpressed in individuals in N_1.

Clearly, the aim was to identify genes with small *P*-values for the corresponding Wilcoxon two-sample test, which is based on the test statistic $Score(g)$. At 10% level of significance, we selected only the genes that had *P*-values less than 0.1/(total number of genes), using the Bonferroni correction under the multiple comparison setup.

Although this method of pre-selection can filter out genes that are predictive individually, it does not help to identify the best predictive combination of genes. For this reason, if the

resultant data set contains very few genes with *P*-values less than 0.1/(total number of genes), the resultant data set will no more be reliable since some valuable information might have already been thrown away. In addition, the Bonferroni significance level is quite conservative. To avoid an excessive loss of features, the first 2000 genes, ordered by their *P*-values were picked if the pre-selection method initially yielded less than 2000 genes.

Training the primary classifier

We used a logistic regression model to fit the training data and to classify the test data. Despite the feature pre-selection, the feature space was yet 4–28% of the total number of probes on the chip (54,675). This was still high compared with the training data sample size. A simple logistic regression model[19] would lead to overfitting.[4] We therefore used an L1-regularized logistic regression model to drive a large number of less significant parameters to 0 and filter out only those genes that played a significant role in classifying the data into case and control groups.

Let $Y_i \in \{0,1\}$ be the random variable that represents the response of the *i*th individual. Now we define the standardized expression value of gene *g* for individual *i* by

$$z_i^{(g)} = \frac{x_i^{(g)} - \mu_g}{\sigma_g},$$

with

$$\mu_g = \frac{1}{n}\sum_{i=1}^{n} x_i^{(g)}$$

and

$$\sigma_g = \frac{1}{n-1}\sum$$

Then, our model is

$$\pi_i := Pr\left(Y_i = 1\right) = \frac{\exp(\alpha + \sum_{g=1}^{p}\beta_g z_i^{(g)})}{1+\exp(\alpha + \sum_{g=1}^{p}\beta_g z_i^{(g)})}.$$

where *p* is the total number of genes under consideration. Hence, the likelihood of the observed data are

$$\mathrm{L}(\alpha,\beta_1,\ldots,\beta_p|y_1,\ldots,y_n) := \Pr\left(Y_1 = y_1,\ldots,Y_n = y_n\right) = \prod_{i=1}^{n}\frac{[\exp(\alpha + \sum_{g=1}^{p}\beta_g z_i^{(g)})]^{1_{\{y_i=1\}}}}{1+\exp(\alpha + \sum_{g=1}^{p}\beta_g z_i^{(g)})}.$$

Therefore, an estimate of the parameter-vector $\theta = (\alpha,\beta_1,\ldots,\beta_p)'$ can be obtained by maximizing the log-likelihood function as

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \log L(\alpha,\beta_1,\ldots,\beta_p|y_1,\ldots,y_n) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n}\left[1_{\{y_i=1\}}\left(\alpha + \sum_{g=1}^{p}\beta_g z_i^{(g)}\right) - \log\left(1+\exp\left(\alpha + \sum_{g=1}^{p}\beta_g z_i^{(g)}\right)\right)\right].$$

As mentioned earlier, we had to avoid overfitting, and thus optimized a penalized log-likelihood with an L1 penalty in β_g as

$$J(\theta) = \log L(\alpha,\beta_1,\ldots,\beta_p|y_1,\ldots,y_n) + \lambda\left(|\alpha| + \sum_{g=1}^{p}|\beta_g|\right).$$

The regularization or tuning parameter λ was fixed to the value that yielded the lowest L1-regularized deviance ($-2J(\theta)$), out of a 30-fold cross-validation on the training data set. **Figure 4** shows the cross-validated deviance estimates and confidence bounds for each proposed λ, as well as the selection of the optimal regularization parameter for the LC (2 classes) task.

Note that this is a convex optimization problem that can be solved efficiently. We used the MATLAB *lassoglm()* function, which uses the coordinate descent algorithm[20] to solve the optimization problem for a given regularization parameter l. After obtaining the estimates of the parameter vector, the probability that an individual with expression value $x^{(g)}$ for gene *g*, belongs to class *1* (i.e., has the response *1*), is given by

$$\hat{\pi}\left(x^{(1)},\ldots,x^{(p)}\right) = \frac{\exp\left(\hat{\alpha} + \sum_{g=1}^{p}\hat{\beta}_g\frac{x^{(g)}-\mu_g}{\sigma_g}\right)}{1+\exp\left(\hat{\alpha} + \sum_{g=1}^{p}\hat{\beta}_g\frac{x^{(g)}-\mu_g}{\sigma_g}\right)}.$$

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank the IMPROVER committee for organizing the Diagnostic Signature Challenge. The challenge provided a great opportunity for us to develop our classification algorithms and validate it using real-life experimental data sets.

References

1. Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, et al.; MAQC Consortium. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 2006; 24:1151-61; PMID:16964229; http://dx.doi.org/10.1038/nbt1239
2. Petricoin EF 3rd, Hackett JL, Lesko LJ, Puri RK, Gutman SI, Chumakov K, et al. Medical applications of microarray technologies: a regulatory science perspective. Nat Genet 2002; 32(Suppl):474-9; PMID:12454641; http://dx.doi.org/10.1038/ng1029
3. Li X, Quigg RJ, Zhou J, Gu W, Nagesh Rao P, Reed EF. Clinical utility of microarrays: current status, existing challenges and future outlook. Curr Genomics 2008; 9:466-74; PMID:19506735; http://dx.doi.org/10.2174/138920208786241199
4. Bishop CM. (2006). Pattern recognition and machine learning, Vol 4 (Springer)
5. Friedman J, Hastie T, Tibshirani R. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn (Springer New York)
6. Spang R. Diagnostic signatures from microarrays: a bioinformatics concept for personalized medicine. BIOSILICO 2003; 1:64-8; http://dx.doi.org/10.1016/S1478-5382(03)02329-1
7. Dudoit S, Fridlyand J, Speed T. Comparison of Dicrimination Methods for the Classification of Tumors Using Gene Expresion Data. J Am Stat Assoc 2002; 97:77-87; http://dx.doi.org/10.1198/016214502753479248
8. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A 2000; 97:262-7; PMID:10618406; http://dx.doi.org/10.1073/pnas.97.1.262
9. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000; 16:906-14; PMID:11120680; http://dx.doi.org/10.1093/bioinformatics/16.10.906

10. Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001; 7:673-9; PMID:11385503; http://dx.doi.org/10.1038/89044
11. Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK. Gene selection: a Bayesian variable selection approach. Bioinformatics 2003; 19:90-7; PMID:12499298; http://dx.doi.org/10.1093/bioinformatics/19.1.90
12. Dettling M, Bühlmann P. Finding predictive gene groups from microarray data. J Multivariate Anal 2004; 90:106-31; http://dx.doi.org/10.1016/j.jmva.2004.02.012
13. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 2009; 25:714-21; PMID:19176549; http://dx.doi.org/10.1093/bioinformatics/btp041
14. Meyer P, Alexopoulos LG, Bonk T, Califano A, Cho CR, de la Fuente A, et al. Verification of systems biology research in the age of collaborative competition. Nat Biotechnol 2011; 29:811-5; PMID:21904331; http://dx.doi.org/10.1038/nbt.1968
15. Meyer P, Hoeng J, Rice JJ, Norel R, Sprengel J, Stolle K, et al. Industrial methodology for process verification in research (IMPROVER): toward systems biology verification. Bioinformatics 2012; 28:1193-201; PMID:22423044; http://dx.doi.org/10.1093/bioinformatics/bts116
16. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003; 4:249-64; PMID:12925520; http://dx.doi.org/10.1093/biostatistics/4.2.249
17. Dettling M, Bühlmann P. Boosting for tumor classification with gene expression data. Bioinformatics 2003; 19:1061-9; PMID:12801866; http://dx.doi.org/10.1093/bioinformatics/btf867
18. Park P, Pagano M, Bonetti M. (2001). A nonparametric scoring algorithm for identifying informative genes from microarray data. Pacific Symposium on Biocomputing 63, 52-63
19. Collett D. (2003). Modelling Binary Data, 2nd edn (Chapman & Hall/CRC)
20. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010; 33:1-22; PMID:20808728