Gesellschaft für Informatik e.V. (GI)

publishes this series in order to make available to a broad public recent findings in informatics (i.e. computer science and information systems), to document conferences that are organized in co-operation with GI and to publish the annual GI Award dissertation.

Broken down into
- seminar
- proceedings
- dissertations
- thematics

current topics are dealt with from the vantage point of research and development, teaching and further training in theory and practice. The Editorial Committee uses an intensive review process in order to ensure high quality contributions.

The volumes are published in German or English.

Information: http://www.gi-ev.de/service/publikationen/lni/

ISSN 1617-5468
ISBN 978-3-88579-267-3

This volume contains papers presented at the 25th German Conference on Bioinformatics held at the Technische Universität Carolo-Wilhelmina in Braunschweig, Germany, September 20-22, 2010. The German Conference on Bioinformatics is an annual, international conference, which provides a forum for the presentation of current research in bioinformatics and computational biology. It is organized on behalf of the Special Interest Group on Informatics in Biology of the German Society of Computer Science (GI) and the German Society of Chemical Technique and Biotechnology (Dechema) in cooperation with the German Society for Biochemistry and Molecular Biology (GBM).

D. Schomburg, A. Grote (Eds.): GCB 2010

173

GI-Edition

Lecture Notes in Informatics

**Dietmar Schomburg,
Andreas Grote (Eds.)**

German Conference on Bioinformatics 2010

**September 20 - 22, 2010
Braunschweig, Germany**

Proceedings

Dietmar Schomburg, Andreas Grote (Editors)

German Conference on Bioinformatics 2010

September 20-22, 2010
Technische Universität Carolo Wilhelmina zu Braunschweig, Germany

Gesellschaft für Informatik e.V. (GI)

Lecture Notes in Informatics (LNI) - Proceedings
Series of the Gesellschaft für Informatik (GI)

Volume P-173

ISBN 978-3-88579-267-3
ISSN 1617-5468

Volume Editors
Prof. Dr. Dietmar Schomburg
Technische Universität Carolo-Wilhelmina zu Braunschweig
Email: d.schomburg@tu-bs.de
Dr. Andreas Grote
Technische Universität Carolo-Wilhelmina zu Braunschweig
Email: andreas.grote@tu-bs.de

printed by Köllen Druck+Verlag GmbH, Bonn

Preface

This volume contains papers presented at the 25th German Conference on Bioinformatics held at the Technische Universität Carolo-Wilhelmina in Braunschweig, Germany, September 20-22, 2010. The German Conference on Bioinformatics is an annual, international conference, which provides a forum for the presentation of current research in bioinformatics and computational biology. It is organized on behalf of the Special Interest Group on Informatics in Biology of the German Society of Computer Science (GI) and the German Society of Chemical Technique and Biotechnology (Dechema) in cooperation with the German Society for Biochemistry and Molecular Biology (GBM). Five outstanding scientists were invited to give keynote lectures to the conference:

- Edda Klipp - 'Cellular stress response and regulation of metabolism'
- Thomas Lengauer - 'HIV Bioinformatics: Analyzing viral evolution for the benefit of AIDS patients'
- Werner Mewes - 'The data deluge: can simple models explain complex biological systems?'
- Stefan Schuster - 'Road games in metabolism - A biotechnological perspective'
- Gregory Stephanopoulos - 'After a decade of systems biology, time for a record card'

Besides the keynote lectures, the scientific program comprised 22 contributed talks presenting 12 regular and 10 short papers. All accepted regular papers are collected in these proceedings. The remaining accepted contributions, i.e. short papers and poster abstracts, are published in a separate volume. We would like to thank the program committee members and all reviewers for their evaluations of the contributions. Furthermore, we would like to thank the local organizers for keeping the conference running. The organizers are grateful to all the sponsors and supporting scientific partners. Last but not least, we would like to thank all contributors and participants of the GCB 2010.

Braunschweig, August 2010

Dietmar Schomburg and Andreas Grote

Organizers

Conference Chair

Dietmar Schomburg, Braunschweig

Local Organizers

Wolf-Tilo Balke (TU Braunschweig)
Sándor Fekete (TU Braunschweig)
Reinhold Haux (TU Braunschweig)
Dieter Jahn (TU Braunschweig)
Frank Klawonn (Ostfalia University of Applied Sciences)
Constantin Bannert (TU Braunschweig)
Antje Chang (TU Braunschweig)
Andreas Grote (TU Braunschweig)
Katharina Hanke (TU Braunschweig)
Adam Podstawka (TU Braunschweig)
Alexander Riemer (TU Braunschweig)
Maurice Scheer (TU Braunschweig)

Programm committee

Mario Albrecht, Saarbrücken
Wolf-Tilo Balke, Braunschweig
Tim Beißbarth, Göttingen
Thomas Dandekar, Würzburg
Sándor Fekete, Braunschweig
Georg Füllen, Rostock
Robert Giegerich, Bielefeld
Reinhold Haux, Braunschweig
Ralf Hofestädt, Bielefeld
Matthias Heinemann, Zürich
Dieter Jahn, Braunschweig
Frank Klawonn, Wolfenbüttel
Edda Klipp, Berlin
Ina Koch, Berlin
Oliver Kohlbacher, Tübingen
Thomas Lengauer, Saarbrücken
Hans-Peter Lenhof, Saarbrücken
Michael Marschollek, Hannover
Werner Mewes, München
Michael Meyer-Hermann Frankfurt
Burkhard Morgenstern, Göttingen
Stefan Posch, Halle
Matthias Rarey, Hamburg
Falk Schreiber, Gatersleben
Stefan Schuster, Jena
Gregory Stephanopoulos, Cambridge USA
Jens Stoye, Bielefeld
Andrew Torda, Hamburg
Martin Vingron, Berlin
Christian von Mering, Zürich
Edgar Wingender, Göttingen
Andreas Ziegler, Lübeck

Sponsors and supporters

Supporting scientific societies

Gesellschaft für Chemische Technik und Biotechnologie e.V. (DECHEMA)
http://www.dechema.de

Gesellschaft für Biochemie und Molekularbiologie e.V. (GBM)
http://www.gbm-online.de

Gesellschaft für Informatik e.V. (GI)
http://www.gi-ev.de

Non-profit sponsors

Technische Universität Braunschweig
http://www.tu-braunschweig.de

Commercial sponsors

Biobase - Biological Databases
http://www.biobase-international.com

CLC bio
http://www.clcbio.com

Convey Computer
http://www.conveycomputer.com

genomatix
http://www.genomatix.de

geneXplain
http://www.genexplain.com

MEGWARE Computer Cluster
http://www.megware.com

Thalia
http://www.thalia.de

Transtec - IT & Solutions
http://www.transtec.de

Table of Contents

RNALfoldz: efficient prediction of thermodynamically stable, local secondary structures

Andreas R. Gruber[1], Stephan H. Bernhart[1],
You Zhou[1,2], and Ivo L. Hofacker[1]
[1] *Institute for Theoretical Chemistry*
University of Vienna, Währingerstraße 17, 1090 Wien, Austria
[2] College of Computer Science and Technology
Jilin University, Changchun 130012, China
{agruber, berni, ivo}@tbi.univie.ac.at, zyou@jlu.edu.cn

Abstract: The search for local RNA secondary structures and the annotation of unusually stable folding regions in genomic sequences are two well motivated bioinformatic problems. In this contribution we introduce RNALfoldz an efficient solution two tackle both tasks. It is an extension of the RNALfold algorithm augmented by support vector regression for efficient calculation of a structure's thermodynamic stability. We demonstrate the applicability of this approach on the genome of *E. coli* and investigate a potential strategy to determine z-score cutoffs given a predefined false discovery rate.

1 Introduction

Over the past decade noncoding RNAs (ncRNAs) have risen from a shadowy existence to one of the primary research topics in modern molecular biology. Today computational RNA biology faces challenges in the ever growing amount of sequencing data. Efficient computational tools are needed to turn these data into information. In this context, the search for locally stable RNA secondary structures in large sequences is a well motivated bioinformatic problem that has drawn considerable attention in the community. RNALfold [HPS04] has been the first in a series of tools that offered an efficient solution to this task. Instead of a straight-forward, but costly sliding window approach a dynamic programming recursion has been formulated that predicts all stable, local RNA structures in $\mathcal{O}(N \times L^2)$, where L is the maximum base-pair span and N the length of the sequence. Since its publication, the RNALfold algorithm has inspired a lot of work in this field, see e.g. Rnall by Wan *et al.* [WLX06] or RNAslider by Horesh *et al.* [HWL$^+$09]. All contributions so far in this field focused on improving the computational complexity of the algorithm, but none of the approaches has ever been used to unravel results of biological significance. In particular, *de novo* detection of functional RNA structures has been addressed, but application on a genome-wide scale with a low false discovery rate seems still out of reach. Even on the moderately sized genome of *E. coli* (4.6 Mb) one is drowning in hundreds of thousands of local structures. Unlike in the well established field of protein coding gene detection where one can exploit signals like codon usage, functional

RNA secondary structures, in general, do not show strong characteristics that make them easily distinguishable from random decoys. Successful approaches for ncRNA detection operating solely on a single sequence [HHS08, JWW$^+$07] are limited to specific RNA classes, where some outstanding characteristics can be harnessed. There is no master plan for the detection of functional RNA structures, but one would certainly want to limit the `RNALfold` output to a reasonable amount. So far, only the minimum free energy (MFE) of the locally stable secondary structures, which is intrinsically computed by the algorithm, has been considered as potential discriminator to limit the number of secondary structures. As demonstrated clearly by Freyhult and colleagues [FGM05] the MFE is roughly a function of the length of the sequence and the G+C content. Even normalizing the MFE by length of the sequence does not serve as a good discriminator between shuffled or coding sequences and functional RNA structures. A strategy that does work, however, is to compare the native MFE E of the RNA molecule to the MFEs of a set of shuffled sequences of same length and base composition [LM89]. This way we can evaluate the thermodynamic stability of the secondary structure. A common statistical quantity in this context is the z-score, which is calculated as follows

$$z = \frac{E - \mu}{\sigma}$$

where μ and σ are the average and the standard deviation of the energies of the set of shuffled sequences. The more negative the z-score the more thermodynamically stable is the structure. Efficient estimation of a sequence's z-score has been a profound problem already addressed in the very beginnings of computational RNA biology. A first strategy to avoid explicit shuffling and folding was based on table look-ups of linear regression coefficients [CLS$^+$90]. Clote and colleagues [CFKK05] introduced the concept of the asymptotic z-score, where the efficient calculation is also solved via table look-ups. The current state-of-the art approach for fast and efficient estimation of the z-score is to use support vector regression [WHS05].

The study by Clote and colleagues and a follow up to Chen *et al.* (1990) [LLM02] also report on the effort to predict thermodynamically stable structures using a sliding window approach. In this contribution we present `RNALfoldz` an algorithm that combines local RNA secondary structure prediction and the efficient search for thermodynamically stable structures. `RNALfoldz` is an extension of the `RNALfold` algorithm augmented by support vector regression for efficient calculation of a sequence's z-score. We demonstrate the applicability of this approach on the genome of *E. coli* and investigate a potential strategy to determine z-score cutoffs given a predefined false discovery rate.

2 Methods

2.1 Fast estimation of the z-score using support vector regression

For the efficient estimation of the z-score we follow the strategy first introduced by Washietl *et al.* [WHS05]. Instead of explicit generation and folding of shuffled sequences in order to determine the average free energy and the corresponding standard deviation support vector regression (SVR) models are trained to estimate both values. As described in detail in the previous work, we used a regularly spaced grid to sample sequences for the training set. Synthetic sequences ranged from 50 to 400 nt in steps of 50 nt. The G+C content, A/(A+T) ratio and C/(C+G) ratio were, however, extended to a broader spectrum, now ranging from 0.20 to 0.80 in steps of 0.05. A total of 17,576 sequences were used for training. For each sequence of the training set 1,000 randomized sequences were generated using the Fisher-Yates shuffle algorithm, and subsequently folded with `RNAfold` with dangling ends option `-d2` [HFS$^+$94]. SVR models for the average free energy and standard deviation were trained using the `LIBSVM` package (`www.csie.ntu.edu.tw/~cjlin/libsvm`). While in the previous work input features and the dependent variables were normalized to a mean of zero and a standard deviation of one, we apply here a different normalization strategy that leads to a significantly lower number of support vectors for the final models. For the regression of the average free energy model the dependent variable is normalized by the length of the sequence, while for the standard deviation it is the square root of the sequence length. The length still remains in the set of input features and is scaled from 0 to 1. Other features remain unchanged. We used a RBF kernel, and optimized values for the SVM parameters were determined using standard protocols for this purpose. Final regression models were selected by balancing two criteria: (i) mean absolute error (MAE) on a test set of 5,000 randomly drawn sequences of arbitrary length (50-400) from the human genome, and (ii) complexity of the model (number of support vectors) , which translates to following procedure: from the top 10% of regression models in terms of MAE we selected the one that had the lowest number of support vectors. For the average free energy regression we selected a model with a MAE of 0.453 and 1,088 support vectors, and for the standard deviation regression a model with a MAE of 0.027 and 2,252 support vectors. To validate our approach we finally compared z-scores derived from the SVR to traditionally sampled z-scores on a set of 1,000 randomly drawn sequences from the human genome. The correlation coefficient (R) is 0.9981 and the MAE is 0.072. This is in fair agreement to results obtained when comparing two sets of sampled z-scores (R: 0.9986, MAE: 0.054, number of shuffled sequences = 1,000).

2.2 Adaption of the `RNALfold` algorithm

RNALfold computes locally stable structures of long RNA molecules. It uses a Zuker type secondary structure prediction algorithm [ZS81] and restricts the maximum base pair span to L bases to keep the structures local. The sequence is processed from the 3' (the sequence length n) to the 5' end. In order to keep the number of back trace operations low and the output at moderate size, we want to avoid backtracing structures that differ only by unpaired regions. Furthermore, only the longest helices possible are of interest. To achieve this, a structure starting at base i is only traced back if the total energy $F(i, n)$ is smaller than that of its 3' neighbor $F(i + 1, n)$ while its 5' neighbor has the same energy: $F(i-1, n) = F(i, n) < F(i+1, n)$. The local minimum structure is found by identifying the pairing partner j of i so that $C(i, j) + F(j + 1, n) = F(i, n)$, i.e. the minimum energy from i to n is decomposed into the local minimum part i, j and the rest of the molecule. Here, $C(i, j)$ stands for the energy of a structural feature enclosed by the base pair i, j. As a result of this, the output of RNALfold contains components, i.e. structures that are enclosed by a base pair, only. Before we actually start the trace back, we evaluate two new criteria: (1) the sequence of the structure traced back has to be within the training parameters of the SVR, and (2) the z-score of the energy of this structure has to be lower than a predefined bound. Criterion (1) is simply imposed by the training boundaries of the SVMs. Boundaries have, however, been chosen carefully to cover a broad range of today's known spectrum of functional RNA structures. 99.79% of the sequences in the Rfam v. 10 full data set match the base composition requirements of the SVR and 90% of Rfam RNA families are in within the sequence length restrictions.

In order to get the exact sequence composition that is needed for the SVR evaluations, the 3' end of the structure (j) has to be computed first. This is done by a first, short backtracing step, where the decomposition $F(i, n) = C(i, j) + F(j + 1, n)$ is used to find j. Subsequently, the average free energy given the base composition of the sequence $s(i, j)$ is computed by calling the corresponding SVR model. The SVR model for the standard deviation has approximately twice the number of support vectors as the average free energy model. To minimize calls of this model, first the minimal standard deviation for the particular sequence length is looked up. We can then, using the free energy of $C(i, j)$, calculate a lower bound of the z-score. Only if this lower bound is below the minimal required z-score, the support vector regression for the standard deviation is called to calculate the actual z-score. If the z-score then still meets the minimal z-score criterion, the structure is fully traced back and printed out.

3 Results

The concept of fast and efficient estimation of the z-score by support vector regression was first introduced by Washietl *et al.* [WHS05], and implemented in the noncoding RNA gene finder RNAz. The speed up of this approach compared to explicit shuffling and folding, which is usually done on 1,000 replicas, is tremendous, at minimum a factor of 1,000. Moreover, computing time is invariant to the length of the sequence, while RNA folding is of complexity of $\mathcal{O}(N^3)$. When considering the z-score as evaluation criterion in the RNALfold algorithm, calculation of the z-score becomes a time consuming factor, as in a worst case scenario it has to be done almost for every nucleotide of the sequence. It is therefore a crucial concern to use support vector models that do not only have good accuracy, but also a moderate number of support vectors (SVs). In this work we extended the z-score support vector regression to cover a broader range of the sequence spectrum, but managed at the same time to build models that have significantly less SVs than the models used by RNAz. This was accomplished by normalizing the dependent variables of the regression, i. e. the average free energy and the standard deviation, by the sequence length. The dependent variables do not strictly linearly correlate with the sequence length and so we have to keep the sequence length as an input feature. Nevertheless, redundant points are created in the training set, which eventually leads to a smaller space to be trained. For the average free energy model and the standard deviation model we were able to achieve a 6.3 and a 2.7 fold reduction, respectively, in the number of SVs compared to the RNAz equivalents.

3.1 Evaluation of **RNALfoldz** predicition accuracy

For the task of predicting local RNA secondary structures one would particularly be interested in following criteria: (i) to which extent can functional ncRNAs be discovered, (ii) how well do the molecule's predicted boundaries match to the real coordinates, and (iii) is there any significant difference between native, biological sequences and random decoys. To address these questions, we applied RNALfoldz to the genome of *E. coli* (Accession number: CP000948). A maximum base-pair span L of 120 nt and a z-score cutoff of -2 was used. Setting the cutoff at -2 is for sure restrictive, but it should still cover a large fraction of the ncRNA repertoire. Both strands were considered. A total of 202,126 structures have been obtained. In comparison, the regular RNALfold returned a total of 1,387,136 structures, 824, 000 of which have a length $\geq$ 50 nt. The RNALfoldz output (a true subset of the RNALfold output) is only a forth of the regular RNALfold output.

The *E. coli* genome Genbank file lists 119 ncRNAs with a maximum length of 120 nt in its current annotation. To investigate the extent annotated ncRNAs are covered in the RNALfoldz output, we define for a RNALfold/RNALfoldz prediction to be counted as hit a minimal coverage of 90% of the ncRNA sequence. Giving this setup a total of 106 and 89 ncRNAs can be found in the RNALfold and RNALfoldz output, respectively. Detailed results for each RNA gene are shown in an online supplementary table. With a z-score cutoff of -2, 17 ncRNAs that were found by RNALfold are not in output set of RNALfoldz. The detection success is directly proportional to the reduction rate of the RNALfold output. Modulating the z-score cutoff affects both quantities (Fig. 1). The failure to detect the 13 ncRNAs that were missed by both RNALfold and RNALfoldz results from the fact that the RNALfold algorithm predicts only self-contained RNA structures. For example, the two ncRNA genes *rprA* and *ryeE* that have only low covering RNALfoldz hits, have indeed multi-component structures at the MFE level (abstract shape notation [GVR04]: [][][][], [][][]). In these cases RNALfoldz will rather produce multiple hits than one single hit covering the whole ncRNA. Overall, our findings confirm that most *E. coli* small ncRNAs are indeed more thermodynamically stable than expected by chance and that the RNALfoldz algorithm is able to detect these structures efficiently.

We further investigated how precisely the RNALfoldz predictions map to the coordinates of the annotated ncRNAs. This is a legitimate issue, but one has to keep in mind that functional RNAs adopt their structure at the transcription level, while in this experiment we used the genomic sequence to detect these structures. So it might easily happen that the RNA is predicted in a bigger structural context than its actual size. The underlying dynamic programming algorithm is the same for RNALfold and RNALfoldz, and hence results discussed here do hold for both versions. In this work we define *noise* as the fraction of the RNALfoldz hit that does not overlap with the annotated ncRNA. In total, 34 out of

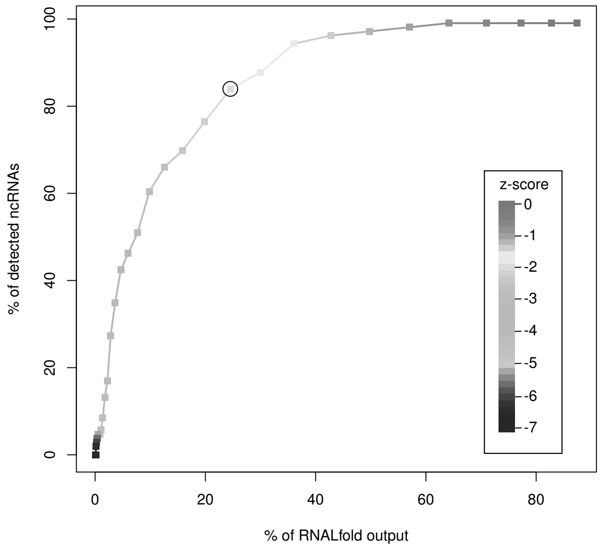


Figure 1: Non-coding RNA detection success vs. reduction of the `RNALfold` output. Given a z-score cutoff of 0 only one structure prediction is missed in the `RNALfoldz` output. With a z-score cutoff of -2 (circle) we see a four-fold reduction of the output, while at the same time covering 84% of the correct `RNALfold` ncRNA predictions.

the 89 predictions have less than 10% noise. Averaged over all hits ($\geq$ 90% coverage) we see *noise* of 18%. Using a classic sliding window approach with a length of 120 nt, one would expect more than 33% noise for a window containing a tRNA sequence (average length of tRNAs in E. coli: 78 nt). In the `RNALfoldz` output we find that 29 out of 73 tRNA predictions have less than 10% noise.

Finally, we address the significance of the predictions when compared to randomized controls. Therefore, we performed the same experiment on randomized sequences generated by (i) mononucleotide shuffling, (ii) simulation with an order-0 Markov model (mononucleotide frequencies) , and (iii) simulation with an order-1 Markov model (dinucleotide frequencies). Shuffling and simulations were done with `shuffle` from Sean Eddy's squid library using default parameters. A detailed comparison of the results of these four experiments is shown in Fig. 2. In general, we observe a tendency to more stable structures in the native sequence than in any randomized sequence. Structures with a z-score $\leq$ -8 are profoundly enriched in the native sequence, which might point to biological relevance of these structures. These are, however, extremes and most ncRNAs will have z-score values in a much higher range.

The value -2 for the z-score cutoff in this experiment was chosen arbitrarily. Moving to an even lower value than -2 will reduce the false discovery rate, but at the same time limit the number of ncRNAs that show such high thermodynamic stability. Using the `RNALfoldz` output from the experiment with randomized sequences (order-1 Markov model), we can calculate an empirical precision or positive predictive value (PPV), which is simply the

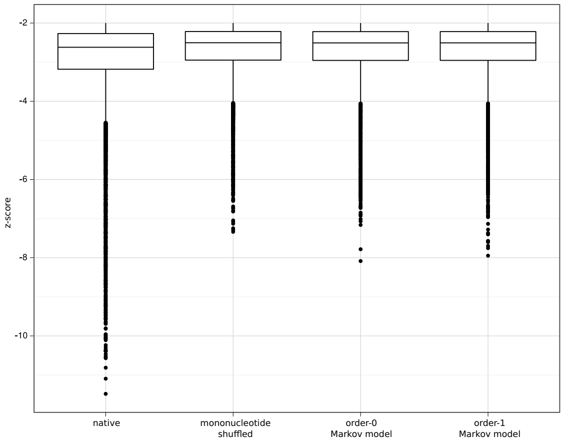


Figure 2: Comparison of the distribution of stable secondary structures from the native *E. coli* genome and randomized controls. The native *E. coli* sequence has a strong tendency to more stable local secondary structures. `RNALfoldz` predictions with a z-score below -8 are exclusively found in the native sequence.

proportion of true positives against all positive results. Assuming that thermodynamic stability is inherently linked to biologically function, we declare any `RNALfoldz` prediction with a z-score below a certain threshold from the native sequence and the randomized sequence as true positive and as false positive, respectively. Using then a PPV of 0.75, which corresponds to 25% estimated false positives, and, hence, a deduced z-score cutoff of -3.86 we can find 25 of the 106 annotated ncRNAs that are detectable with the `RNALfold` algorithm, while reducing the `RNALfoldz` to 21,715 predictions (3% of the `RNALfold` output). We further investigated if we can determine more specific z-score cutoffs when the `RNALfoldz` output is partitioned into different structural classes. This is motivated by the reasonable assumption that, for example, a short stable hairpin is more likely formed by chance than a stable, structurally more complex, multi-branched molecule. Hence, one would set different z-score cutoffs for different structural classes. To investigate this claim we map the MFE structures to the corresponding abstract RNA shape at the highest abstraction level. At this abstraction level only the helix nesting pattern is considered. As an example, the well-known cloverleaf structure of tRNA molecules is then simply represented as `[[][][]]`. The six major structural classes are shown in Tab. 1. We further partition structures according to their length into two classes *short* ($\leq$ 85 nt) and *long*.

Fig. 3 shows structure class specific precision values in dependency of the z-score, for those three classes that show the most deviation from the population precision. Using now class-specific z-score values when filtering the `RNALfoldz` output we can raise our prediction count from 25 to 38 ncRNAs, while keeping the expected false-positive rate fixed at 25%. The total number of `RNALfoldz` predictions increases slightly to 23,225.

Table 1: Major structural classes in the *E. coli* genome

frequency	abstract shape	length class	figure code	class specific z-score cutoff (PPV 0.75)
27%	[[][]]	long		-3.60
26%	[[][]]	short	SC2	-4.14
21%	[]	short	SC3	-4.16
7%	[[][[][]]]	long		-3.80
7%	[[[][]][]]	long		-3.74
4%	[]	long	SC6	-3.35
8%	rest			-3.35

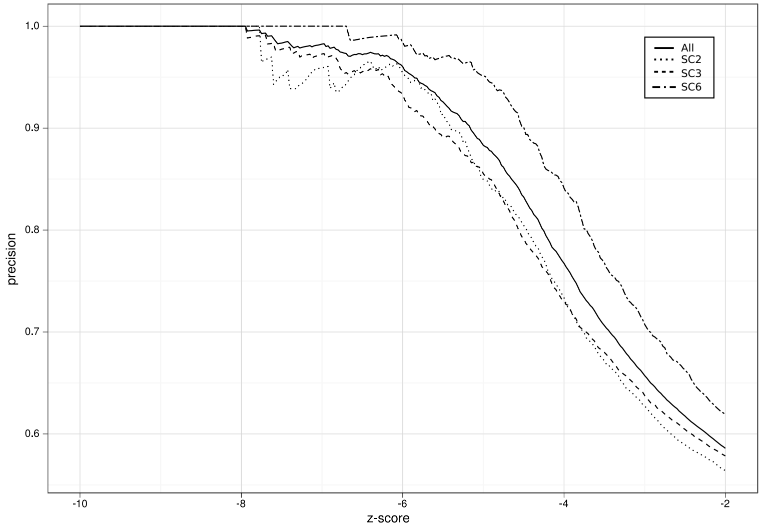


Figure 3: Precision values of different structural classes by the z-score. The solid line represents the whole RNALfoldz output.

3.2 Timing

The overall complexity $\mathcal{O}(N \times L^2)$ of the core algorithm does not change, the z-score calculation just adds a constant factor. We benchmarked both implementations on an Intel Quad Core2 CPU with 2.40 GHz. Detailed results are shown in Tab. 2.

At a maximal base-pair span of 120 nt RNALfold is able to scan at a speed of approx. 250 kb/min. At the same settings and with a minimal z-score cutoff of -2 scanning speed drops to 153 kb/min for RNALfoldz. The performance clearly depends on the number of times the support vector regression has to be called. When moving to a lower z-score cutoff of -4 the scanning speed increases to 207 kb/min.

Table 2: Timing results [sec] for RNALfold and RNALfoldz.

L	RNALfold	RNALfoldz		
		z-score $\leq$ -2	z-score $\leq$ -3	z-score $\leq$ -4
120	1,123	1,842	1,477	1,359
240	2,629	3,922	3,321	3,105

4 Discussion

In this work we have presented an extension of the RNALfold algorithm to predict thermodynamically stable, local RNA secondary structures. Using fast support vector regression models to calculate the z-score this approach comes with only a minor overhead in execution time compared to the former version, while yielding at the same time a much lower number of relevant structures. We have demonstrated that already with a z-score cutoff of -2, approx. 80% of the annotated *E. coli* small ncRNAs can be found in the RNALfoldz output. Comparison to randomized genome sequences showed that the native *E. coli* genome sequence has a strong bias to more stable secondary structures. This shift is, however, not significant enough to qualify RNALfoldz as a stand-alone RNA gene finder with an acceptable false discovery rate. We see the role of RNALfoldz mainly as a first filtering step in a cascade of computational ncRNA detection steps. In particular, the intersection of data from high throughput sequencing, promoter and transcription termination signals (see e.g. [SNS$^+$10]) or G+C content on AT rich genomes with RNALfoldz hits could be of value.

In this contribution, we assume that thermodynamic stability is inherently coupled to biological function. This is certainly true to some extent, but there are also a lot of RNA classes where stability is not a major issue for function, e.g. C/D box snoRNAs or ncRNAs that form interaction with other RNAs. It is therefore highly unlikely that these RNA classes will show up in the RNALfoldz output. In this context, RNALfoldz can, however, be used to define a set of highly stable loci which can then be excluded from further analysis.

It has been noted early on that thermodynamic stability alone is not a sufficient discriminator to distinguish ncRNAs from random sequences [RE00]. This is the main reason why most ncRNA gene finders rely solely on signals from evolutionary conservation of RNA secondary structures, or use thermodynamic stability only as an additional feature. These methods are clearly limited by the comparative genomics data available. Investigation of species that are distantly related to any species sequenced so far, or species specific RNA genes are, hence, out of scope for these methods. The RNALfoldz algorithm presented in this work will not be a magic tool suddenly shedding light on these dark areas. The search for extraordinarily stable structures, however, can help to give first clues to putatively functional RNA secondary structure elements, where other methods fail.

Acknowledgments

This work has been funded, in part, by the Austrian GEN-AU projects “bioinformatics integration network III” and “regulatory non-coding RNAs”. `RNALfoldz` is part of the Vienna RNA package and can be obtained free of charge at: `http://www.tbi.univie.ac.at/~ivo/RNA/`. An electronic supplement located at `http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/RNALfoldz` compiles training and test data.

References

[CFKK05] P Clote, F Ferré, E Kranakis, and D Krizanc. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, 11(5):578–591, 2005.

[CLS+90] J H Chen, S Y Le, B Shapiro, K M Currey, and J V Maizel. A computational procedure for assessing the significance of RNA secondary structure. *Comput Appl Biosci*, 6(1):7–18, 1990.

[FGM05] E Freyhult, P P Gardner, and V Moulton. A comparison of RNA folding measures. *BMC Bioinformatics*, 6:241–241, 2005.

[GVR04] R Giegerich, B Voss, and M Rehmsmeier. Abstract shapes of RNA. *Nucleic Acids Res*, 32(16):4843–4851, 2004.

[HFS+94] I L Hofacker, W Fontana, P F Stadler, L S Bonhoeffer, M Tacker, and P Schuster. Fast Folding and Comparison of RNA Secondary Structures. *Monatsh. Chem.*, 125:167–188, 1994.

[HHS08] J Hertel, I L Hofacker, and P F Stadler. SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, 24(2):158–164, 2008.

[HPS04] I L Hofacker, B Priwitzer, and P F Stadler. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, 20(2):186–190, 2004.

[HWL+09] Y Horesh, Y Wexler, I Lebenthal, M Ziv-Ukelson, and R Unger. RNAslider: a faster engine for consecutive windows folding and its application to the analysis of genomic folding asymmetry. *BMC Bioinformatics*, 10:76–76, 2009.

[JWW+07] P Jiang, H Wu, W Wang, W Ma, X Sun, and Z Lu. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res*, 35(Web Server issue):339–344, 2007.

[LLM02] S Y Le, W M Liu, and J V Maizel. A data mining approach to discover unusual folding regions in genome sequences. *Knowledge-Based Systems*, 15(4):243 – 250, 2002.

[LM89] S Y Le and J V Maizel. A method for assessing the statistical significance of RNA folding. *J Theor Biol*, 138(4):495–510, 1989.

[RE00] E Rivas and S R Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16(7):583–605, 2000.

[SNS+10] J Sridhar, S R Narmada, R Sabarinathan, H Y Ou, Z Deng, K Sekar, Z A Rafi, and K Rajakumar. sRNAscanner: a computational tool for intergenic small RNA detection in bacterial genomes. *PLoS One*, 5(8), 2010.

[WHS05] S Washietl, I L Hofacker, and P F Stadler. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A*, 102(7):2454–2459, 2005.

[WLX06] X F Wan, G Lin, and D Xu. Rnall: an efficient algorithm for predicting RNA local secondary structural landscape in genomes. *J Bioinform Comput Biol*, 4(5):1015–1031, 2006.

[ZS81] M Zuker and P Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–148, 1981.

Efficient sequence clustering for RNA-seq data without a reference genome

Florian Battke[1†*], Stephan Körner[1†], Steffen Hüttner[2], Kay Nieselt[1]
[1] *Center for Bioinformatics, University of Tübingen, Germany*
[2] *Hölle & Hüttner AG, Derendinger Str. 40, 72072 Tübingen, Germany*
[†] *equally contributing authors*

battke@informatik.uni-tuebingen.de

Abstract: New deep-sequencing technologies are applied to transcript sequencing (RNA-seq) for transcriptomic studies. However, current approaches are based on the availability of a reference genome sequence for read mapping. We present Passage, a method for efficient read clustering in the absence of a reference genome that allows sequencing-based comparative transcriptomic studies for currently unsequenced organisms. If the reference genome is available, our method can be used to reduce the computational effort involved in read mapping. Comparisons to microarray data show a correlation of 0.69, proving the validity of our approach.

1 Background

Changes in transcription are the most important mechanism of differentiation and regulation. Until recently, the transcriptional activity of a cell was measured by PCR in the case of few genes, or microarrays were used to investigate the whole transcriptome of an organism or tissue. Both methods require previous knowledge about the organism's transcripts, either in the form of ESTs or a complete reference genome sequence for primer resp. probe design. SAGE (serial analysis of gene expression) [VZVK95] is a method to study transcriptional activity based on sequencing of short transcript fragments. The advent of new *deep sequencing* technologies (also called next-generation or second generation sequencing methods) now allows to study the transcriptome in unprecedented detail by directly sequencing the pool of expressed transcripts. Using RNA-seq [WGS09] and a known reference genome, transcriptional activity can be measured with single-base precision.

Sequencing the pool of expressed transcripts creates millions of short (36-500 bases) sequences, called *reads*. These need to be mapped against the reference genome sequence allowing for mismatches due to sequencing errors or SNPs, which creates a huge computational challenge. Many

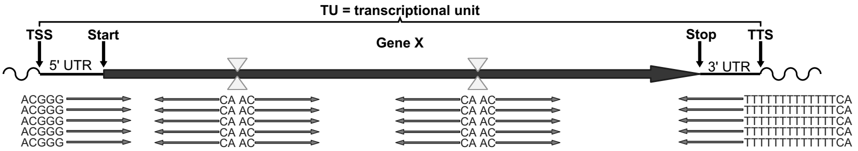


Figure 1: Schematic view of a transcriptional unit. RsaI restriction sites are indicated by yellow triangles. Transcript fragments (red) are sequenced starting from restriction sites in either direction, downstream from the transcription start site (TSS) as well as upstream from the transcription termination site (TTS), resulting in different sequence prefixes (GGG, CA, poly-T).

tools exists for that task, such as SOAP2 [LYL+09], MAQ [LRD08], VMatch [Kur03], RazerS [WER+09], and Bowtie [LTPS09]. Some programs are able to map reads covering splice junctions (TopHat [TPS09], QPALMA [DBOSR08]), others can map reads against several genomes at once, such as GenomeMapper [SHO+09]. Secondly, though more and more reference genomes are made available, the vast majority of organisms remain unsequenced and thus beyond the scope of RNA-seq studies.

Here we present PASSAGE [Hü09], extending the idea of SAGE to create a new efficient method for transcriptome studies in the absence of a reference genome sequence. It makes use of a newly established experimental protocol resulting in reads originating only from well-defined genomic positions. PASSAGE clusters these reads very efficiently to compute expression levels. Comparative studies can also be performed easily based on our method.

2 Material and Methods

Sample Preparation Purified mRNA is incubated with anchored Oligo ($\mathtt{dT}_{13}$) and modified SMART ($\mathtt{dG}_3$) oligonucleotides. These primers contain RsaI restriction sites. Reverse transcription is performed to obtain cDNA which is then amplified using long-distance PCR. After purification steps, the cDNA is cut into *transcript fragments* using the restriction enzyme RsaI. This step replaces the fragmentation step (e.g. by sonication) that is usually performed in RNA-seq protocols. Sequencing adapters are ligated to the fragments, and the fragments are analyzed by deep-sequencing. The universal primer site can be used for different sequencing techniques such as GS FLX$^{\mathrm{TM}}$ (Roche Diagnostics/454) and the Genome Analyzer$^{\mathrm{TM}}$ (Illumina). Barcode sequences can be included in the adapters to allow parallel sequencing of several samples. The resulting reads start with the

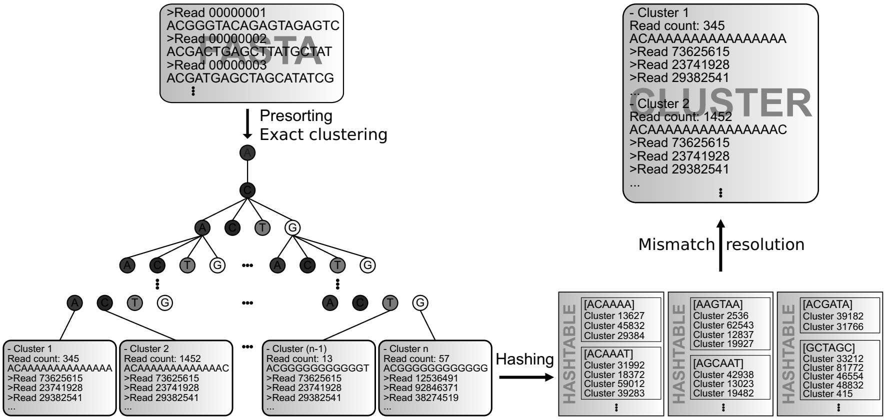


Figure 2: PASSAGE workflow: Reads for one sample (after optional presorting) are used to build a trie, resulting in a clustering of perfectly matching sequences. The reads' sequences are split into three parts and a hash map is built for each such part. These hash maps are used in the mismatch resolution step to cluster all reads with at most two mismatches, resulting in the final clustering output.

optional barcode, followed by a *prefix* and the genomic sequence. Three types of fragments can be distinguished by their prefixes: 5' UTR fragments start with `ACGGG`, 3' UTR fragments with $\texttt{ACT}_{13}$, and internal fragments with the restriction sequence `AC` (see figure 1). This protocol is adapted from [LRR+10] describing a 3'-fingerprint analyzed on a 2-D gel electrophoresis system to next-generation sequencing transcriptomics.

Cluster algorithm Our read clustering algorithm, PASSAGE, employs a three-step process. Starting from a fasta file containing read sequences, this involves *presorting*, *exact clustering* and *mismatch resolution*. The result is a file containing the number of reads contained in each cluster, the cluster's consensus sequence, the IDs of the reads, and a normalized expression estimate (reads per million reads). Result files from multiple samples can be joined into a tabular file containing one column per sample and one row per cluster, which can be analyzed with any standard microarray analysis software.

Presorting Reads are sequenced either from the 3', the 5' or the interior region of a transcriptional unit. This is reflected in different read *prefixes* (see figure 1). Differentiating the reads by these prefixes is not only useful for reducing computational costs, but rather to allocate the reads to the different parts of a transcriptional unit. With the addition of *barcode* sequences, different samples can be analysed in the same sequencing run, adding another common prefix for all reads of that sample. During presorting, reads are sorted according to their barcode (to distinguish different samples) and their prefix.

Exact clustering Based on the presorting result, each prefix is processed as follows. A trie of read sequences is generated such that reads are assigned to a common leaf if their sequences are identical. Since we know that reads either overlap by 100% or not at all, this effectively clusters all reads deriving from the same transcript fragment. Reads are placed into the tree by matching their bases one by one to the corresponding tree path until either a leaf is reached (and the read sequence is completely matched) or a new branch has to be created to accomodate the read's sequence. The result of the tree building step is a list of clusters, each cluster containing identical reads.

Mismatch resolution No current sequencing technology is error-free, thus we can not expect all reads from the same locus to be identical. In order to resolve this, we include a mismatch resolution step in our clustering algorithm. If k mismatches should be allowed, the minimal perfect match length results from equidistant distribution of these mismatches over the clusters' sequences. Thus we partition the clusters' sequences into $k + 1$ parts and create $k + 1$ hash maps. Clusters are placed into these hash maps according to the parts of their sequences. Thus, two clusters differing by at most k mismatches will be found in the same hash bucket in at least one of the hash maps. To ensure similar load factors in the presence of long common sequence prefixes, the first sequence part is slightly longer.

Clusters of identical reads are processed according to their size, starting with the largest cluster (in terms of the number of reads contained). From each hash map, candidate clusters are selected for merging. Ungapped alignments are computed and clusters are merged if their distance is at most k mismatches. Merged clusters are removed from all hash maps and the process is repeated until all clusters have been processed. Analysis showed that usually there is one very large and several smaller clusters for a given locus, and that the reads in the large cluster accurately represent the true genomic sequence. Thus we use the largest cluster's sequence as the consensus sequence for the joined cluster.

3 Results

We illustrate our method using the two closely related yeast species *Candida albicans* and *Candida dubliniensis.* Both are facultative pathogens,

			100pg	1μg
Dataset 1, 40bp	*C. albicans*	Y, 30°	4.1	† 4.6
		· YF, 37°	3.7	5.1
	C. dubliniensis	· HF, 37°	4.9	4.8
		YF, 37°	4.2	4.7
Dataset 2, 76bp	*C. dubliniensis*	· HF, 37°	7.6 / 7.6	∗3.8 / 5.6
		YF, 37°	4.4 / 5.5	∗6.1 / 8.6

Table 1: Conditions and number of reads (millions) sequenced for the two datasets. Y, yeast extract peptone dextrose, is a complete medium for yeast growth. F, fetal calf serum (10%). H, H_2O. The amount of total RNA used for sequencing was either 100pg or 1μg. The second dataset contains replicate sequencings. Hyphae-inducing conditions are marked with (·). Data used for comparison with other tools is indicated with (†), those used for validation using microarrays are marked with (∗).

C. albicans is of higher clinical importance as the most common agent causing candidosis. Both species have a genome size of about 14Mb organized in eight chromosomes and roughly the same number of genes (6185 in *C. albicans*, 5983 in *C. dubliniensis*). Cultures were grown under different conditions to study the induction of yeast or hyphae cell proliferation. RNA-seq data was generated from different amounts of total RNA and different read lengths (see table 1). In total, we analyzed 16 RNA-seq runs, using Passage with a maximum of two mismatches.

To assess the robustness of the protocol, we compared the two sequencings for each sample in dataset 1 by mapping the reads to all annotated genes (using Bowtie with up to two mismatches). More than 80% of the annotated genes found in the 100pg sample were also found in the 1μg sample, with the total number of transcripts being about twice as high in the 1μg samples (mean 4344 vs. 2247).

Data volume reduction Both clustering steps significantly reduce data volume (see figure 3). The efficiency of data reduction depends on the quality of the sequencing process. Fewer mismatches allow more reads to be clustered to their correct cluster and thus increase the reduction factor. In our studies using 16 different datasets (8 with 40-mer reads, 8 with 76-mer reads), exact clustering reduced data volume by about 84% (factor 6.1). Mismatch resolution results in a further reduction by about 58% (factor 2.4), resulting in a total reduction of about 93%. The reduction during perfect clustering can be seen as the result of summarizing the

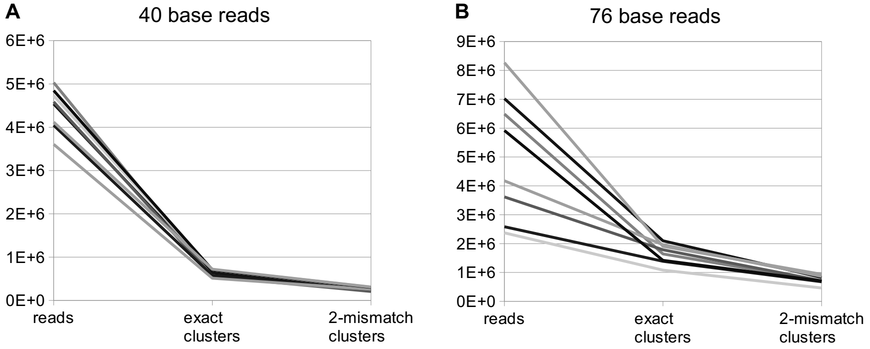


Figure 3: Reduction in data volume (number of reads resp. clusters) achieved by exact clustering and mismatch resolution. 16 datasets were analysed, eight of them with reads of length 40, eight with length 76. Longer reads (B) result in less reduction than shorter reads (A) due to higher error rates.

transcription strength (which varies between conditions) to the number of uniquely sequenced transcripts (which is expected to be more similar for all conditions). During the mismatch resolution step, a reduction is achieved by correcting for the error rate inherent in the sequencing technology, which should also be similar for all experiments.

Runtime analysis Presorting is important to reduce runtime and memory consumption and can be accomplished in $O(n)$ where n is the number of input reads. Exact clustering can also be done in $O(n)$. The time complexity of the mismatch resolution step depends on the average size of the hash buckets and the initial size of the cluster list: If c exact clusters are hashed randomly into buckets, let the average bucket size be ℓ. Merging the clusters can then be done in $O(\frac{c}{\ell} \cdot \ell^2)$. In the worst case, all clusters are hashed into one bucket, yielding $\ell = c$ and $O(c^2)$ runtime. The optimal case would be $\ell = 1$, yielding $O(c)$ runtime.

For real data, we see very small values for ℓ: We found $\ell = 2.5$ for reads of length 76 and $\ell = 4$ for reads of length 40. Thus, for the average case ℓ can be considered constant which results in a runtime of $O(c)$ (see figure 4A). Even in cases where a large number of clusters are collected in one hash bucket, we observe runtime linear with respect to the sum of sizes of the largest buckets in each hash map (see figure 4B). As c is bounded by the number of reads, n, overall average runtime is $O(n)$.

Comparison to other tools PASSAGE was written to cluster short reads and to use the size of the clusters as a measure of transcription.

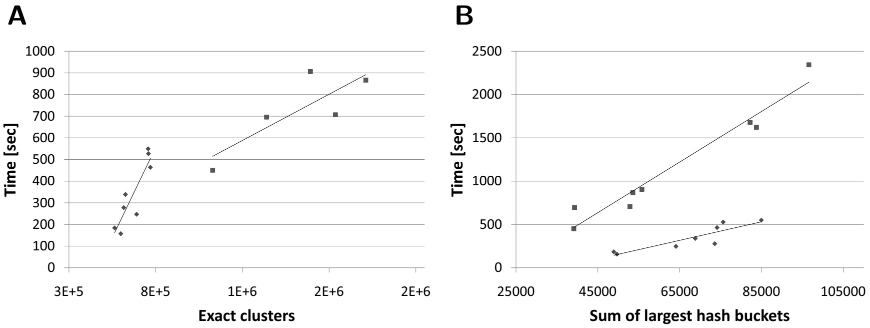


Figure 4: Runtime measurements for the mismatch resolution phase for reads of length 40 bases (diamonds) and 76 bases (squares), respectively, with linear regression curves. On average, runtime is linear in the number of exact clusters used as input (A). In the case of very uneven distribution of clusters to hash buckets, average runtime remains linear with respect to the sum of sizes of the largest hash bucket in each of the three hash maps (B). In both cases, we observe different slopes depending on the length of the reads.

It makes use of the fact that reads either overlap completely or not at all. We believe that no other tool currently offers the same functionality. However, in order to test our algorithm against other tools, we selected the EST assemblers Cap3 [HM99] and Mira3 [CWS99] as well as the short read de novo assemblers Velvet [ZB08] and Locas [KOS$^+$10]. As input we chose a FastA file with approx. 4.6 million Solexa reads of length 40 (184 million bases). Tests were performed on a computer with a 2.5Ghz processor and 8 GB of memory.

We used Bowtie to compute a direct read mapping against the genome of *C. albicans* (assembly 21, obtained from `www.candidagenome.org`) to get the number of "real" clusters. We allowed at most two mismatches (after removing the 3' and 5' prefixes from the respective reads). 3.97 million reads (86%) were mappable and were consequently used for the comparison. Bowtie mapped the reads to 49235 unique mapping positions, thus all methods that produce a significantly lower number of clusters (resp. contigs) combine expression measurements that should be kept separate.

Table 2 shows the runtime and number of assembled clusters for each program used here. It is important to note that these were written for generic assembly tasks while Passage is optimized for our biological protocol. We tried to set parameters such that the results would be most closely comparable to those obtained by Passage. While loading the reads, Cap3's memory consumption grew rapidly beyond the physical limit of our ma-

Program	CPU time	memory	clusters/contigs	*mcl*
Passage	1 min	1200 MB	44,817	40
Cap3[a]	–	>16000 MB	–	–
Mira3	12h 47 min	5600 MB	74,017	40.07
Locas	4h 38 min	5500 MB	2,650	40.08
Velvet	2 min	1300 MB	217	44.55
Bowtie[b]	6 min	21 MB	49,235[c]	–

Table 2: Runtimes and resulting number of clusters/contigs for all tested programs. *mcl*, mean consensus length. [a]Cap3 did not complete due to memory restrictions; [b]Bowtie requires a genomic sequence; [c]unique mapping sites.

chine. The program terminated after filling all available memory. Mira3, Velvet and Locas worked within the limits of our setup. Velvet runs very fast, producing only a very small number of contigs. These contigs are also too long on average, suggesting that it did too good a job of assembling mismatching reads and thus expression estimates derived from Velvet's output are combinations of the real expression values for different transcripts. Locas has a much higher runtime but produces more clusters with a better mean length, yet still far too few to produce correct expression estimates. Mira3's clusters are also close to the optimal length, but the program produces almost twice the number of clusters than Passage and its more generic approach to assembly is reflected in an extremely high runtime. These clusters have extremely vague consensus sequences with often more than 50% ambiguous bases $(r, y, s, w, k, m, b, v, d, h, n, *)$ which again suggests that different clusters have been merged that should have stayed separate.

Furthermore, most assemblers sacrifice specificity (in the detection of overlaps) for speed, while Passage is guaranteed to correctly cluster all reads with $\leq k$ mismatches to the assumed genomic sequence. Passage finds about 4500 clusters less than Bowtie because we do mismatch resolution without a reference genome, sometimes leading to the fusion of two very small clusters from distinct genomic positions with almost identical sequence.

Validation with microarrays We chose two experiments to validate the expression values computed using our approach with microarray data (see table 1). These samples were analyzed using a custom microarray with 50-mer probes for all *C. dubliniensis* ORFs (febit, Heidelberg). Two samples were analyzed using Passage and the resulting clusters were mapped to the *C. dubliniensis* genome using Bowtie. Of 5983 genes, 5144 (86%) genes could be analyzed. We only considered clusters with at least one read in each experiment, resulting in a list of 4377 (73.2%) genes. Fold-changes were computed independently for the microarray and the Passage data. First the fold-change between the two samples was computed for each cluster. The fold-changes of all clusters mapping to a common gene were averaged to obtain a fold-change value for each gene analyzed. The correlation between the fold-changes obtained from the microarray hybridizations and the Passage results was 0.69.

4 Discussion

We present a method for transcriptomic studies based on short RNA sequencing. It is especially useful in the absence of a reference genome. Reference sequences are only available for a tiny fraction of organisms, and while more and more genomes are sequenced, this still remains an issue for many research projects. Using a specialized protocol for the creation of the transcript pool, we greatly reduce the number of different read sequences and facilitate comparison between different samples. It effectively limits sequence overlaps to either complete or no overlap at all. Using this feature, our algorithm can rapidly cluster the reads and estimate expression for the corresponding transcripts in time linear to the number of read sequences.

A comparison to other tools shows that Passage is very fast and produces a sensible number of clusters, which allows to compute reliable expression estimates. We validated the expression levels computed using Passage with microarray data. Our method allows the application of well established software for comparative transcriptomics such as R [R D09] and Mayday [BSN10] to any (currently unsequenced) organisms and meta-transcriptomic samples. If a genome sequence is available, Passage clusters can be mapped against that reference to elucidate genomic locations as well as assign the short cluster to longer (predicted) transcripts. Here, our algorithm also reduces the computational effort necessary for mapping, due to the great reduction in the number of sequences that need to be mapped, thus meeting one of the great challenges of NGS technologies.

Acknowledgements
We are grateful to Kai Sohn at the Fraunhofer IGB for developing the protocol and conducting the biological experiments. Part of this project was funded by a grant of the Baden-Württemberg Ministry of Science.

References

[BSN10] F. Battke, S. Symons, and K. Nieselt. Mayday - Integrative analytics for expression data. *BMC Bioinformatics*, 11(1):121, 2010.

[CWS99] B. Chevreux, T. Wetter, and S. Suhai. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)*, 99:45–56, 1999.

[DBOSR08] F. De Bona, S. Ossowski, K. Schneeberger, and G. Rätsch. Optimal spliced alignments of short sequence reads. *BMC Bioinformatics*, 9(Suppl 10):O7, 2008.

[HM99] X. Huang and A. Madan. CAP3: A DNA sequence assembly program. *Genome Res*, 9(9):868–877, Sep 1999.

[Hü09] Hüttner, S. *Passage – Genexpressionsanalysen.* patent pending, AZ 102009058298.3, Dt. Patentamt München, 2009.

[KOS$^+$10] J.D. Klein, S. Ossowski, K. Schneeberger, D. Weigel, and D.H. Huson. LOCAS – A low coverage assembly tool for resequencing projects. *(manuscript in preparation)*, 2010.

[Kur03] S. Kurtz. The Vmatch large scale sequence analysis software. *Ref Type: Computer Program*, pages 4–12, 2003.

[LRD08] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*, 18(11):1851, 2008.

[LRR$^+$10] E. Lindemann, B. Rohde, S. Rupp, J. Regenbogen, and K. Sohn. A multidimensional electrophoretic system of separation for the analysis of gene expression (Message). *Electrophoresis*, 31(8):1330–1343, 2010.

[LTPS09] B. Langmead, C. Trapnell, M. Pop, and S. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25, 2009.

[LYL$^+$09] R. Li, C. Yu, Y. Li, T.W. Lam, S.M. Yiu, K. Kristiansen, and J. Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966, 2009.

[R D09] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2009.

[SHO$^+$09] K. Schneeberger, J. Hagmann, S. Ossowski, N. Warthmann, S. Gesing, O. Kohlbacher, and D. Weigel. Simultaneous alignment of short reads against multiple genomes. *Genome Biol*, 10:R98, 2009.

[TPS09] C. Trapnell, L. Pachter, and S.L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105, 2009.

[VZVK95] V.E. Velculescu, L. Zhang, B. Vogelstein, and K.W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484, 1995.

[WER$^+$09] D. Weese, A.K. Emde, T. Rausch, A. Döring, and K. Reinert. RazerS—fast read mapping with sensitivity control. *Genome Res*, 19(9):1646, 2009.

[WGS09] Z Wang, M Gerstein, and M Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009.

[ZB08] D.R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, 18(5):821, 2008.

Uncovering the structure of heterogeneous biological data: fuzzy graph partitioning in the k-partite setting

Florian Blöchl[1,2], Mara L. Hartsperger[1,2,*], Volker Stümpflen[1], Fabian J. Theis[1]

[1]Institute for Bioinformatics and Systems Biology, Helmholtz Zentrum München
[2]Equal contributors

Abstract: With the increasing availability of large-scale interaction networks derived either from experimental data or from text mining, we face the challenge of interpreting and analyzing these data sets in a comprehensive fashion. A particularity of these networks, which sets it apart from other examples in various scientific fields lies in their k-partiteness. Whereas graph partitioning has received considerable attention, only few researchers have focused on this generalized situation. Recently, Long et al. have proposed a method for jointly clustering such a network and at the same time estimating a weighted graph connecting the clusters thereby allowing simple interpretation of the resulting clustering structure. In this contribution, we extend this work by allowing fuzzy clusters for each node type. We propose an extended cost function for partitioning that allows for overlapping clusters. Our main contribution lies in the novel efficient minimization procedure, mimicking the multiplicative update rules employed in algorithms for non-negative matrix factorization. Results on clustering a manually annotated bipartite gene-complex graph show significantly higher homogeneity between gene and corresponding complex clusters than expected by chance. The algorithm is freely available at `http://cmb.helmholtz-muenchen.de/fuzzyclustering`.

1 Introduction

With the relatively cheap availability of biological high-throughput methods such as microarrays, machine learning techniques gain more and more importance in the field of bioinformatics. Learning approaches often focus on the analysis of homogeneous data sets that can be represented as a network having vertices of a single type only. However, many real-world networks are heterogeneous and involve objects of multiple, related types, thus forming k-partite graphs consisting of diverse types of vertices. A key question of clustering-based approaches is how to interpret the global organization of these networks as the coexistence of their structural subunits associated with more highly interconnected parts. Identifying these a priori unknown building blocks such as for instance the common genetic origin of different diseases is crucial for the understanding of the structural and functional properties of such networks.

Most available clustering methods cannot be applied to k-partite networks because they do not treat the single node types (partitions) separately and therefore do not represent the global community structure correctly. While this has been addressed in terms of algorithms for some time now [Bar07, GL04, KAKS97, ZHS07, LWZY06], not many possible applications exist yet in bioinformatics, although the field commonly deals with such networks [KHT09]. A particular issue that may hamper application to bioinformatics may be that most existing algorithms identify separated, disjoint clusters by assigning each data point to exactly one cluster [Mac67, JD88], whereas most biological networks consist of highly overlapping cohesive groups of vertices. A single data point can therefore belong to more than only one cluster, e.g. a large fraction of proteins belong to several protein complexes simultaneously [RBDK$^+$08]. So far only a few approaches exist that allow the detection of overlapping clusters by assigning either each data point a degree of belonging to clusters or to several clusters respectively [Bez81, PDFV05].

In order to identify clusters in heterogeneous data and moreover connect these clusters between the different node types, we developed a fuzzy partitional clustering method based on a non-negative matrix factorization (NMF) model [LS99]. We demonstrate that we can identify biological meaningful overlapping clusters in k-partite graphs. We applied our method to a bipartite gene-protein complex graph representing the manually annotated Corum core set [RBDK$^+$08]. The extracted clusters show significantly higher homogeneity between gene and corresponding complex clusters than expected by chance.

2 A multiplicative update rule for fuzzy k-partite clustering

Recently, an algorithm for the partitioning of k-partite graphs has been put forward in [LWZY06]. It clusters each node set of the graph separately; then the clusters are connected via a smaller, weighted k-partite graph. The algorithm consists of an alternating minimization procedure: first the nodes in each layer are clustered in order to minimize the distance to the small representative graph (change). Then the hidden graph (backbone graph) is updated according to the same cost function.

A key assumption made in [LWZY06] is that the assignment in the first step is made in a binary fashion. This hard clustering is a feature that often is achieved by soft clustering algorithms when not forcing explicit cluster overlap [Bez81]. However it can be easily seen that the cost function put forward in [LWZY06] is not fully minimized by this approximation.

Here, we address the minimization using a multiplicative update algorithm. In contrast to the above method, by not choosing any binary assignment a priori, we observe a close to binary assignment mostly in the hidden nodes, whereas the clustering in each node-type is soft. The resulting algorithm is similar in structure to multiplicative algorithms for NMF, with the difference that we address a three-matrix factorization problem, see e.g. [DS06], and have to deal with a multi-summand cost function.

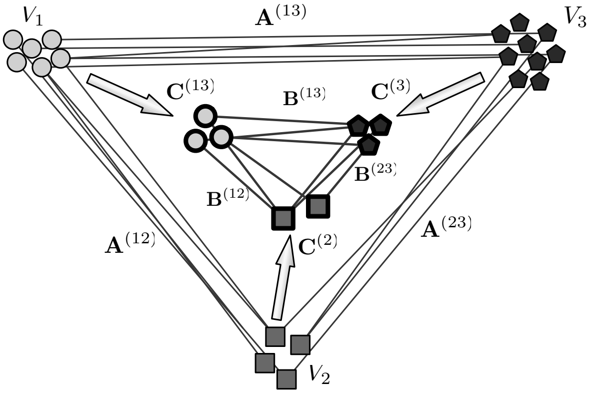


Figure 1: (a) definition of a 3-partite graph G with notation used. (b) approximation of G using a smaller 3-partite graph H defined on fuzzy node clusters.

2.1 Definitions and factorization model

A k-partite graph is a graph $G = (V, E)$ and a partition of the vertices V into k disjoint sets V_i such that no two vertices in the same subset are adjacent. So edges are only allowed between different subsets ('colors'). Let $n_i := |V_i|$ be the number of vertices in partition i. We represent the graph as a set of $n_i \times n_j$ matrices $\mathbf{A}^{(ij)}$ with $1 \leq i < j \leq k$. Commonly, each matrix element is either 0 or 1, but we only restrict the matrices to have non-negative coefficients thereby allowing weighted graphs as well. We can readily include directed instead of undirected k-partite graphs by specifying incidence matrices also for $i > j$. It is easy to see that the following cost function and optimizations generalize to this situation.

We want to approximate G by a smaller cluster network H (*backbone network*), which is defined on fuzzy clusters of each G-partition V_i. For simplicity we for now fix the number of V_i-clusters to m_i. We say that a non-negative $n_i \times m_i$-matrix $\mathbf{C}^{(i)}$ is a *fuzzy clustering* of V_i, if it is right-stochastic i.e. $\sum_l c_{kl}^{(i)} = 1$ for all k. Then we search for a k-partite graph H with $m_i \times m_j$ incidence matrices $\mathbf{B}^{(ij)}$ and fuzzy clusterings $C := (\mathbf{C}^{(i)})_{i=1,\dots,k}$ such that the connectivity explained by H is as close as possible to G after clustering.

We can measure this difference in many different ways. In [LWZY06], this choice has been circumvented by specializing on arbitrary Bregman divergences, which still allow efficient reformulation of gradient-type algorithms without knowing the specific formula. This is also possible in our case of multiplicative update rules, as has been shown for NMF in [DS06]. However, for simplicity, we choose the minimum square distance as cost function. This implies minimization of

$$f(H, C) := \sum_{i<j} \left\| \mathbf{A}^{(ij)} - \mathbf{C}^{(i)}\mathbf{B}^{(ij)}(\mathbf{C}^{(j)})^\top \right\|_F^2 \tag{1}$$

where $\|.\|_F^2$ denotes the squared Frobenius norm, i.e. the square sum of the matrix elements. The model, the used definitions and the approximation are illustrated in figure 1.

2.2 Derivation of the algorithm

We want to minimize $f(H, C)$ from (1) using a local algorithm extending gradient descent. We assumed an undirected k-partite graph, so $\mathbf{A}^{(ij)}$ is undefined for $i > j$. Hence, we now set $\mathbf{A}^{(ij)} := (\mathbf{A}^{(ji)})^\top$ for $i > j$ (and similarly for $\mathbf{B}^{(ij)}$). Then we find

$$\begin{aligned} \frac{\partial f}{\partial b_{rs}^{(ij)}} &= -2\left((\mathbf{C}^{(i)})^\top \mathbf{A}^{(ij)} \mathbf{C}^{(j)} - (\mathbf{C}^{(i)})^\top \mathbf{C}^{(i)} \mathbf{B}^{(ij)} (\mathbf{C}^{(j)})^\top \mathbf{C}^{(j)}\right)_{rs} \\ \frac{\partial f}{\partial c_{rs}^{(i)}} &= -2\sum_{j \neq i}\left(\mathbf{A}^{(ij)} \mathbf{C}^{(j)} (\mathbf{B}^{(ij)})^\top - \mathbf{C}^{(i)} \mathbf{B}^{(ij)} (\mathbf{C}^{(j)})^\top \mathbf{C}^{(j)} (\mathbf{B}^{(ij)})^\top\right)_{rs} . \end{aligned}$$

Minimizing f by alternating gradient descent, we now simply start from an initial guess of $\mathbf{B}^{(ij)}, \mathbf{C}^{(i)}$ and alternate between updates of the $\mathbf{B}^{(ij)}$ and the $\mathbf{C}^{(i)}$ with according learning rates. Such update rules however have two disadvantages: for one, the choice of update rate η (possibly different for $\mathbf{B}$, $\mathbf{C}$ and i, j) is unclear; in particular, for too small η convergence may take too long or may not be achieved at all, whereas for too large η we may easily overshoot the minimum. Moreover, the resulting matrices may become negative. Therefore, we follow Lee and Seung's idea for NMF [LS99] and rewrite this into multiplicative update rules. Hence, let us choose update rates

$$\begin{aligned} \eta_{rs}^{(ij)} &:= \frac{b_{rs}^{(ij)}}{2\left((\mathbf{C}^{(i)})^\top \mathbf{C}^{(i)} \mathbf{B}^{(ij)} (\mathbf{C}^{(j)})^\top \mathbf{C}^{(j)}\right)_{rs}} \\ \eta_{rs}^{(i)} &:= \frac{c_{rs}^{(i)}}{2\left(\sum_{j \neq i} \mathbf{C}^{(i)} \mathbf{B}^{(ij)} (\mathbf{C}^{(j)})^\top \mathbf{C}^{(j)} (\mathbf{B}^{(ij)})^\top\right)_{rs}} \end{aligned}$$

Plugging this into the gradient descent equations, this results in the desired multiplicative update rules

$$b_{rs}^{(ij)} \leftarrow b_{rs}^{(ij)} \frac{\left((\mathbf{C}^{(i)})^\top \mathbf{A}^{(ij)} \mathbf{C}^{(j)}\right)_{rs}}{\left((\mathbf{C}^{(i)})^\top \mathbf{C}^{(i)} \mathbf{B}^{(ij)} (\mathbf{C}^{(j)})^\top \mathbf{C}^{(j)}\right)_{rs}} \tag{2}$$

$$c_{rs}^{(i)} \leftarrow c_{rs}^{(i)} \frac{\left(\sum_{j \neq i} \mathbf{A}^{(ij)} \mathbf{C}^{(j)} (\mathbf{B}^{(ij)})^\top\right)_{rs}}{\left(\sum_{j \neq i} \mathbf{C}^{(i)} \mathbf{B}^{(ij)} (\mathbf{C}^{(j)})^\top \mathbf{C}^{(j)} (\mathbf{B}^{(ij)})^\top\right)_{rs}} \tag{3}$$

2.3 Algorithm formulation and relation to other work

We note that we can readily show that these update rules do not increase the cost function (1). This can be shown via auxiliary functions similar to NMF [LS01] and multi-factor NMF [DS06], which has been applied in a related model for co-clustering of microarray data [CDGS04]. This theoretical result implies convergences of the update rules. However in contrast to early statements in NMF [LS01], this does not necessarily imply convergence

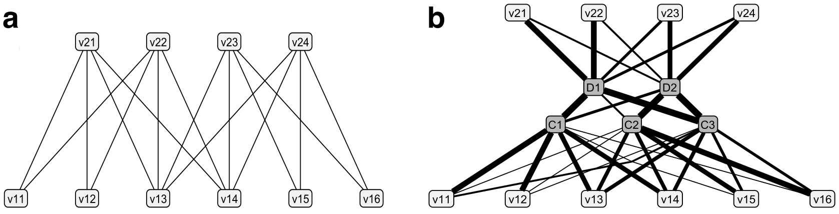


Figure 2: Toy example of a bipartite graph (a) from [LWZY06], with its backbone network and fuzzy clusters (b). Note that neither of the two clusterings are binary.

to stationary points of the Euclidean norm (zero of the differential from (1)), since the update steps may be too small to reach those points. Another possible drawback of such multiplicative updates is the fact that once a matrix entry has been set to zero (which may happen due to zeros in $\mathbf{A}^{(ij)}$ or to numerics), the coefficient will never then be able to become positive again during learning.

We have not yet taken into account the constraint that the cluster matrices $\mathbf{C}^{(i)}$ are required to be right-stochastic i.e. $\mathbf{C}^{(i)}\mathbf{e} = \mathbf{e}$ for $\mathbf{e} = (1, \ldots, 1)$. For simplicity, we force this constraint by regularly projecting each row of $\mathbf{C}^{(i)}$ onto the sphere of the 1-norm. Alternatively, we may introduce this constraint as Lagrange parameter, and get modified cost function with weighted Lagrange parameters. We can still prove non-increasingness of the multiplicative update rule along the lines of [DS06]. The final fuzzy k-partite clustering algorithm is summarized in algorithm 1. An implementation is freely available at `http://cmb.helmholtz-muenchen.de/fuzzyclustering`. In figure 2, we illustrate the feasibility of the algorithm on a small bipartite toy example.

Our algorithm contains two nested loops over the number of partitions. The update steps are fully vectorized and contain only matrix products of non-square matrices. The total time complexity of the algorithm can therefore be estimated as

$$\#\text{iterations} \times \mathcal{O}(k^2\, n_{\text{max1}} n_{\text{max2}}\, m_{\text{max}})\,. \tag{4}$$

Here, n_{max1} and n_{max2} denote the sizes of the largest and the second-largest partition, m_{max} is the maximum number of clusters to extract within any partition. Hence, the algorithm is fast and efficient. The runtime is linear in each partition size and grows only quadratic in the total number of nodes in the case of graphs with similarly large partitions.

In order to extend cost functions in (unipartite) data clustering to include fuzzy clusters, commonly a so-called *fuzzification factor* $m > 1$ is introduced [Bez81,Dun73]. Instead of squared norm minimization of the residuals, a higher residual power is minimized, which results in overlapping non-trivial cluster assignments. However, we will find that already the standard case $m = 1$ may suffice to introduce non-trivial overlapping clusters. This is because we are interested in co-clustering, which is different from standard data clustering where only a unipartite graph and hence $\mathbf{C}^{(i)} = \mathbf{C}^{(1)}$ is assumed.

Algorithm 1: fuzzy k-partite clustering

```
Input: $k$-partite graph $G$ with possibly non-negatively weighted edge matrices $\mathbf{A}^{(ij)}$, $i < j$, number of clusters $m_1, \dots, m_k$
Output: fuzzy clustering $\mathbf{C}^{(i)}$ and $k$-partite cluster graph $H$ given by matrices $\mathbf{B}^{(ij)}$

Initialize $\mathbf{C}^{(i)}, \mathbf{B}^{(ij)}$ to random non-negative matrices.
Normalize $c^{(i)}_{rs} \leftarrow c^{(i)}_{rs} / (\sum_t c^{(i)}_{rt})$ for all $i, r, s$
repeat
    update fuzzy clusters
    for $i \leftarrow 1, \dots, k$ do
        $\mathbf{C}^{(i)} \leftarrow \mathbf{C}^{(i)} \otimes (\sum_{j \neq i} \mathbf{A}^{(ij)} \mathbf{C}^{(j)} \mathbf{B}^{(ij)\top}) \oslash (\sum_{j \neq i} \mathbf{C}^{(i)} \mathbf{B}^{(ij)} \mathbf{C}^{(j)\top} \mathbf{C}^{(j)} \mathbf{B}^{(ij)\top})$
        Normalize $c^{(i)}_{rs} \leftarrow c^{(i)}_{rs} / (\sum_t c^{(i)}_{rt})$ for all $r, s$
    end
    update k-partite cluster graph H
    for $i \leftarrow 1, \dots, k-1$ do
        for $j \leftarrow i+1, \dots, k$ do
            $\mathbf{B}^{(ij)} \leftarrow \mathbf{B}^{(ij)} \otimes (\mathbf{C}^{(i)\top} \mathbf{A}^{(ij)} \mathbf{C}^{(j)}) \oslash (\mathbf{C}^{(i)\top} \mathbf{C}^{(i)} \mathbf{B}^{(ij)} \mathbf{C}^{(j)\top} \mathbf{C}^{(j)})$
        end
    end
until convergence;
Note: $\otimes$ and $\oslash$ symbolize elementwise multiplication and division, respectively.
```

3 Fuzzy clusters and backbone of a gene-complex hypergraph

In order to illustrate the applicability of our method to heterogeneus biological data we employ the Corum core set [RBDK$^+$08] that reflects a non-redundant catalogue of experimentally verified mammalian protein complexes manually annotated at MIPS. A bipartite graph $G = (V, E)$ with $|V| = 4877$ and $|E| = 8738$ was constructed from these data. The two disjoint node sets are represented by protein complexes and their associated genes further referred to as V_c and V_g, respectively. We then focused on a reduced data set G' with $|V'| = 4090$ and $|E'| = 7946$ retrieved by extracting the maximally connected subgraph. The remaining graph consisted of 1728 complex (V_c) and 2362 gene (V_g) vertices.

The determination of the number of clusters for each node type, in which the graph has to be decomposed, is difficult, and even in the case of unipartite k-means does not allow a direct and computationally simple answer. To address this issue we approximated the number of clusters to be found in the complex and the gene partition respectively by limiting the maximal number of clusters k_c for V_c according to $k_c = \lfloor \sqrt{|V_c|/2} \rfloor$, and then scaled the number of clusters k_g for V_g by $k_g = \lceil k_c \sqrt{|V_g|/|V_c|} \rceil$. We calculated the value of the cost function for each pairwise combination starting from k_c=1. Due to random initial conditions, the algorithm is inherently indeterministic. Therefore, we discuss performance over 10 runs each. Figure 3(a) shows the distribution of cost func-

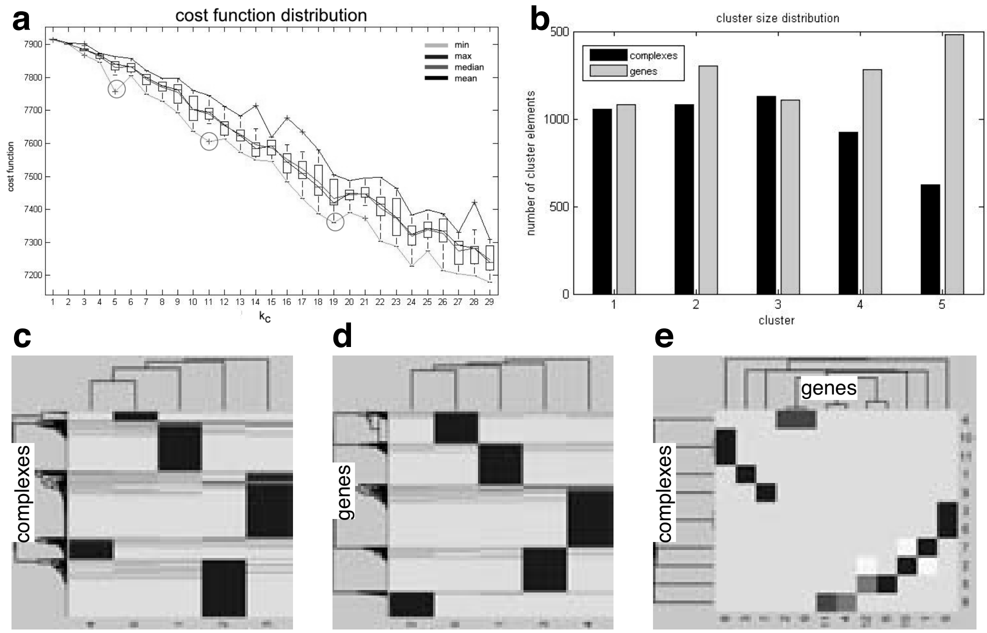


Figure 3: (a) Approximation of cluster numbers k_c, k_g. (b) Distribution of cluster sizes for $k_c = 5, k_g = 5$. Hierarchical clustering of (c) complex and (d) gene clusters (see fig 4(c) for backbone network for $k_c, k_g = 5$). The clustered backbone for $k_c = 11, k_g = 12$ is shown in (e).

tion values for the particular parameter settings. As final parameters k_c and k_g we chose $(k_c, k_g) \in \{(5,5), (11,12), (19,22)\}$, where we observe significant drops of the cost function. With this, we detect organizational structures on different levels of resolution. In the following we will mostly discuss the smallest graph with 5 clusters each (see figure 3(b)).

Figure 3 shows that our method is able to identify overlapping clusters. In the resulting five clusters, the majority of elements is assigned to a single cluster. However, there exists a considerable amount of nodes assigned to several clusters simultaneously, see figures 3(c,d). Almost ten percent of complexes (193) and genes (187) are assigned to two clusters with $p >= 0.3$. For instance, the genes *ITGB2* and *MCRS1* are even part of threes clusters with $p >= 0.3$. This clearly demonstrates the need for a fuzzy approach. The clusters vary strongly in size (figure 3(b)). and their interconnectivity is sparse, see figure 4(c). However, in the case of k_c=11 and k_g=12 we already have a resolution level that is fine enough to see details, and several binary clusters become apparent (figure 3(e)).

In order to evaluate whether both the extracted clusters and their interconnections given by the backbone graph are biologically feasible, we employed FunCat classifications. For all genes we mapped Gene Ontology associations to their according FunCat categories to achieve comparability between the node types (http://mips.gsf.de/proj/funcatDB/). Usually, complexes and genes are annotated with the lowest FunCat category or GO term respectively. In our analysis we took a subset of 13 FunCat main categories. Subcategory annotations were mapped to the according main category terms for consistency reasons.

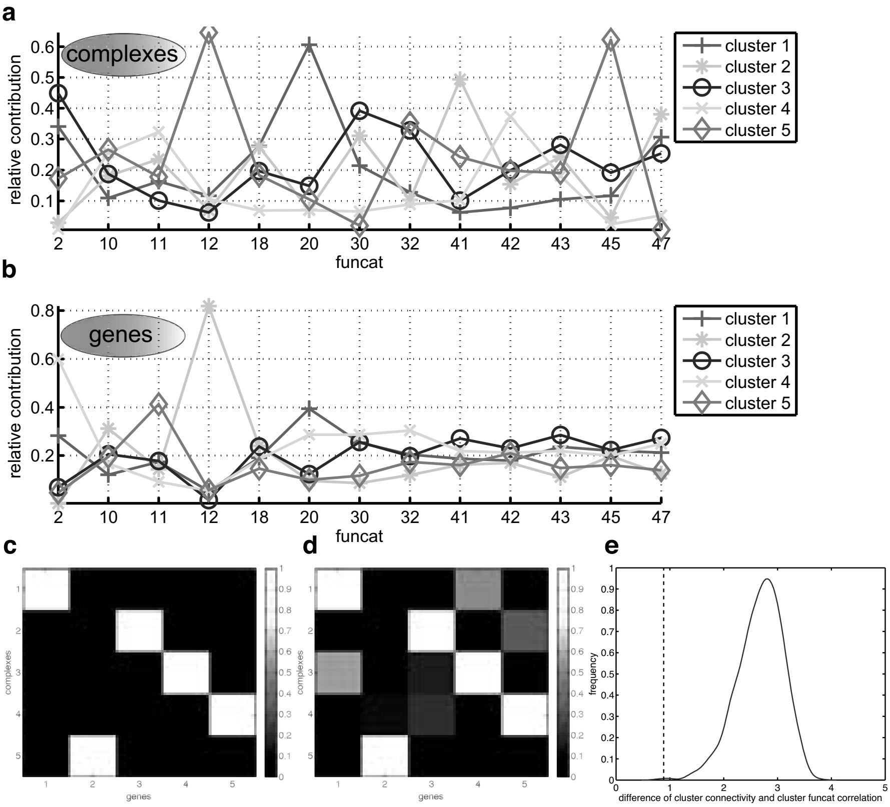


Figure 4: (a,b) FunCat annotation profile for complex and gene clusters. (c) shows the normalized backbone connectivity, and (d) the normalized positive crosscorrelations of the FunCat profiles from (a) and (b). (e) Shows statistics over 1000 random networks, proving significance of the clusters (dashed line) with a p-value of $p < 10^{-3}$.

From figure 4(a) and (b) we see that the extracted clusters can be easily interpreted biologically, as most of them have a high fraction of functional annotations with a certain FunCat term. Moreover, from visual comparison, see figure 4, we see that interconnected clusters also seem to be functionally correlated. In order to quantify this, we determined for each cluster how it is associated with each of the 13 FunCat categories by weighting a cluster elements FunCat classification by its degree of membershipto the particular cluster and calculated Pearson correlation of FunCat annotations of the complex and gene clusters. As expected, we find a high similarity between the clusters interconnectivity and their functional correlation. This shows that our fuzzy partitioning approach yields biologically meaningful results by identifying functionally related clusters.

To evaluate the significance of these results we compared our findings with the results of a random model. Assuming that a random network does not form functionally related clusters, we applied a bipartite randomization procedure to our original network. We generalized the degree-preserving rewiring for complex networks, first introduced by Maslov and Sneppen [MS02]: In every randomization step we randomly picked two edges and exchanged their endpoints of one type (either proteins or complexes) without creating multiple edges or self-loops. This rewiring procedure leads to a loss of degree-correlations between first and second neighbors. Hence, one can observe the degree of randomization by the course of these quantities over the process. This also tells us how many randomization steps are needed. In practice, degree-correlations vanished after around one randomization step per edge. So, for our analyses we used five times this number as in [WAH$^+$08].

We determined the clusters' FunCat profiles and calculated normalized positive correlations. To have a distance measure, we calculated the difference between the normalized backbone connectivity and the normalized positive cross-correlation matrix. Comparing these distances to clusterings using the hard approach from [LWZY06], we found much smaller values. As an example, a histogram is shown in figure 4(e), which illustrates that out of 1000 iterations only a single random entry is smaller than the 0.89, resulting in a p-value $< 10^{-3}$. This shows the significance of our results.

4 Conclusion

In this contribution, we presented a novel computationally efficient and scalable graph partitioning algorithm. Unlike other methods in the field it allows for the identification of overlapping clusters in k-partite graphs of heterogeneous data. It is based on an efficient minimization procedure, mimicking the multiplicative update rules employed in algorithms for non-negative matrix factorization. We verified our approach on a bipartite network of protein complexes where we demonstrated that we successfully identified functionally correlated clusters.

Partitioning on a local level, i.e. aiming at detecting quite small clusters, our algorithm will enable reclassification, annotation or even detection of misclassified elements in heterogeneous data sets. Partitioning into large-scale clusters, we focus on understanding their global organization. For instance, simple bipartite graph analysis has recently brought insights into the organization of microRNA interactions [RKS$^+$10]. At the moment, we extend this work by integrating predictions of microRNA target sites with protein complexes, disease information and different types of annotations.

Acknowledgements. This work was partially supported by the Helmholtz Alliance on Systems Biology (project 'CoReNe') and the Federal Ministry of Education and Research (BMBF) in its MedSys initiative (project 'SysMBO').

References

[Bar07] M.J. Barber. Modularity and community detection in bipartite networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 76(6 Pt 2):066102, Dec 2007.

[Bez81] J.C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algoritms*. Plenum Press, New York, 1981.

[CDGS04] H. Cho, I.S. Dhillon, Y. Guan, and S. Sra. Minimum Sum Squared Residue based Co-clustering of Gene Expression data. In *Proc. SIAM International Conference on Data Mining*, pages 114–125, 2004.

[DS06] I.S. Dhillon and S. Sra. Generalized Nonnegative Matrix Approximations with Bregman Divergences. In *Proc. NIPS 2005*, 2006.

[Dun73] J.C. Dunn. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, 3:32–57, 1973.

[GL04] J. Guillaume and M. Latapy. Bipartite Structure of All Complex Networks. *Information Processing Letters*, 90(5):215–221, 2004.

[JD88] Anil K Jain and R.C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988.

[KAKS97] G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar. Multilevel hypergraph partitioning: application in VLSI domain. In *Proc. DAC '97*, pages 526–529. ACM Press, 1997.

[KHT09] S. Klamt, U. Haus, and F.J. Theis. Hypergraphs and cellular networks. *PLoS Computational Biology*, 5(5), 2009.

[LS99] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative Matrix Factorization. *Nature*, 40:788–791, 1999.

[LS01] D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. In *Proc. NIPS 2000*, volume 13, pages 556–562. MIT Press, 2001.

[LWZY06] B. Long, X. Wu, Z. Zhang, and P.S. Yu. Unsupervised Learning on K-partite Graphs. In *Proc. SIGKDD 2006*, pages 317–326, 2006.

[Mac67] J. B. MacQueen. Some Methods for Classification and Analysis of MultiVariate Observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[MS02] Sergei Maslov and Kim Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, May 2002.

[PDFV05] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, Jun 2005.

[RBDK$^+$08] A. Ruepp, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, M. Stransky, B. Waegele, T. Schmidt, O. Noubibou Doudieu, V. Stümpflen, and H.W. Mewes. CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res*, 36(Database issue):D646–D650, Jan 2008.

[RKS$^+$10] Andreas Ruepp, Andreas Kowarsch, Daniel Schmidl, Felix Buggenthin, Barbara Brauner, Irmtraud Dunger, Gisela Fobo, Goar Frishman, Corinna Montrone, and Fabian J. Theis. PhenomiR: a knowledgebase for microRNA expression in diseases and biological processes. *Genome biology*, 11(1):R6+, January 2010.

[WAH$^+$08] P Wong, S Althammer, A Hildebrand, A Kirschner, P Pagel, B Geissler, P Smialowski, F Bloechl, M Oesterheld, T Schmidt, N Strack, FJ Theis, A Ruepp, and D Frishman. An evolutionary and structural characterization of mammalian protein complex organization. *BMC Genomics*, 9(1):629, Dec 2008.

[ZHS07] D. Zhou, J. Huang, and B. Schoelkopf. Learning with Hypergraphs: Clustering, Classification, and Embedding. In *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

Shape-based barrier estimation for RNAs

Sergiy Bogomolov[1,◇], Martin Mann[2,◇], Björn Voß[3], Andreas Podelski[1]
and Rolf Backofen[2]
[1]*Software Engineering,* [2]*Bioinformatics,* [3]*Genetics and Experimental Bioinformatics at Albert-Ludwigs-Universität Freiburg, Germany*
◇ These authors contributed equally to this work
{bogom,mmann}@informatik.uni-freiburg.de

Abstract:
The ability of some RNA molecules to switch between different metastable conformations plays an important role in cellular processes. In order to identify such molecules and to predict their conformational changes one has to investigate the refolding pathways. As a qualitative measure of these transitions, the barrier height marks the energy peak along such refolding paths. We introduce a meta-heuristic to estimate such barriers, which is an NP-complete problem. To guide an arbitrary path heuristic, the method uses RNA shape representative structures as intermediate checkpoints for detours. This enables a broad but efficient search for refolding pathways. The resulting Shape Triples meta-heuristic enables a close to optimal estimation of the barrier height that outperforms the precision of the employed path heuristic.

1 Introduction

RNA plays a central role in living cells. Numerous RNAs are able to switch between different structures within their life time due to thermodynamics, temperature changes (thermometers), ligand binding (riboswitches) or other signals [FHMS$^+$01]. Such multistable RNAs regulate gene expression directly or are connected to regulatory mechanisms, e.g. splicing [LC93]. For the correct prediction and study of such structural changes it is necessary to identify the lowest energy refolding pathway in the underlying RNA energy landscape. The energy barrier height surmounted along such paths can be used to estimate refolding probabilities [GFW$^+$08] or to study the kinetics of the folding process [WSSF$^+$04].

Maňuch *et al.* have shown that the calculation of the exact barrier height is a hard, NP-complete problem for RNA secondary structure landscapes [MTSC09]. Therefore, exact approaches rely on the full enumeration of the low energy parts of the landscape [SvdPS99, FHSW02, KH05], resulting in exponential runtimes. Heuristics have been introduced to avoid the exponential behaviour while still providing a reasonable estimate of the barrier height. The first greedy approach by Morgan and Higgs considers direct paths only [MH98] which are of minimal length. Subsequently, the barrier estimation was improved via more advanced direct path heuristics [FHMS$^+$01, TOSY06, GFW$^+$08]. In order to avoid the restriction of direct pathways, heuristics were introduced that allow for minor detours in the landscape [LFH09, DLVHC10]. Such methods revealed the high potential of non-direct pathways.

Our *Shape Triples* approach aims to improve the barrier height approximation of arbitrary path heuristics by splitting the pathway prediction $x_s \rightsquigarrow x_t$ into $x_s \rightsquigarrow r \rightsquigarrow x_t$, where r is a defined checkpoint for a detour. We use RNA shapes and their representative structures, the so called *shreps* [GVR04], to define the detour checkpoints r. This is based on the observation that intermediate structures x_i along low barrier pathways can show very different branching patterns compared to the start and target structures x_s, x_t. Since RNA shapes group structures based on their branching pattern, we can use shreps to access the pattern of x_i. By pivoting on the shreps of all shapes, we have a good chance to catch the optimal detour while the number of shapes is very small compared to the number of RNA structures.

The resulting Shape Triples meta-heuristic, i.e. a high-level strategy that guides other path heuristics [Bla09], enables an efficient and precise estimation of barrier heights within RNA energy landscapes.

To evaluate our method, we show for two bistable RNA molecules the increased precision of the meta-heuristic compared to the employed path heuristic for a large number of refolding paths. We further show that in most cases the exact barrier height can be determined using our Shape Triples approach.

2 Preliminaries

In order to formulate our algorithms and results, we introduce the concept of energy landscapes, the barrier height problem, and their application to RNA. This is followed by an overview of RNA shape abstractions.

Energy Landscapes and Barrier Heights

In order to describe and investigate folding processes, the concept of discrete energy landscapes is applied frequently [Wri32, Sta02, FHSW02]. It is defined by a triple $\langle X, E, N \rangle$, i.e. a finite set of *states* X, an associated *energy function* $E : X \rightarrow \mathbb{R}$, and a *neighborhood relationship* $N : X \rightarrow \mathcal{P}(X)$, where $\mathcal{P}$ denotes the powerset. The folding process is mainly influenced by the *local minima* $M \subseteq X$ of the landscape defined by $\forall_{m \in M} \forall_{x \in N(m)} : E(m) \leq E(x)$.

A folding trajectory corresponds to a *walk (or path)* $w = (x_1, \ldots, x_l) \in X^l$ of length l within the energy landscape that respects the neighborhood relation ($\forall_i : x_i \in N(x_{i-1})$). With $W(x_s, x_t)$ we denote the infinite set of *all possible* walks starting in x_s and ending in x_t.

The *barrier height* B denotes the lowest energy peak to make two structures x_s, x_t accessible to each other, i.e.

$$B(x_s, x_t) = \min\{ \max\{ E(x \in w) \mid w \in W(x_s, x_t) \}\} \,. \qquad (1)$$

The barrier height heavily influences the folding probabilities within a certain energy landscape [FFHS00]. It can be used to derive energy landscape abstractions like barrier trees [HS88, FHSW02] and enables studies of folding kinetics [WSSF$^+$04, GFW$^+$08].

The *energy barrier problem* is to determine the exact barrier height B of two given states of an energy landscape.

RNA Secondary Structure Landscapes

In order to investigate the folding behavior of an RNA molecule the energy landscape of its secondary structures can be used [FFHS00, LFH09]. Given the nucleotide *sequence* $S \in \{\mathtt{A}, \mathtt{U}, \mathtt{G}, \mathtt{C}\}^n$ of an RNA of length n, a *secondary structure* x is a set of base pairs $\{(i, j) \mid 1 \leq i < j \leq n\}$ such that (a) S_i, S_j form a Watson-Crick (A-U, G-C) or a G-U base pair, with (b) at most one base pair per position, i.e. $\forall_{(i,j),(k,l)} : j \neq k \wedge (i = k \Leftrightarrow j = l)$, such that (c) all pairs are non-crossing, i.e. $\forall_{(i,j),(k,l)} : i < k < j \Leftrightarrow i < l < j$. The free energy of a given structure x can be calculated by a base pair based decomposition into structural elements [ZS81]. We use the implementation from the Vienna RNA Package[1] v1.7.2 within the Energy Landscape Library[2] v3.2.0 [MWB07]. All energies are given in $\frac{\text{kcal}}{\text{mol}}$ where calculations use parameters "`-d2 -T 37`". For details of the method applied and the energy parameters we refer to literature [ZS81, Hof03].

The neighborhood within an energy landscape reflects small structural changes along the folding process. To this end we apply so called *single moves* [FFHS00], i.e. the insertion or deletion of a single base pair. Thus, the neighborhood of a given structure x is defined by $N(x) = \{x' \mid |\mathtt{bp}(x) - \mathtt{bp}(x')| = 1\}$, using its number of base pairs $\mathtt{bp}(x) = |x|$.

The discrete *energy landscape of an RNA S* is thus defined by X as all secondary structures x of S, E as the free energy function defined by Zuker and Stiegler [ZS81], and the single move neighborhood N.

Maňuch *et al.* have shown the NP-completeness of the energy barrier problem in such RNA energy landscapes [MTSC09].

RNA Shape Abstractions

RNA shapes, introduced by Giegerich *et al.* [GVR04], are a coarse grained model of RNA secondary structures. The shape abstraction is a homomorphic mapping of the secondary structure set X of an RNA into a set of compact representations of the different branching pattern covered by X. Five levels of abstraction are introduced and we denote these $\pi_i(x)$, *the shape abstraction of the i-th level* of a given RNA structure x. For details on the method we refer to literature [GVR04, SVR$^+$06]. Throughout this manuscript we use the `RNAshapes`[3] implementation v2.1.5.

Given an RNA energy landscape $\langle X, E, N \rangle$, we denote with P_i *the set of all shape abstractions of level i of X*, i.e. $P_i = \pi_i(X) = \{\ \pi_i(x) \mid x \in X\}$. Thus each shape $p_i \in P_i$ describes a class of structures of X. The structure with minimal energy within the class is called the *shape representative structure* or *shrep* $r(p_i)$, i.e. $\forall_{x \in X} : (\pi_i(x) = p_i) \rightarrow E(x) \geq E(r(p_i))$.

In the following we will use the RNA shape abstraction concept to generate a new and

[1] Vienna RNA Package available at `http://www.tbi.univie.ac.at/~ivo/RNA/`

[2] ELL available at `http://www.bioinf.uni-freiburg.de/Software/`

[3] `RNAshapes` available at `http://bibiserv.techfak.uni-bielefeld.de/download/`

efficient meta-heuristic to estimate the barrier height between two RNA structures.

3 Methods

Since we want to present a meta-heuristic that employs an arbitrary path heuristic, we briefly review two existing direct path methods for the energy barrier problem, namely the `MH` heuristic by Morgan and Higgs [MH98] as well as a breadth-first-search (`BFS`) approach [FHMS$^+$01]. Both, the `MH` and `BFS` heuristic, can be implemented in our new RNA Shape Approaches presented afterwards. The exhaustive Shape Network approach exploits the potential of the RNA shape abstraction for the energy barrier problem. This is followed by our efficient *Shape Triples* meta-heuristic that enables a fast and precise barrier approximation.

RNA Direct Path Heuristics

Direct path heuristics find an approximate solution to the energy barrier problem for two RNA structures x_s, x_t. Considering only single moves (base pair insertion/deletion), a *direct path* $\hat{w}$ is a walk $w(x_s, x_t)$ of minimal length, i.e. of *base pair distance* $d(x_s, x_t) = |(x_s \cup x_t) \setminus (x_s \cap x_t)|$ [MH98]. In the following the *abbreviation* $B_{DP}(x_s, x_t)$ will be used to denote the barrier height between x_s and x_t estimated by a direct path heuristic.

The `MH` heuristic: Morgan and Higgs introduced a simple greedy heuristic to explore direct paths [MH98]. It uses an iterative conflict-driven scheme of base pair insertions and deletions and evaluates the maximal energy reached within the resulting walk. Applied in several iterations, while storing the path with lowest barrier found, it returns an upper bound on the barrier height. For details on the method refer to the literature [MH98, FHMS$^+$01, GFW$^+$08].

The `BFS` heuristic: Flamm *et al.* improved the greedy `MH` approach using a limited breadth-first-search (`BFS`) [FHMS$^+$01]. Starting from the initial structure x_s, it enumerates all single moves possible in direct walks towards the target structure x_t. From these walks only the best m candidates are considered for extension in the next iteration. This continues until the full walk length of $d(x_s, x_t)$, and thus the target structure x_t, is reached. `BFS` enables better barrier height approximations compared to `MH` to the cost of increasing runtime correlated with m [GFW$^+$08]. In the following, we denote a `BFS` search with cut-off m with $\texttt{BFS}_m$.

Drawbacks of Direct Paths: Direct path heuristics are fast, but at the cost of precision, since only a small "corridor" of the energy landscape is investigated. Thus, the barrier height estimated via direct paths is usually higher than the exact one, i.e. $B_{\text{DP}}(x_s, x_t) \geq B(x_s, x_t)$ [MH98]. Lorenz *et al.* have shown that lowest barrier pathways often contain detours and that rerouting via non-direct structures can significantly improve barrier height approximations [LFH09].

Shape Approaches

The central idea of our Shape Approaches is to use energy minimal shrep structures as intermediate checkpoints to reroute the path calculation of a given path heuristic, i.e. to go from the start structure x_s via shreps to target x_t. The resulting non-direct detour paths are more likely to enable a precise barrier estimate than the employed path heuristic alone. For simplicity, we exemplify the Shape Approaches employing a direct path heuristic as `MH` or `BFS`.

The Shape Network approach: In order to evaluate the potential of any Shape Approach we use the *Shape Network (SN)*, which uses the notion of shapes to create an abstraction of the energy landscape. The Shape Network is a fully connected, labeled graph where each node represents the shrep $r(p_i)$ of a shape $p_i \in P_i$ of a given fixed shape abstraction level i. In the following, we ignore the level identifier i and abbreviate $r(p_i) = r_p$ to ease the presentation. Each edge between two nodes $r_p, r_{p'}$ is labeled with a barrier height approximation via direct paths $B_{\text{DP}}(r_p, r_{p'})$ (e.g. using `MH` or `BFS`).

Utilizing a simple variation of the dynamic programming algorithm by Floyd for the shortest path problem [Flo62], we get the *barrier height approximation $B_F(r_p, r_{p'})$ for any two shreps $r_p, r_{p'}$* via any path within the Shape Network. Thus, using this B_{F} estimate we can get an upper bound $B_{\text{SN}}(x_s, x_t)$ of the barrier height between two RNA structures x_s, x_t including detours via an implicit sequence of shreps by

$$B_{\text{SN}}(x_s, x_t) = \min_{p,p' \in P} \{ \max \left\{ \begin{array}{l} B_{\text{DP}}(x_s, r_p), \\ B_{\text{F}}(r_p, r_{p'}), \\ B_{\text{DP}}(r_{p'}, x_t) \end{array} \right\}, \; B_{\text{DP}}(x_s, x_t) \} \tag{2}$$

The major drawback of the Shape Network approach is the high computational cost to calculate the Shape Network via $|P|^2$ direct path calculations where computation time depends on the heuristic (see direct path section). Afterwards the Floyd algorithm runs efficiently in $O(|P|^3)$ and results in the barrier height approximation B_F between all pairs of shreps. Once B_F is calculated, these approximations can be used to estimate the barrier height between any two structures using B_{SN} from Eq. 2 with $(2 \cdot |P| + 1)$ path calculations each.

Thus, the Shape Network approach is a useful tool when interested in a vast number of barrier heights, e.g. to calculate a barrier tree representation of the energy landscape's minima [FHSW02]. Beyond that, we can convert the Shape Network itself into an even coarser barrier tree abstraction covering the shrep structures that might reflect general properties of the energy landscape. Finally, the Shape Network approach gives a lower bound for meta-heuristics based on the Shape Approach idea.

The Shape Triples approach: In the following, we will introduce our *Shape Triples (ST)* meta-heuristic which enables a fast and efficient barrier height approximation. It is based on the observation that the majority of the barrier paths within the Shape Network are very short. We get already good upper bounds $B_{\text{ST}}(x_s, x_t)$ on the barrier height when only investigating detours with one intermediate shape representative r_p, i.e.

$$B_{\text{ST}}(x_s, x_t) = \min_{p \in P} \{ \max \left\{ \begin{array}{l} B_{\text{DP}}(x_s, r_p), \\ B_{\text{DP}}(r_p, x_t) \end{array} \right\}, \; B_{\text{DP}}(x_s, x_t) \}. \tag{3}$$

Thus, our two Shape Approaches yield new barrier height approximations B_{SN} and B_{ST} between the two structures x_s, x_t. These estimates are related via:

$$B(x_s, x_t) \leq B_{\text{SN}}(x_s, x_t) \leq B_{\text{ST}}(x_s, x_t) \leq B_{\text{DP}}(x_s, x_t). \tag{4}$$

```
function GETBST($x_s, x_t, P$)
    $B \leftarrow B_{\text{DP}}(x_s, x_t)$                                  ▷ initialization of barrier estimate
    for all ($p \in P$) do
        if ($E(r_p) < B$) then                                          ▷ low energy shreps only
            $B \leftarrow \min\{ B,\ \max\{B_{\text{DP}}(x_s, r_p), B_{\text{DP}}(r_p, x_t)\}\}$   ▷ update $B$ if needed
        end if
    end for
    return $B$                                                          ▷ final $B_{\text{ST}}(x_s, x_t)$ estimate
end function
```

Figure 1: Scheme for an efficient calculation of $B_{\text{ST}}(x_s, x_t)$.

In order to calculate $B_{\text{ST}}(x_s, x_t)$ from Eq. 3 we do not have to consider all shrep structures as possible intermediate checkpoints for detours. Every indirect path using a shrep r with $E(r) > B_{\text{DP}}(x_s, x_t)$ will result in a worse barrier height estimation than already given by B_{DP} (see Eq. 3). Thus we can use an adaptive scheme to reduce the computational cost for calculating B_{ST} that considers only shreps with energy below the best barrier height estimation found so far as given in Fig. 1. The scheme can be further improved when using an energy sorted shape/shrep enumeration: as soon as a shrep exceeds the current barrier estimate the iteration can be terminated. Note, the same applies to the Shape Network approach.

4 Results and Discussion

We investigate the Shape Approaches using the RNA molecules L45 and SL from Tab. 1. SL is the spliced leader RNA from *Leptomonas collosoma* taken from [LC93]. It was shown that the ability of this molecule to switch between two metastable structures heavily influences its splicing behavior. L45 is a bistable artificial RNA taken from [LFH09].

In order to evaluate the methods, we study the barrier height error, i.e. the approximated (Eq. 2/3) minus the exact barrier height (Eq. 1). To this end we pick 5000 random pairs (x_s, x_t) of local minima for SL with structural distance $\geq$ 7 and energy ≤ 0. The exact

ID	shape i	2	3	4	5	structures	
L45	$\|P_i\| =$	528	68	57	13	$\|X\| =$	5,999,391,327
SL	$\|P_i\| =$	6305	594	336	49	$\|X\| <$	1.725×10^{18}

L45	S	GGGCGCGGUUCGCCCUCCGCUAAAUGCGGAAGAUAAAUUGUGUCU
	x_s	(((((.....)))))(((((.....)))))(((((.....)))))
	x_t	((((((((((.....(((((.....))))).....))))))))))
SL	S	AACUAAAACAAUUUUUGAAGAACAGUUUCUGUACUUCAUUGGUAUGUAGAGACUUC
	x_s	..((...((((((..(((((.((((...)))).)))))..))).)))..)).....
	x_t	.......................((((((((((((.....)))))..)))))))..

Table 1: RNA shape/structure numbers and sequences S for the energy landscapes investigated. For SL we estimated $|X|$ via sequence length n using the upper bound of $1.07427 \cdot n^{-3/2} \cdot 2.35467^n$ from [CKKS09]. The structures x_s/x_t correspond to the switch structures of the bistable molecules.

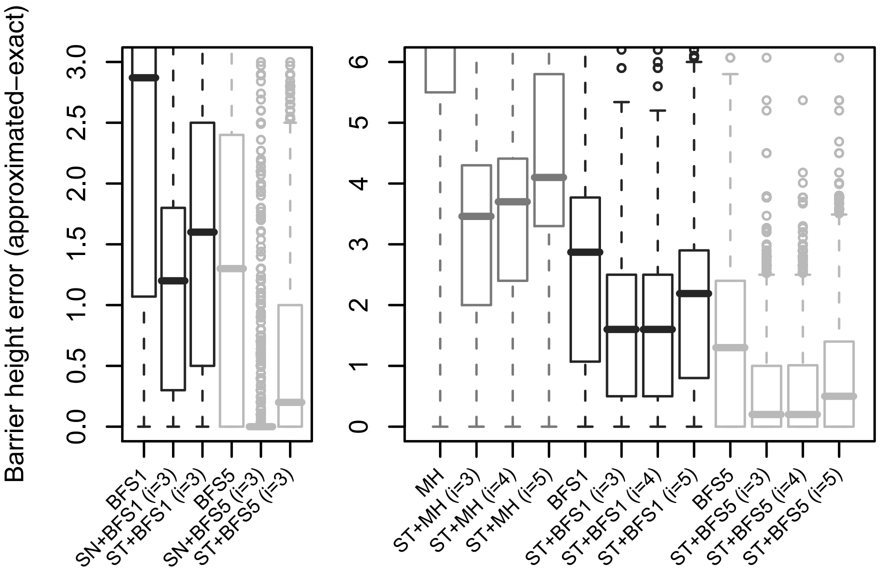


Figure 2: Evaluation of the Shape Approaches for RNA SL. (left) Direct path BFS-heuristic for cut-off 1 and 5 in comparison to Shape Network and Shape Triples approach at shape level i=3. (right) Performance of the Shape Triples approach when applying different direct path heuristics and shape levels i. Boxes cover 50% of the distribution while solid lines mark the median.

barrier height is calculated using an exhaustive approach implemented in the barriers program [FHSW02].

Figure 2 (left side) evaluates the Shape Approaches compared to the BFS direct path heuristics for SL. The Shape Network approach performs best among all methods and finds the exact barrier for $\geq 75\%$ of the pairs (SN+BFS$_5$). This shows the potential of detour pathways using RNA shapes. Furthermore, the much simpler Shape Triples heuristic shows only a slightly higher error on average and still outperforms the direct path heuristic.

The figure also compares (on the right) the performance of the Shape Triples approach for different shape levels and direct path heuristics. Here, BFS clearly beats the MH-heuristic and increasing BFS cut-offs lower the error (as in [GFW$^+$08]). More importantly, the Shape Triples approach always yields better results, depicting the robustness of the method and its independence of the direct path method applied. Finally, increased abstraction (shape level) reduces the precision of the method. This is expected since less detours in the landscape are considered (see Tab. 1 for shape numbers). Nevertheless, the differences get less significant when employing a more precise path heuristic like BFS$_5$ (in green).

Table 2 evaluates the Shape Triples approach for the structure pairs from Tab. 1. In most cases B_{ST} matches or is close to the exact barrier height B and improves the upper bound from direct path results (B_{DP}). Note, even for high shape abstraction levels we gain a significant improvement. First experiments reveal that an increase of the BFS cut-off can further improve our B_{ST} results (data not shown).

				Shape Triples $B_{\mathrm{ST}}(x_s, x_t)$				
ID	$B(x_s, x_t)$			shape i	2	3	4	5
`L45`	**-7.5**	only `BFS`$_5$	-4.87	with `BFS`$_5$	**-7.5**	**-7.5**	-6.4	-6.2
		$B_{\mathrm{DP}}(x_s, x_t) \geq$	-4.87	$\vert P_i\vert\%$	16.7	33.8	31.6	38.5
`SL`	**0.5**	only `BFS`$_5$	2.6	with `BFS`$_5$	**0.5**	**0.51**	**0.51**	2.6
		$B_{\mathrm{DP}}(x_s, x_t) \geq$	1.9	$\vert P_i\vert\%$	5.9	9.4	14.8	18.4

Table 2: Barrier height evaluation for the x_s/x_t structure pairs from Tab. 1. Given is the exact barrier $B(x_s, x_t)$, the estimate via only direct path `BFS`$_5$, the lowest barrier for such direct paths $B_{\mathrm{DP}}(x_s, x_t)$, and the Shape Triples approximations $B_{\mathrm{ST}}(x_s, x_t)$ for different shape level using `BFS`$_5$. $|P_i|\%$ denotes the percentage of $|P_i|$ from Tab. 1 used to calculate B_{ST} (see Methods).

The number of shapes $|P_i|$ grows slowly exponential with increasing sequence length (see Tab. 1) [NS09, LPC08]. Nevertheless, the percentage of shapes considered to calculate B_{ST} drops drastically as shown by $|P_i|\%$ in Tab. 2. Therefore, even for increasing sequence length, the computation effort of the Shape Triples approach remains low.

We compare our results to the κ,λ-neighborhood approach presented in [LFH09]. There, detours are rerouted through energy minimal structures within the κ,λ-neighborhood, i.e. via energy minimal structures within the structural distances κ and λ to the start and target structures, respectively. Using a `BFS`$_{100}$ heuristic (R. Lorenz, pers. commun.), Lorenz *et al.* are able to estimate the exact barrier height of -7.5 for `L45` [LFH09][4]. The Shape Triples approach reproduces the same exact barrier height for different shape levels (see Tab. 2) while using a much faster `BFS`$_5$ with cut-off 5 instead of 100 (see Methods).

5 Conclusion

We have introduced RNA shape based meta-heuristics to estimate the barrier height between RNA structures, an important problem to study multistable RNA molecules. The methods use shape representative structures (shreps) as intermediate checkpoints to reroute a given path heuristic. This enables a broader search in the energy landscape as done by the employed heuristic alone. We have shown that our Shape Triples approach is able to estimate barrier heights close to the optimum using a `BFS`$_5$ heuristic. The approach scales with the number of investigated shreps as shown in Fig. 1. Thus, the use of different shape levels enables a trade-off between barrier precision and computational performance (see Tab. 2) where the latter depends on the performance of the individual path heuristic applied.

While being introduced for direct path heuristics only, the method is applicable to any other path heuristic. Thus, we plan to investigate the use of the RNATABUPATH [DLVHC10], currently using a different RNA energy scheme, that was shown to yield slightly better results than `BFS` by allowing for minor detours. When employing RNATABUPATH

[4] Note, in [LFH09] the energy difference $\Delta E = (B(x_s, x_t) - E(x_s))$ is given. Thus, the barrier height was recalculated by $(E(x_s) + \Delta E)$.

within the Shape Triples approach it may be possible to improve the results even further (see Eq. 4).

We plan to investigate different shrep selection strategies to further speedup the method. Possible directions are the structural distance to start and target structure or a shape distance based evaluation.

Furthermore, the method is basically not restricted to RNA shapes but open to any sampling of low energy structures of the underlying RNA energy landscape. Thus, any scheme for an efficient calculation of such a set of structures can be used to replace the set of shape representatives in the Shape Triples approach (Fig. 1) and might even improve the results.

Therefore, we consider the Shape Triples meta-heuristic to be a very useful tool to combine results from different algorithmic fields to gain very precise barrier height estimates for arbitrary RNA structures.

References

[Bla09] P.E. Black. Metaheuristic. In Dictionary of Algorithms and Data Structures [online], Paul E. Black, ed., U.S. National Institute of Standards and Technology, 30 March 2009. (accessed 15.06.2010) Available from: http://www.itl.nist.gov/div897/sqg/dads/HTML/metaheuristic.html.

[CKKS09] P. Clote, E. Kranakis, D. Krizanc, and B. Salvy. Asymptotics of canonical and saturated RNA secondary structures. *JBCB*, 7(5):869–893, 2009.

[DLVHC10] I. Dotu, W.A. Lorenz, P. Van Hentenryck, and P. Clote. Computing folding pathways between RNA secondary structures. *NAR*, 38(5):1711–1722, 2010.

[FFHS00] C. Flamm, W. Fontana, I.L. Hofacker, and P. Schuster. RNA folding at elementary step resolution. *RNA*, 6(03):325–338, 2000.

[FHMS$^+$01] C. Flamm, I.L. Hofacker, S. Maurer-Stroh, P.F. Stadler, and M. Zehl. Design of multistable RNA molecules. *RNA*, 7(02):254–265, 2001.

[FHSW02] C. Flamm, I.L. Hofacker, P.F. Stadler, and M.T. Wolfinger. Barrier Trees of Degenerate Landscapes. *Z. Phys. Chem.*, 216(2/2002):155–173, 2002.

[Flo62] R.W. Floyd. Algorithm 97: Shortest path. *Commun. ACM*, 5(6):345, 1962.

[GFW$^+$08] M. Geis, C. Flamm, M.T. Wolfinger, A. Tanzer, I.L. Hofacker, M. Middendorf, C. Mandl, P.F. Stadler, and C. Thurner. Folding kinetics of large RNAs. *J of Mol Biol.*, 379(1):160–173, 2008.

[GVR04] R. Giegerich, B. Voss, and M. Rehmsmeier. Abstract shapes of RNA. *NAR*, 32(16):4843–4851, 2004.

[Hof03] I.L. Hofacker. Vienna RNA secondary structure server. *NAR*, 31(13):3429–31, 2003.

[HS88] K.H. Hoffmann and P. Sibani. Diffusion in hierarchies. *Phys. Rev. A*, 38(8):4261–4270, 1988.

[KH05] M. Kubota and M. Hagiya. Minimum basin algorithm: An effective analysis technique for DNA energy landscapes. *LNCS*, 3384:202–214, 2005.

[LC93] K.A. LeCuyer and D.M. Crothers. The Leptomonas collosoma spliced leader RNA can switch between two alternate structural forms. *Biochemistry*, 32(20):53015311, 1993.

[LFH09] R. Lorenz, C. Flamm, and I.L. Hofacker. 2D projections of RNA folding landscapes. In *Proc. of GCB'09*, volume 157 of *LNCS*, pages 11–20, 2009.

[LPC08] W.A. Lorenz, Y. Ponty, and P. Clote. Asymptotics of RNA shapes. *J of Comp. Biol.*, 15(1):31–63, 2008.

[MH98] S.R. Morgan and P.G. Higgs. Barrier heights between ground states in a model of RNA secondary structure. *J of Physics A*, 31:3153–3170, 1998.

[MTSC09] J. Maňuch, C. Thachuk, L. Stacho, and A. Condon. NP-Completeness of the Direct Energy Barrier Problem without Pseudoknots. In *DNA Computing*, volume 5877 of *LNCS*, pages 106–115, 2009.

[MWB07] M. Mann, S. Will, and R. Backofen. The Energy Landscape Library–a platform for generic algorithms. *Proc. of BIRD*, 7:83–86, 2007.

[NS09] Markus Nebel and Anika Scheid. On quantitative effects of RNA shape abstraction. *Theory in Biosciences*, 128(4):211–225, 2009.

[Sta02] P.F. Stadler. Fitness landscapes. In *LNP*, pages 183–204. Springer, 2002.

[SvdPS99] P. Sibani, R. van der Pas, and J.C. Schön. The lid method for exhaustive exploration of metastable states of complex systems. *Comp. Phys. Comm.*, 116(1):17–27, 1999.

[SVR$^+$06] P. Steffen, B. Voss, M. Rehmsmeier, J. Reeder, and R. Giegerich. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22(4):500–503, 2006.

[TOSY06] T. Takeda, H. Ono, K. Sadakane, and M. Yamashita. A Local Search Based Barrier Height Estimation Algorithm for DNA Molecular Transitions. In *DNA Computing*, volume 3892 of *LNCS*, pages 359–370, 2006.

[Wri32] S. Wright. The Roles of Mutation. In *Congress on Genetics*, page 365, 1932.

[WSSF$^+$04] M.T. Wolfinger, W.A. Svrcek-Seiler, C. Flamm, I.L. Hofacker, and P.F. Stadler. Efficient computation of RNA folding dynamics. *J of Physics A*, 37(17):4731–4741, 2004.

[ZS81] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *NAR*, 9(1):133–148, 1981.

Efficient Similarity Retrieval for Protein Binding Sites based on Histogram Comparison

Thomas Fober[1]*, Marco Mernberger[1]*, Gerhard Klebe[2] and Eyke Hüllermeier[1]
[1]*Department of Mathematics and Computer Science*
[2]*Department of Pharmacy*
Philipps-Universität, 35032 Marburg, Germany

Abstract: We propose a method for comparing protein structures or, more specifically, protein binding sites using a histogram-based representation that captures important geometrical and physico-chemical properties. In comparison to hitherto existing approaches in structural bioinformatics, especially methods from graph theory and computational geometry, our approach is computationally much more efficient. Moreover, despite its simplicity, it appears to capture and recover functional similarities surprisingly well.

1 Introduction

With the steady improvement of structure prediction methods, the inference of protein function based on structure information becomes more and more important. The comparison of protein structures, for which quite a number of methods have already been proposed, is a central task in this regard. One class of methods focuses on geometrical aspects and, correspondingly, makes use of tools from computational geometry. As examples of this type of approach, we mention geometric hashing [RW97] and labeled point cloud superposition [FH09]. Another idea is to use graphs as formal models of molecular structures. Here, the focus is more on the physical and chemical properties, which are often modeled as nodes of a graph, while geometrical or topological properties are captured in a more indirect way via weighted edges. Typical examples of this approach include measures based on sub-graph isomorphism [NB07], graph edit distance [FMKH09], and graph kernels [G08].

Geometrical and graph-based approaches are appealing, especially since they produce more than just a numerical degree of similarity. Usually, they also provide useful extra information, e.g., correspondences between basic structural units. The price to pay is a high computational complexity. In fact, many of the aforementioned methods lead to NP-hard optimization problems and scale poorly with the size of the structures. This complexity prevents them from being used in large-scale studies like cluster analysis requiring all-against-all comparisons.

*The first two authors contributed equally to this work.

A possible alternative to methods of the above kind is offered by *feature-based* approaches in which a protein structure is first represented in terms of a fixed number of features, that is, a vector of fixed dimensionality. The comparison of objects is thus reduced to the comparison of feature vectors. Since the original object cannot be recovered from a finite number of features, this transformation normally comes with a significant loss of information. Consequently, it is unclear to what extent the similarity of the original structures is mirrored by the similarity of their respective feature vectors. On the other hand, this approach has an obvious advantage with regard to complexity, as feature vectors can be compared quite efficiently.

In this paper, we propose a feature-based approach to the comparison of protein binding sites. More specifically, our idea is to summarize important information about the geometrical and physico-chemical properties of protein binding sites in terms of histograms. This idea is largely motivated by the successful use of similar approaches in the field of image processing, where the distribution of the brightness or the colors of a picture are represented in terms of histograms [RTG00, VB00]. A similar approach has also been applied in the field of structural bioinformatics [SSS$^+$07] for the analysis of homologous proteins.

2 Modeling Protein Binding Sites

Our approach builds upon CavBase [SKK02], a database for the automated detection, extraction, and storing of protein cavities (hypothetical binding sites) from experimentally determined protein structures. In CavBase, a set of points is used as a first approximation to describe a binding pocket.

The geometrical arrangement of the pocket and its physicochemical properties are first represented by predefined *pseudocenters* – spatial points that represent the center of a particular property. The type and the spatial position of the centers depend on the amino acids that border the binding pocket and expose their functional groups. Currently, CavBase considers seven types of pseudocenters (hydrogen-bond donor, acceptor, mixed donor/acceptor, hydrophobic aliphatic, metal ion, pi, aromatic).

Pseudocenters can be regarded as a compressed representation of areas on the cavity surface where certain protein-ligand interactions are experienced. Consequently, a set of pseudocenters is an approximate representation of a spatial distribution of physicochemical properties.

3 Transforming Protein Binding Sites into Histograms

A histogram h is a partition of a set of observations $\mathcal{O} \subset \mathcal{X}$ into a finite number of discrete units. Formally, h can be represented as a $\mathcal{B} \longrightarrow \mathbb{R}$ mapping, where $\mathcal{B}$ is a finite set of *bins*, and $h(b)$ denotes the number (fraction) of observations falling into bin

b. We call a histogram h normalized if $\sum_{b \in \mathcal{B}} h(b) = 1$. Each bin b is associated with a subset $X[b]$ of the domain $\mathcal{X}$, so that $h(b) = |\mathcal{O} \cap X[b]|$ before normalization and $h(b) = |\mathcal{O}|^{-1} |\mathcal{O} \cap X[b]|$ in the normalized case. The set of bins is assumed to form a partition of $\mathcal{X}$, i.e., $X[a] \cap X[b] = \emptyset$ for $a \neq b$ and $\bigcup_{b \in \mathcal{B}} X[b] = \mathcal{X}$.

To obtain histograms from a protein binding site, we will use two important properties, namely its distribution of pseudocenters and the distribution of distances between pseudocenters, thereby capturing both, the physico-chemical properties as well as the geometry of the binding site.

To combine both pseudocenter and distance information, our representation is based on sets of pairwise distances: $D_{i,j}$ is the set of all distances between pseudocenters of type i and j, with $1 \leq i \leq j \leq n_p$ (n_p denoting the number of pseudocenter types). To obtain a corresponding histogram $h_{i,j}$, we use $\mathcal{B} = \{1, \ldots, d_{\text{max}}\}$ and let $X[b] = [b-1, b[$. All histograms are normalized to ensure equal weights (except empty histograms). Thus, a structure is represented by a set of $n = n_p(n_p+1)/2$ histograms.

4 Distance Measures

Consider two structures represented, respectively, by histograms $g_1, \ldots, g_n$ and $h_1, \ldots, h_n$. Moreover, let δ be a distance measure suitable for comparing histograms. The overall distance between the two structures can then be obtained by aggregating the distances $\delta(g_i, h_i)$, for example in terms of the Euclidean norm of the vector

$$\left(\delta(g_1, h_1), \ldots, \delta(g_n, h_n)\right) \ .$$

In the literature, two types of distance measures on histograms are distinguished, namely *bin-by-bin* and *cross-bin* measures. The former are rather simple and only compare values in the same bin. The distance between two histograms is then defined by the sum of distances for all bins. Cross-bin measures, on the other hand, also compare values in different bins. In order to aggregate these distances, they also require the existence of a *ground distance* on $\mathcal{B}$; in our case, we can simply define $|a - b|$ as distance between bins a and b.

Since cross-bin measures proved superior to bin-by-bin measures in a previous study [FH10], we focus on the former type. More precisely, we consider the Quadratic Form Distance,

$$d_{QF}(g, h) = \sqrt{(\vec{g} - \vec{h})^T A (\vec{g} - \vec{h})} \ ,$$

where A is a matrix whose entries $a_{i,j}$ specify the similarity between bins b_i and b_j with

$$a_{i,j} = 1 - \frac{d_{i,j}}{\max_{i,j}\{d_{i,j}\}} \ ,$$

the Earth Mover's Distance,

$$d_{EMD}(g,h) = \begin{cases} \min\left\{\sum_{\mathcal{B}_n} f_{i,k} \,|\{f_{i,k} : (i,k) \in \mathcal{B}_n\}\right\} \\ \text{subject to:} \\ \sum_{k:(i,k)\in\mathcal{B}_n}(f_{i,k} - f_{k,i}) = g(b) - h(b) \quad \forall\, b \in \mathcal{B} \\ f_{i,k} \geq 0 \qquad\qquad\qquad\qquad\qquad \forall\,(i,k) \in \mathcal{B}_n \end{cases}$$

and Cumulative Distributions. The latter approach replaces the original histogram h by the corresponding cumulative distribution, defined by $H(b) = \sum_{a\leq b} h(a)$, and then measures the distance on these distributions. Here, we use the Kolmogorov-Smirnov distance

$$d_{KS}(g,h) = \max_{b\in\mathcal{B}}\{|G(b) - H(b)|\}$$

and the match distance

$$d_M(g,h) = \sum_{b\in\mathcal{B}} |G(b) - H(b)|.$$

5 Experimental Results

In our experiments, we first used a dataset from a previous study designed to assess the performance of global structural alignment methods. This dataset contains 355 protein binding sites comprising two classes of proteins, ATP binding and NADH binding proteins. Binding sites known to bind the corresponding ligands in similar conformation were derived from CavBase; in case of multiple binding sites belonging to the same structure, only one representative was selected at random. See [FMKH09] for a more thorough description of the dataset.

As a second, more complex dataset (Table 1), we selected a number of different, highly populated functional enzyme classes according to the ENZYME database [BWF$^+$00]. Protein structures belonging to the selected classes were derived from the Protein Data Bank and corresponding cavities where extracted from CavBase.

Since CavBase may contain multiple cavities for the same protein, not all of them being functionally important, we selected only those binding sites that contained at least two residues belonging to the catalytic center of the protein according to the catalytic activity atlas annotation (CSA) version 2.2.12 [PBT04]. In case of multiple instances for the same structure, we took the binding site with the largest number of catalytic residues.

5.1 Classification Performance on a Two-Class Problem

As a first proof-of-concept, we assessed the performance of our distance measure on a two-class classification problem, namely of ATP- versus NADH-binding proteins. More precisely, we used a k-nearest neighbor (k-NN) classifier combined with different cross-bin measures to discriminate the two classes. As performance criteria, we measured the

EC number	Function	Number of proteins
2.1.1.45	thymidylate synthase	153
3.4.21.4	trypsin	373
3.4.23.16	HIV-1 retropepsin	291
3.4.24.27	thermolysin	70
1.9.3.1	cytochrome-c oxidase	233
4.2.1.1	carbonate dehydratase	316
3.4.25.1	proteasome endopeptidase	167
2.6.1.1	aspartate transaminase	106

Table 1: Dataset of 8 different EC classes.

accuracy of the methods in terms of their classification rates (determined through leave-one-out cross validation) as well as their efficiency in terms of runtime.

For comparison, we also applied kernel methods (the shortest path (SP) kernel [BK05], the random walk (RW) kernel [G08] and the fingerprint (FP) kernel [FMM$^+$09]), graph-based methods (the iterative graph alignment (IGA) [WHKK07] and the evolutionary graph alignment (GAVEO) [FMKH09]) and geometric approaches (the labeled point cloud superposition (LPCS) [FH09]).

Table 2 summarizes the results of these approaches. As can be seen, there are clear differences in terms of performance: The highest classification accuracy is achieved by LPCS, followed by the fingerprint kernels. The graph-alignment methods (IGA and GAVEO) perform less strongly, and the worst classification rates are produced by the graph kernels.

The runtime reported in the table includes the time needed for an all-against-all comparison of the 355 structures and the time needed to perform a leave-one-out cross validation. As can be seen, all methods require at least one day.

k	RW	SP	LPCS	FP	IGA	GAVEO
1	0.597	0.606	0.935	0.842	0.766	0.789
3	0.597	0.628	0.916	0.882	0.718	0.766
5	0.597	0.634	0.890	0.873	0.724	0.780
runtime (h)	1149.88	171.14	361.58	35.98	2136.88	> 5000

Table 2: Classification rates and runtime in hours of a k-NN classifier using different values of k and different distance measures.

Table 3 summarizes the results for our histogram approach using different cross-bin distance measures and bins of size 1 (as they will be used in the whole work). Interestingly, the accuracy values are quite high, even outperforming some of the competitor methods, although LPCS still performs best. However, considering the runtime efficiency of the histogram approach, the results show that we can retrieve comparably good results within only a fraction of the time.

k	d_{QF}	d_M	d_{KS}	d_{EMD}
1	0.862	0.865	0.859	0.772
3	0.856	**0.882**	0.854	0.749
5	0.845	0.865	0.837	0.732
runtime (h)	0.785	0.470	0.472	11.53

Table 3: Classification rates of cross-bin measures on the NADH/ATP data set.

Rank	pdb code	Protein	score
1	2ACK	Acetylcholinesterase (AChE)	0
2	1AX9	Acetylcholinesterase (AChE)	0.180
3	1GQS	Acetylcholinesterase (AChE)	0.203
⋮	⋮	⋮	⋮
98	2V98	Acetylcholinesterase (AChE)	0.402
99	1ZGC	Acetylcholinesterase (AChE)	0.404
100	1G6R	Aspartate aminotransferase (mAspAT)	0.405

Table 4: Top ranks retrieved by querying the CavBase with the main pocket of 2ACK. Omitted entries contained exclusively acetylcholinesterases.

5.2 Database Querying

In a second experiment, we applied our approach on the task of querying the complete CavBase for similar structures. Given the simplicity of the approach, one may doubt its suitability for a task of this kind.

We chose the main pocket of acetylcholinesterase from *T. californica* (pdb code: 2ACK) as a query structure. This protein has previously been used to query the CASTp database with a similarity measure that combines structural similarity with evolutionary conservation [BAL03]. Binkowski et al. retrieved further acetylcholinesterase structures on all top ranks, a result they attributed to the uniqueness of the protein structure.

Table 4 shows some results of our query using the match distance. Surprisingly, and despite the simplicity of our approach, the top 99 ranks are exclusively occupied by other acetylcholinesterase structures before the first false positive shows up on position 100. This is consistent with the results of Binkowski et al. and suggests that important information is indeed captured by our histogram representation.

5.3 Discriminating Enzyme Classes

The third experiment investigates whether our approach can be used to discern binding pockets of different enzyme classes. To this end, we selected several highly populated enzyme classes from the Protein Data Bank and calculated the corresponding distance matrix using our histogram approach with the match distance.

k	d_{QF}	d_M	d_{KS}
1	0.941	0.944	0.945
3	0.920	0.919	0.926
5	0.905	0.912	0.916

Table 5: Classification accuracy on the multi-class enzyme dataset.

Since the class information is known, we visualize the distance matrix by means of a heat map, which is shown in Figure 1. Again, it can be seen that important information is captured by the histogram approach, as several classes show a high similarity within the class.

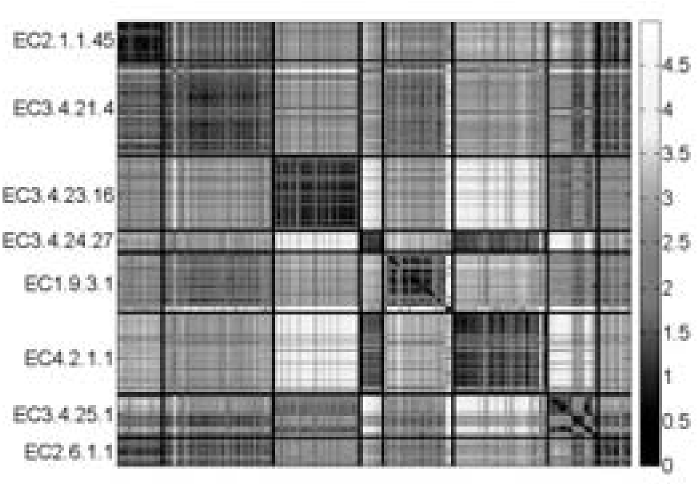


Figure 1: Heat map depicting the distance matrix based on match distance for the EC dataset. Different EC classes are seperated by black lines.

Based on the above distance matrix, we additionally performed a hierarchical clustering using repeated bisection and subsequent k-way refinement. Comparing the resulting clustering with the original EC class yields a Rand index of $R = 0.8633$, indicating that the clustering is in good agreement with the real class structure.

Finally, the distance matrix was again used for a nearest neighbor classification, this time on a multi-class problem. Table 5 shows the classification accuracies for a leave-one-out cross validation, using different distance metrics.

6 Conclusions

In this paper, we have introduced a very simple though extremely efficient method for comparing protein structures in terms of a histogram-based representation. The main interest of the paper is probably less the method itself, but more its strong performance in our experimental studies on classification and retrieval. In light of the simplicity of the representation and the distinctive loss of information it implies, this performance was unexpected. On the other hand, it is true that similar representations have been used quite successfully in other fields, too, where the loss of information is arguably not smaller.

Due to its runtime efficiency and scalability, our approach is amenable to applications that cannot be tackled by other methods. It can be used as a kind of filter, for example, to preselect structures from very large datasets, thereby reducing the amount of data to be processed afterward by more complex structure comparison algorithms. Using the method for clustering, as we have already done in our experiments, is another example. Indeed, the need for an all-against-all comparison does usually prevent the use of computationally complex methods here.

Acknowledgements: The authors like to thank the reviewers for useful suggestions that helped to improve the paper and, moreover, for bringing the approach of Sander *et al.* to their attention.

References

[BAL03] T.A. Binkowski, L. Adamian, and J. Liang. Inferring Functional Relationships of Proteins from Local Sequence and Spatial Surface Patterns. *Journal of Molecular Biology*, 332(2):505–526, 2003.

[BK05] K.M. Borgwardt and H.P. Kriegel. Shortest-Path Kernels on Graphs. In *International Conference on Data Mining*, pages 74–81, Houston, Texas, 2005.

[BWF$^+$00] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[FH09] T. Fober and E. Hüllermeier. Fuzzy Modeling of Labeled Point Cloud Superposition for the Comparison of Protein Binding Sites. In *Proc. IFSA/EUSFLAT–2009*, pages 1299–1304, Lisbon, Portugal, 2009.

[FH10] T. Fober and E. Hüllermeier. Similarity Measures for Protein Structures based on Fuzzy Histogram Comparison. In *WCCI–2010, World Congress on Computational Intelligence*, Barcelona, 2010.

[FMKH09] T. Fober, M. Mernberger, G. Klebe, and E. Hüllermeier. Evolutionary Construction of Multiple Graph Alignments for the Structural Analysis of Biomolecules. *Bioinformatics*, 25(16):2110–2117, 2009.

[FMM$^+$09] T. Fober, M. Mernberger, V. Melnikov, R. Moritz, and E. Hüllermeier. Extension and Empirical Comparison of Graph-Kernels for the Analysis of Protein Active Sites. In *Lernen, Wissen, Adaptivität*, pages 30–36, Darmstadt, Germany, 2009.

[G08] T. Gärtner. *Kernels for Structured Data*. World Scientific, Singapore, 2008.

[NB07] M. Neuhaus and H. Bunke. *Bridging the Gap between Graph Edit Distance and Kernel Machines*. World Scientific, New Jersey, 2007.

[PBT04] C.T. Porter, G.J. Bartlett, and J.M. Thornton. The Catalytic Site Atlas: A Resource of Catalytic Sites and Residues Identified in Enzymes using Structural Data. *Nucleic Acids Research*, 32:129–133, 2004.

[RTG00] Y. Rubner, C. Tomasi, and L. J. Guibas. The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[RW97] I. Rigoutsos and H. Wolfson. Geometric Hashing. *IEEE Computational Science Engineering*, 4:1070–9924, 1997.

[SKK02] S. Schmitt, D. Kuhn, and G. Klebe. A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology. *Journal of Molecular Biology*, 323(2):387–406, 2002.

[SSS$^+$07] O. Sander, T. Sing, I. Sommer, A.J. Low, P.K. Cheung, P.R. Harrigan, T. Lengauer, and F.S. Domingues. Structural Descriptors of gp120 V3 Loop for the Prediction of HIV-1 Coreceptor Usage. *PLoS Computational Biology*, 3(3):555–564, 2007.

[VB00] C. Vertan and N. Boujemaa. Using Fuzzy Histograms and Distances for Color Image Retrieval. In *Challenge of Image Retrieval*, pages 1–6, Brighton, United Kingdom, 2000.

[WHKK07] N. Weskamp, E. Hüllermeier, D. Kuhn, and G. Klebe. Multiple Graph Alignment for the Structural Analysis of Protein Active Sites. *IEEE Transactions on Computational Biology and Bioinformatics*, 4(2):310–320, 2007.

Repeat-aware Comparative Genome Assembly

Peter Husemann* and Jens Stoye
AG Genominformatik, Technische Fakultät, Bielefeld University, Germany

{phuseman, stoye}@techfak.uni-bielefeld.de

Abstract: The current high-throughput sequencing technologies produce gigabytes of data even when prokaryotic genomes are processed. In a subsequent assembly phase, the generated overlapping reads are merged, ideally into one contiguous sequence. Often, however, the assembly results in a set of contigs which need to be stitched together with additional lab work. One of the reasons why the assembly produces several distinct contigs are repetitive elements in the newly sequenced genome. While knowing order and orientation of a set of non-repetitive contigs helps to close the gaps between them, special care has to be taken for repetitive contigs. Here we propose an algorithm that orders a set of contigs with respect to a related reference genome while treating the repetitive contigs in an appropriate way.

1 Introduction

The sequencing of genomes has become easier and cheaper with the current massively parallel sequencing methods [Mar08]. Following a shotgun approach, these methods fragment the genome randomly into small parts. The ends of those fragments are sequenced and referred to as *reads*, both reads of one fragment form a *mate pair*. In a subsequent assembly phase, an assembler software tries to merge overlapping parts of the reads into longer contiguous sequences [Pop09], the so called *contigs*. If the size of the fragments is known, the mate pairs have a defined distance and this information can also help in the assembly. Ideally, the result is a single sequence which resembles the complete genome. However, there are a few obstacles that lead to several contigs instead of a single one. Besides non-random fragmentation and areas of an unusual GC content, repeats play a major role in this process [MKS10]. For the latter the assembly software can not distinguish between the reads from different occurrences of the repeating region. Thus, reads with the same sequence from several distinct origins are merged together to a single *repetitive contig*.

In general, the output of the assembly phase is a set of contigs and maybe a rough scaffold derived for example from mate pair information. In order to retrieve the complete genomic sequence, the gaps between the contigs have to be filled by running additional experiments in the lab. If two contigs are known to be adjacent, then it is possible to close the gap with a (long range) PCR or with primer walking for even larger gaps. For the tedious process of gap closing it is therefore beneficial to know the *layout* of the contigs, meaning the order and orientation of them. Mate pair information can again help in this situation, but it fails for gaps longer than the fragment size, and it is also not reliable on repetitive sequences.

Fortunately, many genomes have already been sequenced completely and if one of them is related to the newly sequenced genome then it can be used as a reference to estimate a proper layout. There are a few programs that deal with this task: Projector2 [vHZKK05] maps the contigs on a single template genome and designs primer pairs for gap closure, OSLay [RSH07] finds an optimal syntenic layout, other programs like PGA [ZZLB08] and treecat [HS10a] are even able to utilize several reference genomes to predict a consensus layout. Still problematic for the above mentioned programs are major rearrangements in the reference genomes as well as repeating regions. To cope with the latter, all programs except for treecat employ or at least suggest a repeat masking step.

In this paper we address the problem of repeating contigs in prokaryotic genomes with the goal to find a better layout for a set of contigs according to a closely related reference genome. Therefore we present a novel algorithm that includes repeating contigs as often as necessary in a computed layout. This greatly reduces the complexity of the layout graph while not removing the repetitive contigs.

After introducing the basic concept of a contig adjacency graph in Section 2, we address in Section 3 how repeating contigs can be discovered from a given set of contigs. In Section 4 we present a specially designed algorithm that uses the repeat information to estimate a more appropriate layout for the contigs. Section 5 contains an evaluation of the algorithm and a comparison of the repeat integration with other contig ordering programs.

2 Contig adjacency graph

Let $\Sigma = \{A, C, G, T\}$ be the alphabet of nucleotides. We denote by Σ^* the set of all finite strings over Σ, by $|s| := \ell$ the *length* of string $s = s_1 \dots s_\ell$, and by $s[i, j] := s_i \dots s_j$ with $1 \leq i \leq j \leq \ell$ the *substring* of s that starts at position i and ends at position j.

Suppose we are given a set of contigs $\mathcal{C} = \{c_1, \dots, c_n\}$, $c_i \in \Sigma^*$, and a reference genome $g \in \Sigma^*$ that has already been finished. The *contig adjacency graph* for these sequences is then the weighted graph $G_{\mathcal{C},g} = (V, E)$ that contains for each contig $c_i \in \mathcal{C}$ two vertices: l_i as the *left connector* and r_i as the *right connector* of contig c_i, thus $V = \{l_1, \dots, l_n, r_1, \dots, r_n\}$. A function $contig(v)$ refers to the contig for which vertex v represents the left or right connector. The graph $G_{\mathcal{C},g}$ is fully connected, $E = \binom{V}{2}$, and we term $A = \{\{v, v'\} \mid contig(v) \neq contig(v')\}$ the set of *adjacency edges* that connect the contigs among each other.

The edge weights are given by a function $w : E \to \mathbb{R}_0^+$, and we are particularly interested in the weights of the adjacency edges A. These shall provide a score of how likely the involved connectors are adjacent with respect to the reference genome. One method to calculate the weights based on matching the contigs onto the reference genome is described in [HS10a]. There, the pairwise matches from different contigs are used to calculate a score for the adjacency of the involved contig connectors. While the scores increase with the length of the corresponding matches, they are weighted by their distance. A high weight $w(\{r_i, l_j\})$, for example, supports that the right connector (or head) of contig c_i is adjacent to the left connector (or tail) of contig c_j. We call the sum of all edge weights of a particular node $v \in V$ the *total support* of that node, denoted by $\mathcal{S}_v = \sum_{v' \in V} w(\{v, v'\})$. To estimate how significant an adjacency edge $e = \{v, v'\} \in A$ is for a given contig connector $v \in V$, we consider the *relative support*: $\mathcal{S}_v^{\mathsf{rel}}(e) = \frac{w(e)}{\mathcal{S}_v}$. Intuitively, this fraction tells how specific the connection is for the given contig connector. A single high weight edge results in a relative support close to one, while many equally good connections will lower the value. Note that in general $\mathcal{S}_v^{\mathsf{rel}}(\{v, v'\}) \neq \mathcal{S}_{v'}^{\mathsf{rel}}(\{v, v'\})$.

Given a contig adjacency graph, a natural task is to find a subgraph of it that contains all relevant adjacencies in order to ease the gap closure phase of the sequencing project. We call any subgraph with this property a *layout graph* of a set of contigs. A basic approach could be to find a tour of maximal weight that contains each contig once and in a specified direction. This leads essentially to the problem of finding a longest Hamiltonian cycle in $G_{\mathcal{C},g}$ and is thus NP hard. Moreover, a meaningful biological result can differ from it, especially if some contigs appear several times on the genome. In this case, a repetitive contig has to be included several times into an adequate layout. In the next section we describe how to detect such repetitive contigs.

3 Repeat detection

Our approach for contig ordering distinguishes between repetitive and non-repetitive contigs. Assuming that repeating regions are conserved between closely related species, we can determine if a contig $c \in \mathcal{C}$ is repetitive by matching it onto a given reference genome g. A *match* of c in g is represented as a pair $m = \big((s_b, s_e), (t_b, t_e)\big)$ where the indices denote the starting and ending positions of the two substrings $c[s_b, s_e]$ and $g[t_b, t_e]$. An alignment of the substrings is supposed to yield a high score. Reverse complement matches can be modeled with this notation as well, but are left out for simplicity.

Given a set of matches $\mathcal{M}_{c,g}$ of contig c on the reference genome g, we can determine which matches are repetitive and from this derive if the contig occurs repetitively: We call a match $m = \big((s_b, s_e), (t_b, t_e)\big) \in \mathcal{M}_{c,g}$ *repetitive* if there exists another match $m' = \big((s'_b, s'_e), (t'_b, t'_e)\big) \in \mathcal{M}_{c,g}$ such that (i) the contig substring of m is included in the substring of m' ($s_b \geq s'_b$ and $s_e \leq s'_e$), and (ii) the match positions on the reference are not overlapping ($\{t_b, \ldots, t_e\} \cap \{t'_b, \ldots, t'_e\} = \emptyset$). The exact positions of the matches may vary for different matching procedures and/or scoring functions, so we allow for condition (i) a slack of ρ_1 times the length of m. By default we use a value of 10% for ρ_1.

We call a contig *repetitive* if it has at least one repetitive match m of sufficient length. Sufficient means that at least a fraction of ρ_2 of the contig is covered by the repetitive match: $s_e - s_b \geq \rho_2 \cdot |c|$. As default we set $\rho_2 = 0.9$. We call all contigs that are not repetitive for the sake of a shorter notation *regular contigs*.

Contigs that are repetitive on the newly sequenced genome are not necessarily repetitive on the employed reference genome. To extend, as well as verify, the prediction of repetitive contigs, one can use the information on how many reads have been merged to form a contig provided from the assembly phase. For each contig the average read coverage can be calculated, and by looking at the deviation from the median of these values, it can be observed if a contig is over- or underrepresented with reads. Highly overrepresented contigs are most likely repetitive since the reads gathered from all repeat occurrences are merged to a single contig. Even more, the ratio with respect to the median can serve as a rough estimate for the number of occurrences of a repetitive contig.

4 Repeat-aware layout algorithm

In this section we adopt the contig adjacency discovery algorithm that was proposed in [HS10a] to be aware of repetitive contigs and include them appropriately. The overall strategy is to distinguish between regular and repetitive contigs and to process both sets one after another. The absence of repetitive contigs in the first set implies that most contigs should have exactly two neighbors. Following this observation, we describe in Section 4.1 an algorithm for creating a basic layout graph. In the subsequent Section 4.2 we address how this layout graph can be augmented with the repetitive contigs in a meaningful way.

4.1 Layouting the non-repetitive parts of the genome

To devise a basic layout graph for the regular contigs, two steps are necessary:

1. The contig adjacency graph that contains the edge weights for all contig connectors has to be computed. Note that the graph is created for repetitive *and* regular contigs, thus the procedure starts with matching *all* contigs onto the given reference genome. The score calculation is performed as in [HS10a] with the difference that repetitive matches are ignored for the calculation of scores between regular contigs. This helps to reduce misleading edges for a contig caused for example if it is flanked by repeats. Of course, for repetitive contigs all matches are used.

2. In the second step, the calculated edge weights can be used to extract the adjacencies with the highest support and collect them in a layout graph. We want to discover those adjacencies from the contig adjacency graph that are most likely present in the true order of the regular contigs. Since we do not resolve repetitive contigs at this stage, the result should be a set of linear chains of the contigs which can also be present in the form of one or several cycles. Our algorithm processes all edges between regular contigs in decreasing weight order and greedily integrates them one by one into an initially empty layout graph, except if any of the involved contig connectors is already used. With this heuristic approach we generate multiple fragments of good adjacencies that are in general joined to larger chains during the course of the algorithm.

This procedure finds for most regular contigs appropriate neighbors. However, if very small contigs lie between two large contigs then we sometimes observe a *shadowing effect*,

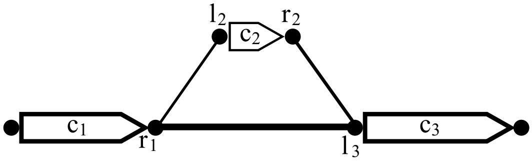


Figure 1: Shadowing effect: If a small contig c_2 is on a reference genome located between the larger contigs c_1 and c_3, in the contig adjacency graph the correct edges to c_2 can have a lower weight than the edge $\{r_1, l_3\}$.

as illustrated in Figure 1: The adjacency edge between the large contigs can have a high weight that shadows the edge weights to the small contig. Thus, the algorithm would not include the small contig into the layout graph. This behavior is generally unwanted but, as we will see in Section 4.2, it can be advantageous for small repetitive contigs. That is why we do not abandon the effect, e. g. by ignoring the size of the matches in the weight function. Instead, we compensate the shadowing effect for the affected regular contigs by integrating them into the initial layout as good as possible. Therefore, we look at all edges that were not integrated in step 2. Again, in decreasing weight order we include an edge if any of the two connectors is still unused in the layout. To control that only very specific edges edges are incorporated, we test if the additional edge has a high relative support $\mathcal{S}^{\text{rel}}$ of at least τ_1. Although the shadowing edge stays in the layout, in most cases the correct edges from the small contig will also be included, resulting in a triangle shape of connections as in Figure 1.

4.2 Adding the repetitive contigs

Starting with the basic layout graph of the previous subsection, the task is now to include the repetitive contigs into the layout. For genome finishing, the gain through repetitive contigs is only limited since they are not well suited for a primer-based closing of gaps. Primers for the repetitive sequence will bind unspecifically to several regions on the genome and should thus be avoided. Nonetheless, we believe that it is very helpful in the finishing phase of a sequencing project for a researcher to be informed which repetitive contigs interrupt the gap between two regular contigs. However, the order of the repetitive contigs in a gap plays, to our opinion, only a secondary role because this information can not directly help in the finishing process: If both primers are based on repetitive contigs, this will produce even more unpredictable results. Our idea in Algorithm 1 is therefore to place each repetitive contig as often as necessary between the corresponding regular contigs into the basic layout graph.

The important edges that we want to integrate in our basic layout are those which connect a repetitive contig with a regular one, see line 1. We demand that the relative support of these edges with respect to the repetitive contig is higher than a threshold τ_2. This avoids the incorporation of arbitrarily weak edges. The edges between repetitive contigs are not

Algorithm 1: Repetitive contigs integration algorithm

```
Input: set of contigs $\mathcal{C}$, set of repetitive contigs $\mathcal{R} \subset \mathcal{C}$, contig adjacency graph
        $G = (V, E)$, basic layout graph $G_L$
Output: repeat-aware layout graph $G_L$ of the contigs
1  let $E_{\text{rep}} = \{\{v, v'\} \mid contig(v) \in \mathcal{R}, contig(v') \notin \mathcal{R}$ and $\mathcal{S}_v^{\mathsf{rel}}(\{v, v'\}) > \tau_2\}$
2  foreach edge $e \in E_{\text{rep}}$, sorted by decreasing weight $w(e)$ do
3      if $e = \{v_1, l\}$ contains the left connector $l$ of a contig $c \in \mathcal{R}$ then
4          let $r$ be the right connector of contig $c$
5          if exists $v_2 = \arg\max_{v \in V} \{w(\{v_1, v\}) \mid \{r, v\} \in E_{\text{rep}}\}$ then
6              duplicate $l$ and $r$ to $l'$ and $r'$
7              $V_L = V_L \cup \{v_1, l', r', v_2\}$
8              $E_L = E_L \cup \{\{v_1, l'\}, \{l', r'\}, \{r', v_2\}\}$
9              remove $\{v_1, l\}$ and $\{r, v_2\}$ from $E_{\text{rep}}$
10         end
11     else // $e = \{r, v_1\}$ contains the right connector of a contig $c \in \mathcal{R}$
12         perform lines 4 to 10 analogously
13     end
14 end
```

considered in this approach, as motivated above. For the interesting edges, we try to find for each involved regular contig connector a suitable counterpart that is also connected to the other end of the repetitive contig, as shown in lines 4 to 10 for the left connectors. This procedure is based on the following observation: As illustrated in Figure 2, a repetitive contig $c \in \mathcal{R}$ usually has several good edges for its right and its left connector leading to different regular contigs. The problem is to determine which edges belong to a particular repeat occurrence on the reference genome. The shadow effect, which was an obstacle for regular contig ordering, becomes here now an advantage. In the example of Figure 2, the edge $\{r_1, l_5\}$ has a high weight if the contigs c_1 and c_5 are only separated by the occurrence of the relatively small repeating contig c. The strategy is hence to search, for any regular node that is connected to one side of a repetitive contig, for a counterpart that

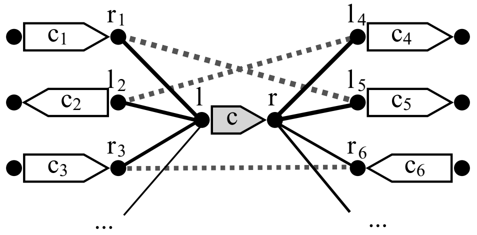


Figure 2: Typical scenario for the adjacency edges of a repetitive contig $c \in \mathcal{R}$. The dashed lines depict the best edge from a contig connector on the right to a contig connector on the left.

is connected to the other side, such that the edge from the node to the counterpart has the highest weight.

This way, we find for each significant occurrence of a repetitive contig the two surrounding regular contigs with respect to the reference genome. For all occurrences we add two new connectors of the repetitive contig and the appropriate edges to the layout graph.

5 Results

We evaluated our algorithm on two biological datasets. The aim was to correctly place the repetitive contigs between the regular contigs allowing them to appear more than once. None of the programs for comparative contig arrangement that we are aware of was designed to handle repetitive contigs explicitly. Still, we applied some of them to our data in order to see whether they would recover a part of the connections nevertheless. After introducing the datasets, we will explain the evaluation procedure and then show the results.

Dataset For the evaluation we prepared a set of contigs and acquired two reference genomes. All genomic sequences belong to the *Corynebacteria* genus. The contigs originate from a 454 sequencing run of the *Corynebacterium urealyticum* strain DSM7109 conducted at the Center for Biotechnology (CeBiTec) of Bielefeld University. The assembly yielded 223 contigs with a total size of 2 316 966 bases. We discarded all ‘contigs’ with a size of less than 500 bases resulting in a set of 69 contigs. This step was taken since many small contigs that consist of only two or three reads can be very confusing in the mapping process. Nevertheless, the N50 contig size, which is a more robust characterization for the size distribution of contig sets than the mean or median, stays the same for both sets.
As reference sequences we took the already finished genome of *Corynebacterium urealyticum* strain DSM7109 [TTT$^+$08] (NCBI Number NC_010545) and the closely related genome of *Corynebacterium jeikeium* K411 (NC_007164).

To ease the evaluation we renumbered and renamed the 69 contigs. Therefore, the contigs were mapped onto the perfect reference, *C. urealyticum*, and ordered according to their matches using the tool r2cat [HS10b]. The program revealed that 15 contigs are repetitive while the remaining 54 contigs were regular according to Section 3. The regular contigs were renumbered consecutively in their true order such that adjacencies can easily be seen. The repetitive contigs were numbered in arbitrary order and prefixed with the letter ’r’. As the next step we manually inspected the matches using r2cat and noted for each pair of adjacent regular contigs which repetitive contigs have an occurrence between them.

Experimental Setup The manually annotated list of repeating contigs between regular contigs serves for the experiments as a standard of truth. To see which of these connections can actually be discovered, we applied the following programs on the described datasets: Projector2 [vHZKK05], OSLay [RSH07], PGA [ZZLB08], treecat [HS10a] and our new

Table 1: Experimental results of ordering the repetitive *C. urealyticum* contigs. Each program was applied on the described datasets. The results of PGA were varying, since it is a randomized algorithm, so we give the mean values for applying the program 20 times.

	perfect reference				*C. jeikeium* reference			
Program	TP	FP	TPR	PPV	TP	FP	TPR	PPV
Projector2	10.0	0.0	0.06	1.00	1.0	5.0	0.01	0.17
OSLay	15.0	1.0	0.09	0.94	4.0	2.0	0.03	0.67
PGA	24.4	16.1	0.15	0.60	20.4	27.9	0.13	0.42
treecat	29.0	1.0	0.18	0.97	20.0	8.0	0.13	0.71
repcat	140.0	7.0	0.89	0.95	54.0	37.0	0.34	0.59

algorithm *repcat* (repeat-aware contig arrangement tool) which is derived from treecat. All programs were used with their standard parameters as proposed in the publication or as pre-given on the web-service unless otherwise stated. At the webform of Projector2 we switched off repeat masking for contigs and target genome and reduced the minimum contig size to 500 bases such that all contigs could be considered. PGA, unfortunately, filters all contigs smaller than 3.5 kb and discards this way all repetitive contigs, so we modified their perl script to include all contigs. The algorithm from *repcat* is adopted from treecat and inherits some of its parameters for the contig adjacency graph creation: We set the standard deviation of the insertion/deletion size to $\sigma_1 = 2\,000$ bases and the lost fragment weighting factor φ to 0. Furthermore, for the repeat detection we used the default parameters stated in Section 3 and selected for the layouting the experimentally evaluated parameters $\tau_1 = 90\%$ and $\tau_2 = 0.1\%$.

In the following evaluation, we assess how well repetitive contigs can be integrated into an ordering. As motivated in Section 4.2, the interesting connections are those from a repetitive contig to a regular contig. Therefore, we extracted those connections from the output of the programs and compared them with our manually annotated standard of truth. If a repetitive contig is present in between a gap, we count the connections to the corresponding regular contigs as true positives (TP). If an adjacency is not given in our list, we count this as a false positive (FP). For example, if the repetitive contigs r008, r012, and r013 are between 048 and 049, we count a connection from r013 to 048 as TP, whereas a connection from r003 to 049 is a FP, except if for example r003 would occur between 049 and 050 as well.

Evaluation We ran all programs on both datasets and counted the TP and FP values as described above. The manually annotated repeat list revealed that the 15 repetitive contigs occurred in 79 instances on the genome. For each occurrence two true positive connections could be predicted, so the sum of all positive predictions is $\mathrm{P} = 158$. Given TP, FP and P we calculated the *sensitivity* (also called *true positive rate*, $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}}$) and the *precision* (also called *positive predictive value*, $\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$) of the predicted connections. Table 1 shows the resulting values for each program and dataset.

We would like to note here that a direct comparison of the values between the different programs has to be handled with caution. While our algorithm was specially built to include repeating contigs as often as necessary into a layout, the objectives of Projector2 and OSLay are to devise a linear ordering where each contig occurs exactly once. These programs can generate at most two true positive connections per repetitive contig, that is a maximum of $\text{TP}_{max} = 30$ for our dataset. PGA combines five such linear layouts and achieves thus at most $\text{TP}_{max} = 150$ correct connections. The program treecat is not restricted by this number but will stop to add edges if both connectors of a repetitive contig have been integrated into a layout. Our new algorithm can in principle predict two true positive connections for *every occurrence* of a repetitive contig, thus gaining a clear advantage over the other programs in this setting. Regardless of that, we believe that an appropriate placing of repetitive contigs is very helpful for a sequencing project and the alternative to mask and discard repeats is not a sufficient solution.

The results for the perfect reference show that the predictions are in general quite accurate. Projector2 and OSLay find only a fraction of their TP_{max}. PGA and treecat recover some more true positives, but PGA surprises with a rather high number of false positives. Our new algorithm *repcat* recovers nearly 90% of the possible true positive connections. This is somehow expected, since it is the only of the applied methods that handles repeats explicitly.

For the more realistic reference *C. jeikeium*, the true positives decrease for all programs while the false positives increase. Especially the results of our algorithm are hit by this tendency, although it still finds many more of the correct repeat adjacencies than any of the other programs.

6 Conclusion

In the context of ordering contigs to assist the gap closure of prokaryotic sequencing projects, we propose a novel algorithm that includes an explicit handling of repetitive contigs. While the common objective of related applications is to find a linear layout, this is obviously not feasible for repetitive contigs. Hence, our approach orders the non-repetitive contigs first and then integrates all repetitive contigs in between the gaps, as often as necessary. We believe that this strategy is more adequate than discarding all repetitive contigs since it allows to assess which of these sequences should be expected in the gaps.

In this setting, the contig adjacency graph turns out to be a valuable concept that is flexible enough to be extended to handle repetitive contigs. However, the reduction to a layout graph always contains the risk of losing important adjacencies. An interactive visualization of the whole graph could help to unleash its true potential.

Considering the problem of repetitive contigs, a next step could be to verify to which degree of synteny an ordering of repetitive elements is feasible and if the information from several reference genomes helps or maybe even confuses.

Acknowledgments

We thank Eva Trost and Andreas Tauch for providing the sequence data.

References

[HS10a] P. Husemann and J. Stoye. Phylogenetic comparative assembly. *Algorithms Mol. Biol.*, 5(1):3, 2010.

[HS10b] P. Husemann and J. Stoye. r2cat: synteny plots and comparative assembly. *Bioinformatics*, 26(4):570–571, 2010.

[Mar08] E. R. Mardis. The impact of next-generation sequencing technology on genetics. *Trends Genet.*, 24(3):133–141, 2008.

[MKS10] J. R. Miller, S. Koren, and G. Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, 2010.

[Pop09] M. Pop. Genome assembly reborn: recent computational challenges. *Brief. Bioinform.*, 10(4):354–366, 2009.

[RSH07] D. C. Richter, S. C. Schuster, and D. H. Huson. OSLay: optimal syntenic layout of unfinished assemblies. *Bioinformatics*, 23(13):1573–1579, 2007.

[TTT$^+$08] A. Tauch, E. Trost, A. Tilker, U. Ludewig, S. Schneiker, A. Goesmann, W. Arnold, T. Bekel, K. Brinkrolf, I. Brune, S. Götker, J. Kalinowski, P.-B. Kamp, F. Pereira Lobo, P. Viehoever, B. Weisshaar, F. Soriano, M. Dröge, and A. Pühler. The lifestyle of *Corynebacterium urealyticum* derived from its complete genome sequence established by pyrosequencing. *J. Biotechnol.*, 136(1-2):11–21, 2008.

[vHZKK05] S. A. F. T. van Hijum, A. L. Zomer, O. P. Kuipers, and J. Kok. Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies. *Nucleic Acids Res.*, 33:W560–W566, 2005.

[ZZLB08] F. Zhao, F. Zhao, T. Li, and D. A. Bryant. A new pheromone trail-based genetic algorithm for comparative genome assembly. *Nucleic Acids Res.*, 36(10):3455–3462, 2008.

CASOP GS: Computing intervention strategies targeted at production improvement in genome-scale metabolic networks

Katrin Bohl[1,2,†], Luís F. de Figueiredo[1,†], Oliver Hädicke[3], Steffen Klamt[3],
Christian Kost[2], Stefan Schuster[1] and Christoph Kaleta[1,*]
[1] Department of Bioinformatics, Friedrich Schiller University Jena,
Ernst-Abbe-Platz 2, D-07743 Jena, Germany,
[2] Department of Bioorganic Chemistry, Max Planck Institute for Chemical
Ecology, D-07745 Jena, Germany,
[3] Max Planck Institute for Dynamics of Complex Technical Systems,
Sandtorstrasse 1, D-39106 Magdeburg, Germany
[*]Corresponding author e-mail: christoph.kaleta@uni-jena.de
[†]Both authors contributed equally

Abstract: Metabolic engineering aims to improve the production of desired biochemicals and proteins in organisms and therefore, plays a central role in Biotechnology. However, the design of overproducing strains is not straightforward due to the complexity of metabolic and regulatory networks. Thus, theoretical tools supporting the design of such strains have been developed. One particular method, CASOP, uses the set of elementary flux modes (EFMs) of a reaction network to propose strategies for the overproduction of a target compound. The advantage of CASOP over other approaches is that it does not consider a single specific flux distribution within the network but the whole set of possible flux distributions represented by the EFMs of the network. Moreover, its application results not only in the identification of candidate loci that can be knocked out, but additionally proposes overexpression candidates. However, the utilization of CASOP was restricted to small and medium scale metabolic networks so far, since the entire set of EFMs cannot be enumerated in such networks. This work presents an approach that allows to use CASOP even in genome-scale networks. This approach is based on an estimation of the score utilized in CASOP through a sample of EFMs within a genome-scale network. Using EFMs from the genome-scale metabolic network gives a more reliable picture of the metabolic capabilities of an organism required for the design of overproducing strains. We applied our new method to identify strategies for the overproduction of succinate and histidine in *Escherichia coli*. The succinate case study, in particular, proposes engineering targets which resemble known strategies already applied in *E. coli*. **Availability:** Source code and an executable are available upon request.

1 Introduction

Using microorganisms to overproduce certain metabolites and proteins is the central objective of metabolic engineering [Lee09]. While many applications consider the improvement of the production of native compounds, there is also an increasing number of attempts in which entire heterologous pathways have been engineered [NKL10, AS09]. Small molecule compounds whose production have been engineered span from alcohols and lipids, for instance, used in bio-fuel production, to plastics and pharmaceuticals. Usually, the design of strains that overproduce a desired target compound involves a large number of modifications of the metabolic and regulatory network including gene knockouts/knockins and the overproduction of proteins [KJSW10]. Yet the complexity of metabolic and regulatory networks associated with the cost and effort to manipulate organisms is still a major challenge in the development of improved production designs. A large set of theoretical tools has been developed that aim at simplifying this process [BPM03, KR10, TS10]. A common feature of these methods is the prediction of metabolic flux distributions prior and after perturbations. These predictions are then combined with either deterministic or stochastic procedures that try to identify knockout and knockin combinations that improve the production of the target compound with as few genetic modifications as possible.

Recently, a new method called **C**omputational **A**pproach for **S**train **O**pti-mization aiming at high **P**roductivity (CASOP, [HK10]) based on the concept of elementary flux modes (EFMs, [SDF99]) has been proposed. However, this approach has been limited to small and medium-scale metabolic networks, so far, since the enumeration of all EFMs can only be performed in such networks [KS02]. In this work we want to outline an approach that allows us to circumvent this limitation of CASOP. Instead of computing the scores utilized in CASOP from the entire set of EFMs we compute them from a subset of the EFMs in a genome-scale network. This subset of EFMs is obtained from a sampling procedure that is similar to a previously described method to enumerate EFMs in genome-scale metabolic networks [KdFBS09]. Using our approach, CASOP can be applied even to genome-scale networks.

2 Methods

2.1 CASOP

Using the EFMs of a reaction network, CASOP calculates for each reaction a Z_2-score that indicates whether the flux through this reaction needs to be increased or decreased, in order to improve the production of a particular target compound. Similar to other methods, the reasoning behind CASOP is that the organism tries to optimize its growth yield. However, in contrast to most methods, CASOP does not assume that the organism attains the optimal flux, but rather uses a combination of optimal and, to a certain extent, sub-optimal flux distributions.

Computing CASOP-scores

In the following we give a short overview over CASOP. For a more detailed description see [HK10]. In order to determine scores for the knockout or overexpression of enzymes, CASOP considers two versions of a metabolic network that contain a biomass reaction defining the proportion of building blocks the organism requires for its reproduction. The first network corresponds to the wild-type model. In the second network, the biomass reaction is coupled with the production of the target metabolite such that, in weights, 10% of biomass and 90% of the target metabolite are consumed. EFMs are computed in both networks. Afterward each EFM i is assigned a weight ν_i that depends on its yield in the biomass reaction, $Y^i_{Biomass/S}$ (ratio between carbon source inflow and flux through the (modified) biomass reaction). The weights of the EFMs are adjusted using a parameter k such that increasing values of k attribute higher weights to EFMs with higher yields. In this work we used a value of $k = 5$.

Afterward, a reaction importance measure is computed for each reaction in both networks as the sum of the weights of the EFMs containing this reaction. As the name suggests, the reaction importance measure allows to assess the impact of a perturbation of an enzyme catalyzing it on the production of biomass and/or the target metabolite. If one reaction has a high importance within the network containing the production of the target metabolite, but a low importance in the other network, this reaction is a candidate for overexpression since increasing the flux through it increases the flux through EFMs producing the target metabolite. In contrast, a reaction that has a low importance for the production of the target metabolite, but a high importance for sole biomass production can be removed, since it favors the flux through EFMs that do not produce the target metabolite. Hence, the Z_2-scores of CASOP, that indicate candidates for knockout and overexpression, are computed as the difference between the reaction importances of each reaction between the two networks. These scores take values between -1 and +1. A positive score indicates a reaction that is candidate for overexpression and a negative score indicates a knockout candidate. Please note that, in contrast to [HK10] we split reversible reactions in irreversible forward and backward directions. Thus, reversible reactions are assigned two CASOP scores allowing us to assess the role of forward and backward direction separately.

Building on the Z_2-scores, the CASOP procedure then knocks out the enzymes in silico that catalyzes the reaction with the most negative score by removing all EFMs containing this reaction (or other reactions catalyzed by this enzyme). Subsequently, the Z_2-scores are recomputed for the reduced set of EFMs and the procedure is iterated.

Assessing the production of the desired product

CASOP allows one to assess the impact of a genetic modification on the production of a specific target metabolite. However, no statement about the change of the production after several consecutive modifications, such as multiple knockouts, is possible. In order to observe the improvement in the production of the target metabolite, we introduce the measure Y_M which allows us to assess the relative change in yield of metabolite M after several knockouts. We make use of the weights ν_i that CASOP assigns to each EFM i (see [HK10]) in the network in which the production of the target metabolite is not associated with biomass production. Given a set of n EFMs in this network with the individual yields in the target metabolite $Y^i_{M/S}$ of each EFM i, we derive Y_M as

$$Y_M = \sum_{i=1}^{n} \nu_i \cdot Y^i_{M/S}.$$

Since we multiply the weight of each EFM with the production of the target metabolite, Y_M can be considered as a weighted average of the yields of the EFMs in the target metabolite. If Y_M increases after a knockout, we expect this knockout to increase the production of the target metabolite M. Note that Y_M does not correspond to an actual yield, but serves as an indicator of the effect of a knockout strategy.

2.2 Enumeration of EFMs in genome-scale metabolic networks

Until recently, the computation of EFMs has been limited to small and medium-scale metabolic networks. However, fluxes within small-scale networks might be inconsistent with the corresponding fluxes within the underlying genome-scale network [KdFS09]. Several approaches for the computation of EFMs in genome-scale metabolic networks have been developed. One approach, the so-called K-shortest procedure, computes EFMs in increasing number of reactions [dFPR$^+$09]. Another approach, the EFMEvolver [KdFBS09] uses a genetic algorithm to sample large numbers of EFMs in these networks more efficiently.

Here we used a more direct approach than EFMEvolver to compute EFMs. The similarity between both methods concerns the linear programming formulation to compute a single EFM given a metabolic network (for more details see [KdFBS09]). However, instead of using a genetic algorithm, we used an iterative procedure to enumerate EFMs. Starting from an initial EFM using the target reaction, one of its reactions is selected randomly. Subsequently this reaction is blocked by setting its flux to zero and therefore, a new EFM is computed by solving the linear programming formulation. Iterating this procedure, several EFMs are obtained while the number of blocked reactions increases. If no EFM is found given a particular set of blocked reactions, the last reaction is removed from this set. Additionally, with a small probability, all reactions are removed from the set of blocked reactions. This procedure allows one to increase the diversity of the EFMs that are detected since resetting the set of blocked reactions corresponds to initializing a new independent sampling procedure. More details on the sampling procedure will be given elsewhere.

3 Results and Discussion

We applied our method to two cases: the production of succinate from glucose (studied in [HK10]) and the production of histidine from fructose in *Escherichi coli*. In each case, we started with an initial sample of 10^6 EFMs for the two networks that are required in our procedure. As a genome-scale metabolic model of *E. coli*, we used *i*AF1260 [FHR$^+$07]. Besides the carbon source, we supplied the network with the following compounds: NH_4^+, NO$_3^-$, $S0_4^{2-}$, Fe^{2+}, Fe^{3+}, CO$_2$, H$^+$, K$^+$, Ca^{2+}, cobalt, molybdate, Na$^+$, Pi, O$_2$, H$_2$O, Cl$^-$, Cu^{2+}, Mg^{2+}, Mn^{2+} and Zn^{2+} that are required for the survival of the cell.

3.1 Case study I: Succinate production

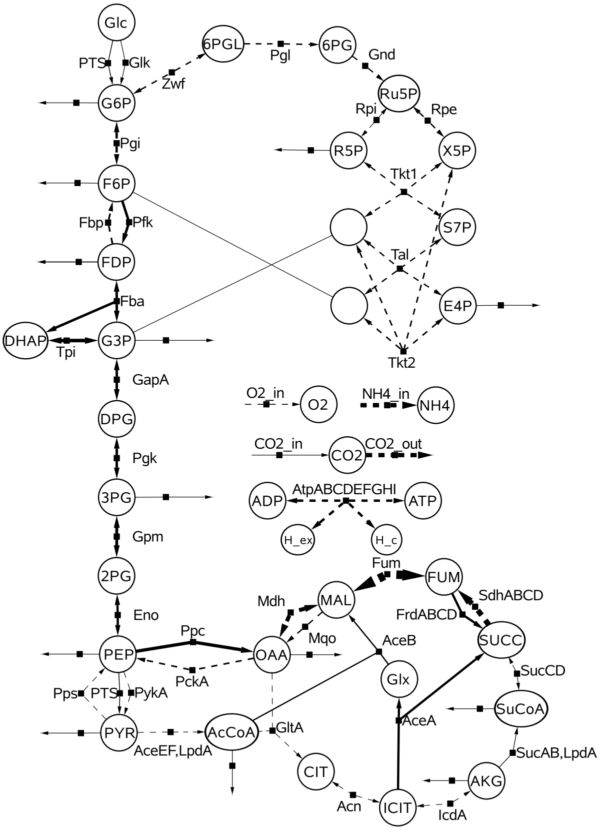


Figure 1: Z_2 scores of central metabolism of the wild-type network for succinate overproduction. The width of the arrows corresponds to the values of the scores. Dashed lines indicate negative scores, bold lines positive scores. Metabolite nodes connected by straight lines are identical. A list of abbreviations can be found in the supplementary material of [FHR$^+$07].

The Z_2-scores for reactions within central metabolism are displayed in Fig. 1. While the relative scores for many reactions matched those discussed in [HK10] there were some differences. For instance, reactions of the glyoxylate shunt have high overexpression ratings, while this was not the case in [HK10]. The importance of such a modification to increase succinate production has been demonstrated by [LBS05]. Additionally, the overexpression of Ppc, as indicated by our analysis, is also known to improve succinate production [LBS05]. Most interestingly, fumarase (Fum) that reversibly converts fumarate into malate received the highest knockout rating. This case exemplifies the advantage of computing the Z_2-scores of both directions of reversible reactions independently. In [HK10] both directions of reversible reactions were not considered independently and, in consequence, the score of Fum was relatively low. However, knocking out fumarase increases Y_{SUCC} almost ten-fold (Fig. 2A). This strong increase in production is probably due to the fact that this deletion interrupts the TCA cycle. In consequence, the concentration of fumarate increases which entails an increase in the concentration of the desired target metabolite succinate. Furthermore, fumarate, which is a side-product of several biosynthetic pathways, can only be disposed through conversion into succinate after this knockout if fumarate is not excreted.

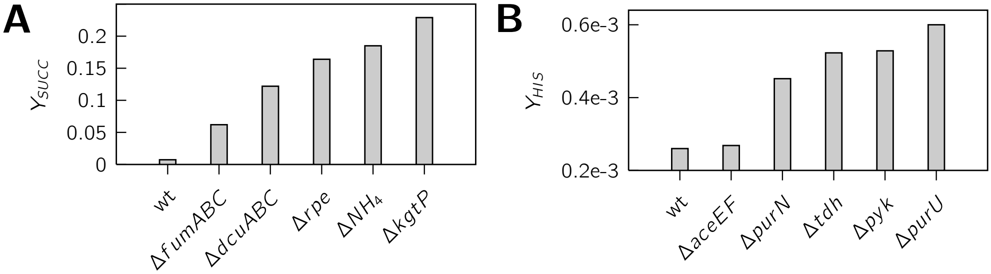


Figure 2: Y_M in the two case-studies. Knockouts are cumulative from left to right. **A** Succinate production. **B** Histidine production.

After knocking out fumarase, the fumarate transporter (Dcu) that exists in 3 isoforms, received the lowest Z_2-score. Knocking out the corresponding genes yielded a strain in which succinate production is coupled to growth. That is, biomass can only be produced when co-producing succinate. Interestingly, after this knockout, the Z_2-score of the succinate dehydrogenase SdhABCD, which is known to improve succinate production [LBS05] and had the second lowest score in the wild-type, indicates that there is no influence of a SdhABCD knockout on the production of succinate anymore. Thus, the knockout of fumarase and the fumarate transporter appears to represent an alternative knockout strategy to the knockout of succinate dehydrogenase.

In the next step, the ribulose-5-phosphate-3-epimerase (Rpe) was suggested for knockout. The forth proposed knockout involves the inflow reaction of ammonium. Knocking out this reaction corresponds to removing ammonium from the growth medium. This modification is not lethal, since we provided nitrate as alternative nitrogen source, but at the expense of reducing the growth rate [BZ90]. Moreover, nitrate can only serve as nitrogen source in the absence of oxygen [LK87]. Indeed, oxygen inflow is also assigned a relatively low Z_2-score. This indicates that the utilization of nitrate as electron acceptor and ammonium source under anaerobic conditions can improve succinate production. The fifth proposed knockout removed the export of α-ketoglutarate further reducing the number of possible pathways to excrete TCA cycle intermediates besides succinate.

3.2 Case study II: Histidine production

As a second case study we examined the production of histidine from fructose (Fig. 2B). In the first step, pyruvate dehydrogenase was suggested as knockout (Fig. 3). In the second step, the phosphoribosylglycinamide formyltransferase (PurN) was suggested as knockout. Removing this reaction drastically increased Y_{HIS} (Fig. 2B). This knockout illustrates the need of considering all reactions within a genome-scale metabolic network as knockout candidates. The reaction catalyzed by PurN consumes 10-Formyltetrahydrofolate (10-FTHF) as a co-factor. However, 10-FTHF is also required for histidine biosynthesis. In purine biosynthesis, the reaction catalyzed by PurN can also be catalyzed by the transformylase PurT that uses formate rather than 10-FTHF. Thus, knocking out PurN increases the 10-FTHF pool available for histidine biosynthesis. Furthermore, a strain with a PurN knockout grows slower than the wild-type [BAH^{+}06], indicating that the capacity of purine production might be reduced. This is of additional advantage for histidine production, since 5-Phospho-α-D-ribose-1-diphosphate (PRPP) is a common precursor of histidine and purine biosynthesis. In the following two steps, threonine dehydrogenase and pyruvate kinase were knocked out. Especially, the knockout of the threonine dehydrogenase is of interest, since it removes one of the two pathways of glycine biosynthesis from threonine. Thus, glycine biosynthesis via serine might be increased which in turn increases the cellular 10-FTHF pool whose major source is glycine biosynthesis via serine. In the fifth step, the formyltetrahydrofolate deformylase PurU that converts 10-FTHF to formate and tetrahydrofolate was knocked out.

3.3 Influence of sample sizes on Z_2-scores

In order to test the reliability of the Z_2-scores we obtained using a sample of 10^6 EFMs (Sample A), we recomputed the scores for independent samples with a higher number of EFMs: $2 \cdot 10^6$ EFMs (Sample B) and $3.7 \cdot 10^6$ EFMs (Sample C). The maximum deviations over the five knockouts between sample A and B increased over the knockout depth from 0.05 to 0.09 after the forth knockout. In all cases this maximum deviation was smaller between sample B and sample C. Here, the maximum deviation was 0.07. Slight deviations occurred in the order by which the reactions were knocked out in the three samples. After the forth knockout, the export of pyruvate rather than α-ketoglutarate received the lowest Z_2-score in the larger samples. Thus, the Z_2-scores are relatively robust if sample sizes are sufficiently large. However, for greater knockout depths, larger samples of EFMs might be required.

4 Conclusions

In this work we have presented CASOP GS as an approach that allows one to apply CASOP to genome-scale metabolic networks. Furthermore, we have introduced a mea-

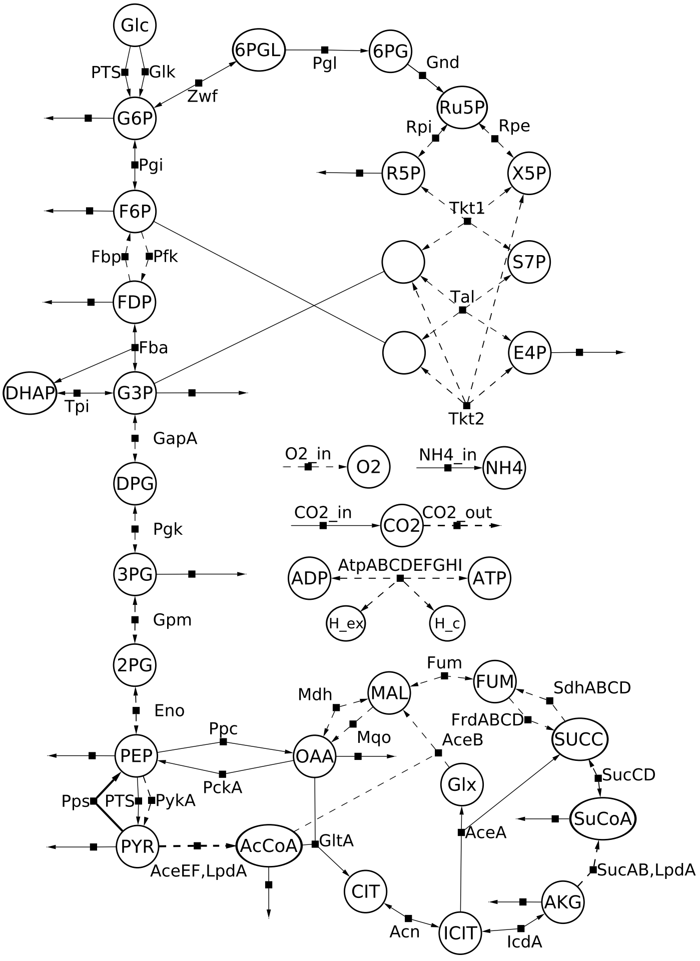


Figure 3: Z_2 scores for histidine production. For details see Fig. 1.

sure that allows one to assess the changes in the production of a target metabolite after multiple genetic modifications. Besides these improvements of CASOP, our approach offers several important advantages over other theoretical methods for strain improvement such as OptKnock [BPM03], OptGene [PRFN05] and other recently proposed approaches [KR10, TS10].

First, and most importantly, our approach provides the user with a ranking of reactions whose removal/overexpression improves the production of a target metabolite. Thus, rather than presenting a complete knockout strategy, the user has the possibility to choose, which reaction is most suitable for knockout or overexpression. This is of particular importance for the incorporation of prior knowledge about difficulties and side-effects of certain gene-manipulations. For example the principal knockout candidate might require removing a gene whose deletion is known to cause pleiotropic effects (e.g. a slow growth rate), while the second rated knockout might yield a strain with a only slightly reduced growth rate.

Second, some approaches only consider a specific part of the metabolic network due to computational limitations. In contrast, our approach takes all reactions within an organism into account. In consequence, we do not only identify candidates for knockouts in the primary metabolism, but also in other parts of the metabolism. This is of particular importance for the overproduction of histidine, since reactions from nucleotide and amino acid metabolism appear to be suitable knockout targets.

Third, most approaches concentrate only on knockouts, while our approach, since it is an extension of CASOP, also proposes overexpression candidates to increase the production of the target metabolite. This is important since the overexpression of genes is frequently used for strain improvement.

CASOP GS offers many advantages over other approaches for the design of production strains. However, a shortcoming is that the regulatory network is not considered. In order to circumvent this problem, we are currently working on an improved version that takes into account regulatory rules by only allowing for EFMs that are consistent with the regulation of metabolism. This, regulation will be implemented in the form of Boolean logic. Moreover, the proposed knockouts of the histidine case study are currently being implemented in *E. coli* in order to validate our results.

References

[AS09] H. Alper and G. Stephanopoulos. Engineering for biofuels: Exploiting innate microbial capacity or importing biosynthetic potential? *Nat Rev Microbiol*, 7(10):715–723, Oct 2009.

[BAH$^+$06] T. Baba, T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol*, 2:2006.0008, 2006.

[BPM03] A. P. Burgard, P. Pharkya, and C. D. Maranas. Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng*, 84(6):647–657, Dec 2003.

[BZ90] H. J. Brons and A. J. Zehnder. Aerobic nitrate and nitrite reduction in continuous cultures of *Escherichia coli* E4. *Arch Microbiol*, 153(6):531–536, 1990.

[dFPR$^+$09] L. F. de Figueiredo, A. Podhorski, A. Rubio, C. Kaleta, J. E. Beasley, S. Schuster, and F. J. Planes. Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics*, 25(23):3158–3165, Dec 2009.

[FHR$^+$07] A. M. Feist, C. S. Henry, J. L. Reed, M. Krummenacker, A. R. Joyce, P. D. Karp, L. J. Broadbelt, V. Hatzimanikatis, and B. Ø. Palsson. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol*, 3:121, 2007.

[HK10] O. Hädicke and S. Klamt. CASOP: A computational approach for strain optimization aiming at high productivity. *J Biotechnol*, 147(2):88–101, May 2010.

[KdFBS09] C. Kaleta, L. F. de Figueiredo, J. Behre, and S. Schuster. EFMEvolver: Computing elementary flux modes in genome-scale metabolic networks. In I. Grosse, S. Neumann, S. Posch, F. Schreiber, and P. Stadler, editors, *Lecture Notes in Informatics - Proceedings*, volume P-157, pages 179–189, Bonn, 2009. Gesellschaft für Informatik.

[KdFS09] C. Kaleta, L. F. de Figueiredo, and S. Schuster. Can the whole be less than the sum of its parts? Pathway analysis in genome-scale metabolic networks using elementary flux patterns. *Genome Res*, 19(10):1872–1883, Oct 2009.

[KJSW10] S. Kind, W. K. Jeong, H. Schröder, and C. Wittmann. Systems-wide metabolic pathway engineering in *Corynebacterium glutamicum* for bio-based production of diaminopentane. *Metab Eng*, Apr 2010. In print.

[KR10] J. Kim and J. L. Reed. OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. *BMC Syst Biol*, 4(1):53, Apr 2010.

[KS02] S. Klamt and J. Stelling. Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep*, 29(1-2):233–236, 2002.

[LBS05] H. Lin, G. N. Bennett, and K.-Y. San. Metabolic engineering of aerobic succinate production systems in *Escherichia coli* to improve process productivity and achieve the maximum theoretical succinate yield. *Metab Eng*, 7(2):116–127, Mar 2005.

[Lee09] S. Y. Lee. Systems biotechnology. *Genome Inform*, 23(1):214–216, Oct 2009.

[LK87] E. C. C. Lin and D. R. Kuritzkes. *Escherichia coli and Salmonella typhimurium - Cellular and Molecular Biology*, volume I, chapter 16 - Pathways for anaerobic electron transport, pages 201–221. ASM, Washington, 1987.

[NKL10] D. Na, T. Y. Kim, and S. Y. Lee. Construction and optimization of synthetic pathways in metabolic engineering. *Curr Opin Microbiol*, Mar 2010. In print.

[PRFN05] K. R. Patil, I. Rocha, J. Förster, and J. Nielsen. Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics*, 6:308, 2005.

[SDF99] S. Schuster, T. Dandekar, and D. A. Fell. Detection of elementary flux modes in biochemical networks: A promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol*, 17(2):53–60, Feb 1999.

[TS10] N. Tepper and T. Shlomi. Predicting metabolic engineering knockout strategies for chemical production: Accounting for competing pathways. *Bioinformatics*, 26(4):536–543, Feb 2010.

Predicting miRNA targets utilizing an extended profile HMM

Jan Grau[1,*], Daniel Arend[1], Ivo Grosse[1], Artemis G. Hatzigeorgiou[2], Jens Keilwagen[3], Manolis Maragkakis[1,2], Claus Weinholdt[1], and Stefan Posch[1]

[1] Institute of Computer Science, Martin Luther University Halle–Wittenberg, Germany
[2] Institute of Molecular Oncology, Biomedical Sciences Research Center 'Alexander Fleming', Vari, Greece
[3] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany

Abstract: The regulation of many cellular processes is influenced by miRNAs, and bioinformatics approaches for predicting miRNA targets evolve rapidly. Here, we propose conditional profile HMMs that learn rules of miRNA-target site interaction automatically from data. We demonstrate that conditional profile HMMs detect the rules implemented into existing approaches from their predictions. And we show that a simple UTR model utilizing conditional profile HMMs predicts target genes of miRNAs with a precision that is competitive compared to leading approaches, although it does not exploit cross-species conservation.

1 Introduction

miRNAs are short ($\sim$ 22 nt) endogeneous RNAs that bind to partially complementary sites on mRNA target sequences. They induce cleavage of the miRNA-mRNA duplex or repress translation of the bound mRNA [BSRC05]. Hence, miRNAs influence gene expression and introduce a novel level of gene regulation. For instance, several miRNA signatures have already been successfully associated with human cancers. In animals, miRNAs preferentially bind to the 3' untranslated region (UTR) of the mRNA, and for binding a high complementarity between miRNA and target is required only at the 5' end of the miRNA. Computational miRNA target prediction plays a key role in deciphering the functional role of miRNAs. Several dozen programs have been therefore developed in the last years, and in the following, we describe the main idea behind some of the most widely used programs.

[LSJR$^+$03] propose an algorithm for the prediction of targets of vertebrate miRNAs called TargetScan. TargetScan requires perfect complementarity between positions 2 and 8 at the 5'-end of the miRNA and a potential target, and the free energy of binding between miRNA and target is computed. Predictions are verified using orthologous UTR sequences from other organisms. [LBB05] propose a refined version called TargetScanS, which demands a shorter region of the target to be complementary to nucleotides $2-7$ of the miRNA. TargetScan 5.0 [FFBB09] additionally considers the distance from the 3' UTR and AU content.

In contrast to TargetScan, miRanda [EJG+03] does not require perfect complementarity at the seed region, but uses an algorithm similar to Smith-Waterman sequence alignment with similarity scores of $+5$ for `G:C` and A:U basepairs, $+2$ for `G:U` basepairs, and -3 for mismatches, and the scores for the first 11 positions of the alignment are weighted by a factor of 2. Potential target sites (TSs) are filtered for a minimum similarity score and a minimum free energy.

PicTar [KGP+05] searches for perfectly complementary seed regions of 7 nt starting from position 1 or 2 of the miRNA. Mismatches in the seed region are allowed if these do not increase the free energy. Additionally, a filter with respect to the free energy of the complete miRNA-mRNA duplex is applied.

DIANA-microT [MRS+09] prefers perfect complementarity of 7 to 9 nt starting from position 1 or 2 of the miRNA. However, if the considered TS shows good complementarity to the 3' end of the miRNA, the length of this seed region may be reduced to 6 nt, and single `G:U` basepairs are allowed. DIANA-microT uses orthologous UTRs from up to 27 organisms for assessing the conservation of TSs. Finally, the score of a potential UTR target is computed as a weighted average of all predicted TSs.

In contrast to previous approaches, we propose a fully statistical approach for predicting TSs of given miRNAs that is capable of learning rules of miRNA-TS binding from data sets comprising pairs of miRNAs and associated TSs. This approach employs an extension of profile hidden Markov models (HMMs) [KBM+94], which we call *conditional profile HMM* (CoProHMM), and learns parameters by the discriminative maximum supervised posterior (MSP) principle [CdM05, GKK+07]. Since all parameters of CoProHMMs are learned from training data, this approach is not biased towards heuristic assumptions about miRNA-TS interaction like the existence or length of a seed region.

2 Methods

In the following, we introduce CoProHMMs for modeling the binding between miRNA and TS. We describe how we learn CoProHMMs from data, and explain how we combine several predictions of a learned CoProHMM to predict target genes of a given miRNA.

2.1 Conditional profile HMMs

At the basis of the CoProHMM modeling miRNA TSs, we use a standard profile HMM architecture [KBM+94], which is illustrated in Fig. 1. This architecture is also referred to as "plan9" due to its 9 transitions at each layer of the model. We define a total of K match states M_k, which emit a nucleotide of the TS with a probability that is conditional on the nucleotide at position k of the miRNA. Here, we use $K = 22$, since this is the length of a typical miRNA and, hence, the model covers all positions of the miRNA that are potentially interacting with the TS. If a TS and the associated miRNA are perfectly complementary, we anticipate that only match states are visited for emitting the complete sequence of the TS. Otherwise, silent delete states D_k allow for the insertion of gaps into

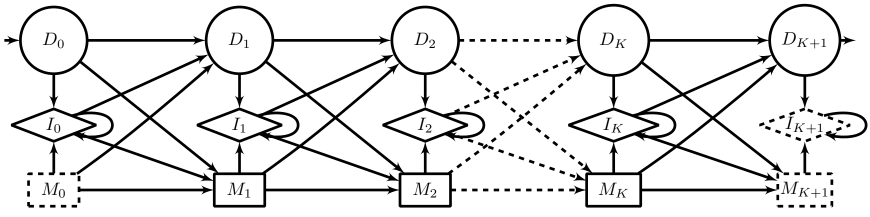


Figure 1: Plan9 architecture of the proposed CoProHMMs. Circles represent silent delete states that do not emit nucleotides of the TS, diamonds represent insert states that emit nucleotides of the TS without considering the nucleotides of the miRNA, and rectangles represent match states that emit nucleotides of the TS with probabilities conditional on the nucleotides of the miRNA. Admissible paths start at D_0 and end at D_{K+1}. States with dashed borders are not visited in admissible paths.

the TS, insert states I_k allow for including gaps in the miRNA, and match states also allow to replace nucleotides. In Fig. 1, edges represent transition probabilities not fixed to 0. From each node of column k, we can reach node I_k in the same column, and nodes M_{k+1} and D_{k+1} in the next column. Each admissible path starts at D_0 and ends at D_{K+1}. Hence, the states M_0, I_{K+1}, and M_{K+1} are never visited in admissible paths, and are only included to simplify recursive definitions in the following.

We parameterize the transition probabilities and the emission probabilities by normalized exponentials [Mac98, BB01] using real-valued parameters, since this allows for an unconstrained numerical optimization of the parameters with respect to the discriminative MSP principle.

According to the plan9 architecture, we define the transition probability $P_T(V|S_k, \boldsymbol{\beta}_{T,S_k})$ of going from node $S_k \in \{I_k, M_k, D_k\}$ to node V given parameters $\boldsymbol{\beta}_{T,S_k}$ as

$$P_T(V|S_k, \boldsymbol{\beta}_{T,S_k}) = \begin{cases} \frac{\exp(\beta_{V|S_k})}{\sum_{\tilde{V} \in \{I_k, M_{k+1}, D_{k+1}\}} \exp(\beta_{\tilde{V}|S_k})} & \text{if } V \in \{I_k, M_{k+1}, D_{k+1}\} \\ 0 & \text{otherwise} \end{cases},$$

where $\boldsymbol{\beta}_{T,S_k} = (\beta_{I_k|S_k}, \beta_{M_{k+1}|S_k}, \beta_{D_{k+1}|S_k}), \beta_{V|S_k} \in \mathbb{R}$.

In contrast to standard profile HMMs, we use conditional probabilities depending on the nucleotides of the miRNA for the emissions of the match states. For match state M_k, we define the conditional emission probability $P_{M_k}(a|r_k, \boldsymbol{\beta}_{M_k})$ of symbol a in the TS given the k-th symbol r_k of the miRNA and parameters $\boldsymbol{\beta}_{M_k}$ as

$$P_{M_k}(a|r_k, \boldsymbol{\beta}_{M_k}) = \frac{\exp(\beta_{a|r_k,M_k})}{\sum_{\tilde{a} \in \Sigma} \exp(\beta_{\tilde{a}|r_k,M_k})}, \tag{1}$$

where $\boldsymbol{\beta}_{M_k} = (\beta_{A|A,M_k}, \beta_{C|A,M_k}, \ldots, \beta_{U|U,M_k}), \beta_{a|b,M_k} \in \mathbb{R}$. Finally, we parameterize the emission probability $P_{I_k}(a|\boldsymbol{\beta}_{I_k})$ of symbol a at insert state I_k given parameters $\boldsymbol{\beta}_{I_k}$ in analogy to equation (1).

We define*forward* variables $\mathcal{F}_{S_k}(\ell, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta})$ as the probability of observing the first ℓ symbols of the TS sequence $\boldsymbol{x}$ and visiting node S_k in state interval $s(\ell, \boldsymbol{x}|\boldsymbol{r})$ given parameters $\boldsymbol{\beta}$ and the sequence $\boldsymbol{r}$ of the miRNA, i.e.,

$$\mathcal{F}_{S_k}(\ell, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta}) = P\left(x_1, \ldots, x_\ell, S_k \in s(\ell, \boldsymbol{x}|\boldsymbol{r})|\boldsymbol{r}, \boldsymbol{\beta}\right). \tag{2}$$

A node S_k is visited in state interval $s(\ell, \boldsymbol{x}|\boldsymbol{r})$ if it is contained in a path from D_0 to D_{K+1}, and the symbols x_1 to x_ℓ have been emitted either by predecessors of S_k in the path or by S_k itself, whereas $x_{\ell+1}$ is emitted by a successor of S_k in this path. We use these forward variables for defining the likelihood $P(\boldsymbol{x}|ts, \boldsymbol{r}, \boldsymbol{\beta}_{ts})$ of TS $\boldsymbol{x}$ given the class ts of TS, the sequence of the miRNA $\boldsymbol{r}$, and parameters $\boldsymbol{\beta}_{ts}$, i.e.

$$P(\boldsymbol{x}|ts, \boldsymbol{r}, \boldsymbol{\beta}_{ts}) = \mathcal{F}_{D_{K+1}}(L, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta}_{ts}). \tag{3}$$

Using this definition, the likelihood $P(\boldsymbol{x}|ts, \boldsymbol{r}, \boldsymbol{\beta}_{ts})$ is not necessarily normalized over all possible sequences $\boldsymbol{x} \in \Sigma^L$ of given length L.

Similar to original profile HMMs, we recursively derive the forward variables of match state M_k using its predecessors $S_{k-1} \in \{I_{k-1}, D_{k-1}, M_{k-1}\}$ from the previous column of the plan9 architecture (cf. Fig. 1) as

$$\begin{aligned}\mathcal{F}_{M_k}(\ell, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta}) = {} & P_{M_k}(x_\ell|r_k, \boldsymbol{\beta}_{M_k}) \\ & \sum_{S_{k-1}} \mathcal{F}_{S_{k-1}}(\ell-1, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta})\; P_T(M_k|S_{k-1}, \boldsymbol{\beta}_{T,S_{k-1}}).\end{aligned} \tag{4}$$

In analogy, we derive the forward variables of insert states and delete states.

We initialize the forward variables as follows: We can observe D_0 only before the emission of the first symbol. Hence, we set $\mathcal{F}_{D_0}(\ell, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta})$ to 1 if $\ell = 0$ and to 0 otherwise. We cannot reach M_0 in any admissible path and, thus, $\mathcal{F}_{M_0}(\ell, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta}) = 0$. Finally, we set $\mathcal{F}_{S_k}(0, \boldsymbol{x}|\boldsymbol{r}, \boldsymbol{\beta}) = 0$ for all emitting states S_k.

2.2 Discriminative training

For learning the parameters of the CoProHMM discriminatively, we need an additional background model. Here, we use a homogeneous Markov model of order 1 with parameters $\boldsymbol{\beta}_{bg}$ that do not depend on the miRNA $\boldsymbol{r}$, i.e.,

$$P(\boldsymbol{x}|bg, \boldsymbol{r}, \boldsymbol{\beta}_{bg}) = P_{hMM(1)}(\boldsymbol{x}|\boldsymbol{\beta}_{bg}). \tag{5}$$

We derive the class posterior of class $c \in \{ts, bg\}$ using the likelihoods $P(\boldsymbol{x}|c, \boldsymbol{r}, \boldsymbol{\beta}_c)$ of equations (3) and (5) as

$$P\left(c\,|\,\boldsymbol{x}, \boldsymbol{r}, \boldsymbol{\beta}\right) = \frac{P(c|\boldsymbol{\beta})P(\boldsymbol{x}|c, \boldsymbol{r}, \boldsymbol{\beta}_c)}{\sum_{\tilde{c}} P(\tilde{c}|\boldsymbol{\beta})P(\boldsymbol{x}|\tilde{c}, \boldsymbol{r}, \boldsymbol{\beta}_{\tilde{c}})}, \tag{6}$$

where $P(c|\boldsymbol{\beta})$ denotes the a-priori probability of class c, which we parameterize in analogy to equation (1).

For Bayesian inference, we define a prior on the parameters $\boldsymbol{\beta}$. For the homogeneous Markov model of class bg, we use a transformed product-Dirichlet prior [Mac98] with equivalent sample size (ESS) [HGC95] $\alpha_{bg} \cdot K$. We define another transformed product-Dirichlet prior with ESS α_{ts} for the parameters of the CoProHMM, which is the product of independent transformed Dirichlet priors for each set of transition parameters and each set of emission parameters. We use Dirichlet priors, since these are conjugate to the likelihood of the homogeneous Markov model and to the distribution of transitions and (conditional) emissions. Hence, their hyper-parameters can be intuitively interpreted as pseudo counts. In the following studies, we use $\alpha_{bg} = \alpha_{ts} = 4$.

We learn all parameters $\boldsymbol{\beta}$ on a set of labelled training data $(\boldsymbol{x}_1, \boldsymbol{r}_1, c_1), \ldots, (\boldsymbol{x}_N, \boldsymbol{r}_n, c_N)$. These training data comprise a sufficient number of TSs, i.e. $c_n = ts$, and non-TSs of several miRNAs. Learning the parameters on the TSs of multiple miRNAs conjointly is motivated by the expectation that by this means, CoProHMM may detect general rules of miRNA-TS binding, that could not be detected if we, for instance, learned a standard profile HMM on the TSs of a single miRNA.

We optimize the parameters with respect to the discriminative MSP principle [CdM05, GKK+07], i.e.,

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left[\prod_{n=1}^{N} P\left(c_n \,|\, \boldsymbol{x}_n, \boldsymbol{r}_n, \boldsymbol{\beta}\right) \right] q\left(\boldsymbol{\beta} \,|\, \alpha_{bg}, \alpha_{ts}\right), \tag{7}$$

where $q\left(\boldsymbol{\beta} \,|\, \alpha_{bg}, \alpha_{ts}\right)$ denotes the product-Dirichlet priors on the parameters $\boldsymbol{\beta}$. This optimization must be carried out numerically, which we accomplish by a quasi-Newton second order method.

2.3 Predicting target genes

In the following, we describe how we utilize a CoProHMM for predicting target genes of a miRNA $\boldsymbol{r}$. We assume that the CoProHMM has already been trained on a set of miRNAs – not necessarily including $\boldsymbol{r}$ – and associated TSs and non-TSs. To this end, we extract the UTR $\boldsymbol{y}_n$ of each gene n. Using a sliding window of width $|\boldsymbol{r}|$, we apply the CoProHMM to each sub-sequence of $\boldsymbol{y}_n$ and compute the log-likelihood according to equation (3) given miRNA $\boldsymbol{r}$. For each UTR, we consider the I sub-sequences yielding the largest log-likelihoods $s_{n,i}$, which end at positions $q_{n,i}$. Let $d_n = q_{n,1}$ and $d'_n = |\boldsymbol{y}_n| - q_{n,1}$ be the distance of the sub-sequence with the largest log-likelihood to the 3' and 5' end of the UTR, respectively. Let $(p_{n,1}, \ldots, p_{n,I})$ denote the positions $(q_{n,1}, \ldots, q_{n,I})$ sorted ascendingly. Let $\boldsymbol{z}_n = (s_{n,1}, \ldots, s_{n,I}, d_n, d'_n, p_{n,1}, \ldots, p_{n,I})$ denote the vector of these features representing UTR $\boldsymbol{y}_n$.

By inspecting histograms of the scores $s_{n,i}$, we find that these may be modeled by a mixture of two Gaussian densities, i.e.,

$$P(s_{n,i}|\boldsymbol{\beta}^s_{c,i}) = P(u^s\!=\!1|\boldsymbol{\beta}^{s,m}_{c,i})\, \mathcal{N}(s_i|\mu_{1,i,c}, \kappa_{1,i,c}) + P(u^s\!=\!2|\boldsymbol{\beta}^{s,m}_{c,i})\, \mathcal{N}(s_i|\mu_{2,i,c}, \kappa_{2,i,c}),$$

where $\boldsymbol{\beta}^s_{c,i} = (\boldsymbol{\beta}^{s,m}_{c,i}, \mu_{1,i,c}, \kappa_{1,i,c}, \mu_{2,i,c}, \kappa_{2,i,c})$, $\mu_{k,i,c}$ and $\kappa_{k,i,c}$ denote the mean and the log-precision of Gaussian density k, respectively, and the component probabilities $P(u^s = u|\boldsymbol{\beta}^{s,m}_{c,i})$ are parameterized in analogy to equation (1).
To allow for variability in TS positioning, we model d_n and d'_n each by a mixture of two gamma densities, i.e.,

$$P(d_n|\boldsymbol{\beta}^d_c) = P(u^d = 1|\boldsymbol{\beta}^{d,m}_c)\, \mathcal{G}(d_n|\alpha^d_{1,c}, \beta^d_{1,c}) + P(u^d = 2|\boldsymbol{\beta}^{d,m}_{c,i})\, \mathcal{G}(d_n|\alpha^d_{2,c}, \beta^d_{2,c}),$$

where $\boldsymbol{\beta}^d_c = (\boldsymbol{\beta}^{d,m}_c, \alpha^d_{1,c}, \beta^d_{1,c}, \alpha^d_{2,c}, \beta^d_{2,c})$, and $\alpha^d_{k,c}$ and $\beta^d_{k,c}$ denote the log-shape and log-rate of gamma density k, respectively. We define the density $P(d'_n|\boldsymbol{\beta}^{d'}_c)$ in analogy.

We model the distances $p_{n,i+1} - p_{n,i}$ by another gamma density, i.e.,

$$P(p_{n,i+1} - p_{n,i}|\boldsymbol{\beta}_c^p) = \mathcal{G}(p_{n,i+1} - p_{n,i}|\alpha_c^p, \beta_c^p),$$

where $\boldsymbol{\beta}_c^p = (\alpha_c^p, \beta_c^p)$.

The complete likelihood of $\boldsymbol{z}_n$ representing UTR $\boldsymbol{y}_n$ of gene n employing convenient independence assumptions amounts to

$$P(\boldsymbol{z}_n|c, \boldsymbol{\beta}_c) \propto \prod_{i=1}^{I} P(s_{n,i}|\boldsymbol{\beta}_{c,i}^s)\, P(d_n|\boldsymbol{\beta}_c^d)\, P(d'_n|\boldsymbol{\beta}_c^{d'}) \prod_{i=1}^{I-1} P(p_{n,i+1} - p_{n,i}|\boldsymbol{\beta}_c^p). \quad (8)$$

In the following studies, we use $I = 5$.

In analogy to equation (6), we define the class posterior in terms of likelihoods $P(\boldsymbol{z}_n|c, \boldsymbol{\beta}_c)$ and a-priori class probabilities $P(c|\boldsymbol{\beta})$. As for the training of the TS model, we optimize the parameters with respect to the discriminative MSP principle (cf. equation (7)) using a training data set of target and non-target genes. In this case, we use beta priors on the parameters of the component probabilities, normal-gamma priors on the parameters of the Gaussian densities, and the conjugate prior according to the definition of the exponential family for the gamma densities. Again, we use an ESS of 4 for both classes. We finally predict target genes based on the class posterior.

3 Results & Discussion

In the following, we first investigate if CoProHMMs can learn characteristics of TSs from data. To this end, we use TSs predicted by existing approaches. Second, we evaluate the utility of CoProHMMs for the prediction of target genes of miRNAs on benchmark data.

3.1 Pilot study: Learning CoProHMMs from predictions

We learn CoProHMMs on the predictions of miRanda and TargetScan to investigate if CoProHMMs can learn the rules implemented into these approaches from their predictions. We choose miRanda and TargetScan, because their approaches differ notably. If CoProHMMs can detect such characteristics from predictions, we might expect that they are also capable of learning novel or refined rules of miRNA-TS binding from experimentally verified TS.

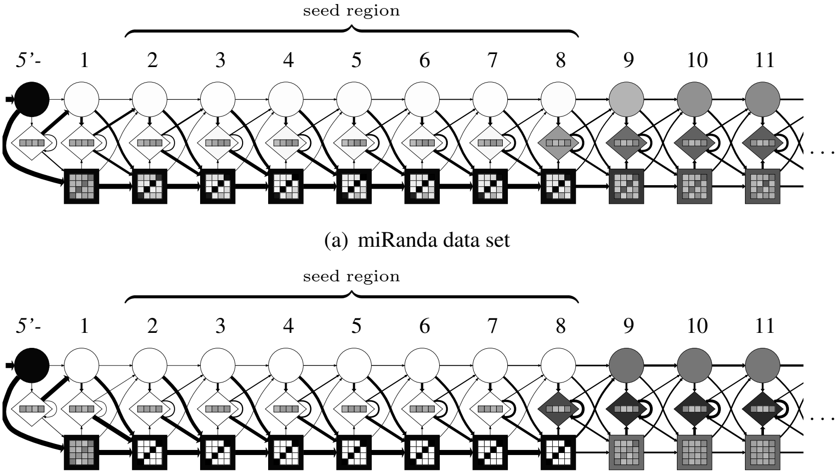
We extract all human TSs and associated miRNAs predicted by TargetScan and miRanda from miRNAMap[1] [HCT$^+$08]. For TargetScan, we use all 244,389 TSs, while we randomly sample 500,000 TSs from the predictions of miRanda. We generate a non-target data set by randomly selecting miRNAs from the mature human miRNAs listed at miRBase[2] [GJSvDE08]. As non-TSs of these miRNAs, we randomly draw 500,000 sub-

[1] ftp://mirnamap.mbc.nctu.edu.tw/miRNAMap2/miRNA_Targets/Homo_sapiens/miRNA_targets_hsa.txt.tar.gz

[2] http://www.mirbase.org

sequences of length $|\boldsymbol{r}| \pm 3$ from 3'-UTRs of human genes according to NCBI Genbank[3] human genome build 37.1.

We present a graphical representation of the CoProHMMs learned on the miRanda data set and the TargetScan data set in Fig. 2. Here, we depict only the most interesting region around the seed, while the complete CoProHMMs for miRanda and TargetScan as well as other approaches are available online[4]. For the states, we use the same shapes as in Fig. 1. The thickness of outgoing edges represents the transition probabilities to the successors of a node. We illustrate the emission probabilities of insert states by a row of grayscale boxes, where the first box corresponds to A, the second box corresponds to C, the third box corresponds to G, and the fourth box corresponds to U. The darker a box, the higher is the corresponding emission probability. In analogy, the conditional emission probabilities of match states are represented by a matrix comprising such rows, where each row corresponds to the conditional probability distribution given one nucleotide of the miRNA. The probabilities of visiting a state are visualized by the darkness of the background of each node. The darker the background of a node the higher the probability of visiting this node.



(a) miRanda data set

(b) TargetScan data set

Figure 2: CoProHMMs learned on the miRanda data set (a) and TargetScan data set (b).

Considering the CoProHMM learned on the miRanda data set, we recover many rules built into miRanda. From the conditional emission probabilities of the match states, we observe a general tendency to complementary base pairings between the TS and the miRNA. This tendency is especially pronounced for the match states in the seed region, but can also be observed for the match states at position 1 and positions 9 to 11. We also detect a slight preference for `G:U` wobble basepairs. These observations are most likely a result of the Smith-Waterman like alignment employed by miRanda. Additionally, miRanda assigns a

[3] http://www.ncbi.nlm.nih.gov

[4] http://www.jstacs.de/index.php/MiRNAs

weight of 2 to the first 11 positions of the alignment, which is reflected by the increased probabilities of visiting match states in the seed region, although this preference already begins to decline at position 8 of the learned CoProHMM.

As a second example, we consider the CoProHMM learned on the TargetScan data set in Fig. 2(b). Notable differences between the CoProHMM for the TargetScan data set and the miRanda data set can be observed for the conditional emission probabilities at the match states. At positions 2 to 8 of Fig. 2(b), we find complementary basepairs almost exclusively, while a slight preference for complementary basepairs is present at the bordering positions 1 and 9. In contrast, the remaining positions exhibit only very slight preferences for specific basepairs. Again, these findings are closely related to the main characteristics built into TargetScan. The perfect complementarity at positions 2 to 8 of the CoProHMM reflects the requirements of TargetScan. We also observe a preference for complementary basepairs at positions 1 and 9, which most likely can be attributed to the fact that initial perfect matches in the seed region may be elongated to either side in TargetScan.

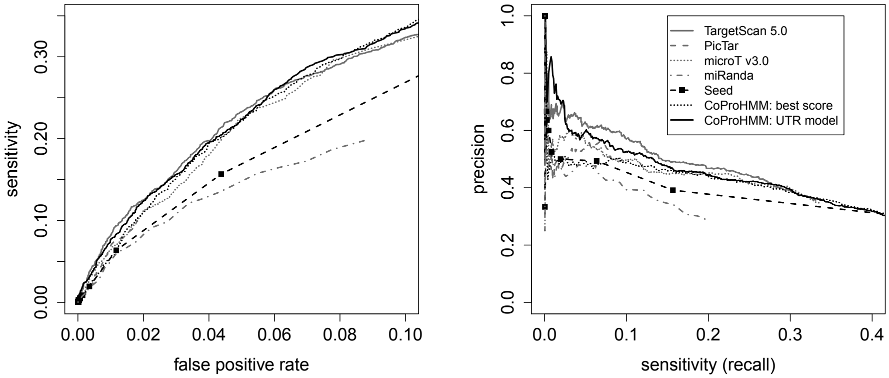
These findings suggest that CoProHMMs are indeed capable of recovering the rules built into miRanda and TargetScan from prediction and, hence, may also be capable of inferring the rules underlying miRNA-TS binding from experimentally verified TSs, once these become available in sufficient quantity.

3.2 Benchmark study: Predicting miRNA target genes

We investigate the utility of CoProHMMs for the prediction of miRNA target genes using the pSILAC data of Selbach *et al.*, which have also been used in recent benchmark studies [SST^{+}08, AMP^{+}09]. To this end, we learn a CoProHMM using a foreground data set that comprises 12 verified TSs and 667 predicted TSs within UTRs of verified target genes extracted from mirecords[5] v. 1 [XZC^{+}09]. As these TSs are too few to reliably learn the models, we also include the TargetScan data set and 405,569 TSs predicted by DIANA-microT. We use predictions of these two approaches, since they yield reasonable precisions in the benchmark studies. We use the same background data set as in the pilot study. We assign a weight of 500 to all verified TSs and a weight of 50 to all predicted TSs in verified target genes to reflect our increased confidence in these data, while we assign a weight of 1 to all other TSs. All TSs of miRNAs contained in the Selbach benchmark data set are excluded when training the CoProHMM to allow for unbiased evaluation.

We extract the UTRs of all genes considered in [SST^{+}08] according to [AMP^{+}09]. For these genes, Selbach *et al.* measured the influence of overexpression or underexpression of a miRNA on the abundance of the corresponding proteins for 5 different miRNAs. For each of these miRNAs, we partition the UTRs into target and non-target UTRs using a threshold of -0.2 on the protein log-fold changes. We assess the performance of the UTR model using the predictions of the CoProHMM in a 5-fold cross validation. In each iteration of the cross validation, we train the parameters of the UTR model on the numeric vectors $\boldsymbol{z}_n$ obtained for 4 of the 5 miRNAs, and we compute the log-likelihood ratios using this trained UTR model for the numeric vectors obtained for the remaining miRNA.

[5] `http://mirecords.biolead.org/download_data.php?v=1`



(a) ROC curve

(b) Precision-recall curve

Figure 3: ROC curve (a) and precision-recall curve (b) of the classifier using the UTR model (solid black line) and the classifier using the best score of the CoProHMM within each UTR sequence (dotted black line) compared to other approaches.

In analogy to [AMP$^+$09], we finally use all log-likelihood ratios to compute sensitivity, precision, and false positive rate for different thresholds.

In Fig. 3, we compare the performance of the classifier using the UTR model (solid black line) to other approaches by means of the precision-recall curve and the ROC curve. As a reference, we also include the performance of a classifier that only uses the best score of the CoProHMM over each UTR sequence, i.e., $s_{n,1}$, (dotted black line). Considering Fig. 3(a), we find that even this classifier using only the best score yields a substantially higher sensitivity than miRanda and Seed for a broad range of false positive rates. Surprisingly, the classifier using the simple UTR model, which does not exploit conservation across species, achieves comparable or slightly improved sensitivities compared to miRanda, Seed, PicTar, and microT, while it performs only slightly worse than TargetScan 5.0 for false positive rates below 0.06.

Turning to the precision-recall curve in Fig. 3(b), we find a similar picture. Notably, the classifier using the UTR model again achieves comparable or even higher precisions than miRanda, Seed, PicTar, and microT. However, it can outperform TargetScan 5.0 only for very low sensitivities and yields lower precisions for sensitivities between 0.03 and 0.28.

The performance of both classifiers using CoProHMMs is astonishing, because, in contrast to most of the other approaches, they do not exploit conservation across different species. Hence, the inclusion of cross-species conservation into CoProHMMs and the proposed UTR model, and the integration of CoProHMMs into other approaches might be a worthwhile direction of future research.

4 Conclusions

miRNAs are involved in the regulation of many cellular processes, and the prediction of miRNA targets is one of the most active fields of bioinformatics. Here, we propose a novel statistical model called conditional profile HMM (CoProHMM) for learning the rules of miRNA-TS interaction from data. We demonstrate that CoProHMMs are capable of reconstructing patterns of miRNA-TS binding built into existing programs from predictions of these approaches.

Conservation is key feature of most miRNA target prediction approaches leading to higher precision at the expense of sensitivity. Interestingly, we find in a benchmark study that a simple UTR model utilizing CoProHMMs yields a competitive precision compared to leading approaches for predicting target genes, although it does not exploit conservation across species.

We anticipate that the number of experimentally verified TSs will rapidly increase in the next years. Only recently, [CZMD09, HLB$^+$10] have independently published novel biological data that shed light on miRNA targeting. Briefly, the two experimental approaches use in-vivo crosslinking, Ago2 immunoprecipitation and cDNA sequencing, and have been able to determine TSs of several miRNAs with high accuracy. Since the power of statistical approaches like CoProHMMs highly depends on the quality of the training data, we might speculate that the performance of CoProHMMs will even increase using these data. Additionally, CoProHMMs might be a suitable approach to extract new and refined rules of miRNA-TS binding from such verified TSs.

We make an implementation of CoProHMMs and the UTR model available to the scientific community with the next release of the open source Java library Jstacs[6].

References

[AMP$^+$09] Panagiotis Alexiou, Manolis Maragkakis, Giorgos L. Papadopoulos, Martin Reczko, and Artemis G. Hatzigeorgiou. Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics*, 25(23):3049–3055, 2009.

[BB01] Pierre Baldi and Søren Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, London, 2nd edition, 2001.

[BSRC05] Julius Brennecke, Alexander Stark, Robert B. Russell, and Stephen M. Cohen. Principles of MicroRNA–Target Recognition. *PLoS Biology*, 3(3), 2005.

[CdM05] Jesús Cerquides and Ramon López de Mántaras. Robust Bayesian Linear Classifier Ensembles. In *Proceedings of the 16th European Conference on Machine Learning*, volume 3720 of *Lecture Notes in Computer Science*, pages 72–83. Springer, 2005.

[CZMD09] Sung Wook Chi, Julie B. Zang, Aldo Mele, and Robert B. Darnell. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–486, 07 2009.

[EJG$^+$03] Anton Enright, Bino John, Ulrike Gaul, Thomas Tuschl, Chris Sander, and Debora Marks. MicroRNA targets in Drosophila. *Genome Biology*, 5(1):R1, 2003.

[6] http://www.jstacs.de

[FFBB09] Robin C. Friedman, Kyle Kai-How Farh, Christopher B. Burge, and David P. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19(1):92–105, 2009.

[GJSvDE08] Sam Griffiths-Jones, Harpreet Kaur Saini, Stijn van Dongen, and Anton J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Research*, 36(suppl_1):D154–158, 2008.

[GKK$^+$07] Jan Grau, Jens Keilwagen, Alexander Kel, Ivo Grosse, and Stefan Posch. Supervised posteriors for DNA-motif classification. In Claudia Falter, Alexander Schliep, Joachim Selbig, Martin Vingron, and Dirk Walther, editors, *German Conference on Bioinformatics*, volume 115 of *Lecture Notes in Informatics (LNI) - Proceedings*, Bonn, 2007. Gesellschaft für Informatik.

[HCT$^+$08] Sheng-Da Hsu, Chia-Huei Chu, Ann-Ping Tsou, Shu-Jen Chen, Hua-Chien Chen, Paul Wei-Che Hsu, Yung-Hao Wong, Yi-Hsuan Chen, Gian-Hung Chen, and Hsien-Da Huang. miRNAMap 2.0: genomic maps of microRNAs in metazoan genomes. *Nucleic Acids Research*, 36(suppl_1):D165–169, 2008.

[HGC95] David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. In *Machine Learning*, pages 197–243, 1995.

[HLB$^+$10] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. 141(1):129–141, 04 2010.

[KBM$^+$94] Anders Krogh, Michael Brown, I. Saira Mian, Kimmen Sjölander, and David Haussler. Hidden Markov Models in Computational Biology : Applications to Protein Modeling. *Journal of Molecular Biology*, 235(5):1501 – 1531, 1994.

[KGP$^+$05] Azra Krek, Dominic Grun, Matthew N Poy, Rachel Wolf, Lauren Rosenberg, Eric J Epstein, Philip MacMenamin, Isabelle da Piedade, Kristin C Gunsalus, Markus Stoffel, and Nikolaus Rajewsky. Combinatorial microRNA target predictions. *Nature Genetics*, 37(5):495–500, 05 2005.

[LBB05] Benjamin P. Lewis, Christopher B. Burge, and David P. Bartel. Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell*, 120(1):15 – 20, 2005.

[LSJR$^+$03] Benjamin P. Lewis, I-hung Shih, Matthew W. Jones-Rhoades, David P. Bartel, and Christopher B. Burge. Prediction of Mammalian MicroRNA Targets. *Cell*, 115(7):787 – 798, 2003.

[Mac98] David J. C. MacKay. Choice of Basis for Laplace Approximation. *Machine Learning*, 33(1):77–86, 1998.

[MRS$^+$09] M. Maragkakis, M. Reczko, V. A. Simossis, P. Alexiou, G. L. Papadopoulos, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas, and A. G. Hatzigeorgiou. DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Research*, 37(suppl_2):W273–276, 2009.

[SST$^+$08] Matthias Selbach, Bjorn Schwanhausser, Nadine Thierfelder, Zhuo Fang, Raya Khanin, and Nikolaus Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, 09 2008.

[XZC$^+$09] Feifei Xiao, Zhixiang Zuo, Guoshuai Cai, Shuli Kang, Xiaolian Gao, and Tongbin Li. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Research*, 37(suppl_1):D105–110, 2009.

Quantitative Comparison of Genomic-Wide Protein Domain Distributions

Arli A. Parikesit[1*], Peter F. Stadler[1−5], Sonja J. Prohaska[1],
[1]Bioinformatics Group, Dept. Computer Science,
and Interdisciplinary Center for Bioinformatics, University of Leipzig,
Härtelstraße 16-18, D-04107 Leipzig, Germany
[2]Max Planck Institute for Mathematics in the Sciences,
Inselstraße 22, D-04107 Leipzig, Germany
[3]Fraunhofer Institute for Cell Therapy und Immunology,
Perlickstr.1, D-04103 Leipzig, Germany
[4]Institute for Theoretical Chemistry, University of Vienna,
Währingerstrasse 17, A-1090 Vienna, Austria
[5]The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico
[*]Corresponding Author

Abstract: Investigations into the origins and evolution of regulatory mechanisms require quantitative estimates of the abundance and co-occurrence of functional protein domains among distantly related genomes. Currently available databases, such as the SUPERFAMILY, are not designed for quantitative comparisons since they are built upon transcript and protein annotations provided by the various different genome annotation projects. Large biases are introduced by the differences in genome annotation protocols, which strongly depend on the availability of transcript information and well-annotated closely related organisms.

Here we show that the combination of *de novo* gene predictors and subsequent HMM-based annotation of SCOP domains in the predicted peptides leads to consistent estimates with acceptable accuracy that in particular can be utilized for systematic studies of the evolution of protein domain occurrences and co-occurrences. As an application, we considered four major classes of DNA binding domains: zink-finger, leucine-zipper, winged-helix, and HMG-box. We found that different types of DNA binding domains systematically avoid each other throughout the evolution of Eukarya. In contrast, DNA binding domains belonging to the same superfamily readily co-occur in the same protein.

1 Introduction

The expression of genomically encoded information is subject to tight regulation and control in all organisms that have been studied in detail. These regulatory rules are implemented in a highly complex network of several biochemically distinct mechanism that act at multiple levels of the gene expression cascade. They include specific chromatin states, the action of transcription factors, regulated mRNA export, alternative splicing, translational control, post-transcriptional and post-translational modifications, and controlled degradation of both RNA and polypeptides. Surprisingly, it appears that different phylogenetic clades emphasize certain types of mechanisms while reducing or even abolishing others. Regulation in eubacteria, for example, appears to be dominated by transcription factors networks, trypanosomes use the post-transcriptional processing of large polycistronic transcripts, ciliates utilize extensive amplification of DNA in creating their macro-nuclei, and crown group eukaryotes have evolved an elaborates system of histone modifications. An understanding of the diversity of life thus requires the investigation of the origin(s) and evolution of these different regulatory mechanisms and their interplay.

The most direct approach towards this goal is the comprehensive reconstruction of the evolutionary histories of the many protein families that play a role in the various modes of evolution. In practice, however, this is an exceedingly difficult and tedious task, since homologies even between highly conserved proteins become hard to establish in comparisons across kingdoms or even across the three domains of life. This is not only for technical reasons: Proteins are composed of recognizable protein domains that implement well-defined functions such as catalytic activities, specific binding, and anchoring in membranes. Over large time-scales, these components have been combined in a combinatorial fashion to produce new functionalities, so that individual proteins often have multiple ancestors that contributed different domains [MBE$^+$08, KAK00]. A more modest approach thus aims at tracing the *distribution* of protein domains comparatively. In a recent study of chromatin evolution, we demonstrated that this is indeed feasible [PSK10]. More detailed insights can be gained from considering domain combinations. For instance, Itoh *et al.* [INK$^+$07] showed that there are many animal-specific or even vertebrate-specific domain-combinations. Network analysis of domain co-occurrences, furthermore, demonstrates a growing core of combinations in multicellular organisms [WA05].

Typically, studies of this type are based on existing annotation. For instance, the protein annotation compiled in KEGG, ENSEMBL and Pfam [FMSB$^+$06] domains were used in [INK$^+$07], ref. [PSK10] was based on the SUPERFAMILY database [WPZ$^+$09], whose HMM models in turn are based on the SCOP (Structural Classification of Proteins) domain definitions [AHC$^+$08].

We recently attempted to investigate the origins of the proteins associated with the microRNA pathway using a rather straightforward approach: For each of the most prominent proteins associated with the microRNA pathway (Drosha, Dicer, DGCR8, TRBP, and TRBP), we searched the SUPERFAMILY database for putative homologs. To this end, we collected the functional domains of these proteins from the literature and then identified the SUPERFAMILY peptide entries in which these known domains co-occurred. Somewhat surprisingly, this approach did not recover the phylogenetic distributions reported in detailed, homology-based studies [CR07, MDB08]. Apart from domains that were missing completely (such as PIWI), we observed that many domains are annotated only in a small subset of the species that are expected to contain them. We concluded from this pilot study that existing peptide annotations are a problematic data source for quantitative cross-species comparisons. The issues are twofold:

1. A comprehensive analysis of the evolution of gene *function* requires a reasonably complete collection and annotation of protein domains. Of course, the current

knowledge is not complete, and there are still novel functional domains yet to be discovered. Interestingly in that regard, co-occurrence data can help to detect undescribed and divergent protein domains [TGMB09]. Furthermore, most protein domains in well-studied model organisms are evolutionarily very old, suggesting that the innovation of protein domains is a relatively infrequent phenomenon [BBHS10]. For example, a recent study showed that the majority of "plant-specific" DNA binding domains originated much earlier then the comparably recent expansion into the diverse gene families present in higher plants [SeoopstfDbd08].

2. The annotation of protein domains is performed on protein sequences retrieved from sequence databases. For each species, these "protein models" are constructed by combining the genomic DNA sequence, EST and cDNA data, and computational predictions. Large differences in EST and/or cDNA coverage as well as in the computational procedures imply that domain annotations can be very different even for phylogenetically closely related species. For example, the current version (1.73) of SUPERFAMILY annotates 64225 domains in human, but only 45312 in chimpanzee, 21208 in gorilla and 14748 in the alpaca, although one would expect a very similar gene complement throughout the eutherian mammals.

In this contribution, we focus on the second issue and investigate strategies to construct inventories of protein domains that avoid the biases arising from gene annotation. While it would certainly be desirable to obtain a complete set of protein domains encoded in any given genome, this is not feasible at present. Our goal here is thus more moderate: we are content with estimates that are consistent between different genomes and thus allow quantitative comparisons. To this end, we re-annotate protein domains using the following three different collections of (putative) polypeptides for each genome: (1) computational translations of annotated transcripts available in sequence databases, (2) conceptual translations of the entire genomic DNA in all 6 reading frames, and (3) protein predictions generated by a *de novo* gene predictor.

2 Materials and Methods

As test system we use the genomes of three apes (human GRCh37.57, chimp CHIMP2.1.57, and gorilla gorGor3.57). The genomes were downloaded from the ENSMBL website (`www.ensembl.org`), version 57. Transcript files were downloaded from the cDNA section of the corresponding genome builds. The three ape species are so similar that we can expect a virtually identical complement of protein domains. Even in very rapidly evolving gene families, such as the KRAB-ZNF family of transcriptional repressors [NHZS10], the copy numbers differences in between primates are restricted to a few percent. The most extreme case are olfactory receptors [Nii09], where the number of functional copies differs by up to 25% between human and chimp due to massive gene loss [GN08]. This difference, however, will not be clearly detectable at domain level, since many of the very recent pseudogenes are expected to yield inconspicuous hits to the HMM domains models. In contrast to expected similarity of the great apes, their transcriptome and proteome

Table 1: Summary statistics of source data. The number of domains refers to query set of 100 randomly selected SCOP entries. n.d.: not determined.

Species	Human	Chimpanzee	Gorilla	Yeast
Data set	RCh37.57	CHIMP2.1.57	gorGor3.57	SGD1.01.57
	number of peptides investigated			
transcripts	76592	34142	27325	5885
genscan	118894	96615	113532	4197
	number of detected domains			
transcripts	5551	3769	3386	621
genscan	3392	2796	3323	614
genomic translation	23	n.d.	n.d.	409

annotations differ by nearly a factor of three, Tab. 1.

Gene predictions were performed using genscan [BK97, BK98]. To this end, the chromosomes were split into fragments between 500kb and 600kb since genscan does not accept larger input files. The sequences of the predicted genes were extracted directly from the genscan output. The chromosome fragments were constructed with substantial overlaps to avoid artifacts arising from incomplete gene predictions at fragment boundaries, leading to redundant predictions within the overlapping regions. These were removed before further analysis.

We also tested GeneMark [LTHCM05] as an alternative gene predictor and obtained comparable results. We decided to focus on genscan because: (1) it has been reported to perform well across distantly related species (teleost fishes, nematodes, amphioxus, and fungi) without retraining its internal model [Kor04], (2) because it is much faster than the alternatives, and (3) because it is the mostly widely used gene predictor [MMNH04].

Protein domains are represented as Hidden Markov Models (HMMs) [Edd96, DEKM98, Edd98]. In order to save computations resources we randomly selected 100 domains from the SUPERFAMILY database [WPZ$^+$09], version 1.73 (10.01.2010) for the statistical analysis. We used HMMER 3.0rc1 to map the HMMs to the protein sequences with the the same E-value cut-off as the SUPERFAMILY: $E \leq 10^{-4}$. In case of overlapping HMM hits, we retain only the best-scoring match.

3 Results

Scatter-plots of the number of domain occurrences measured on the set of annotated transcript and on the *de novo* gene predictions shows a significant correlation, Fig 1. In contrast, an attempt to estimates the domain numbers by running the HMMs on translated genomic DNA failed miserably: only a small fractions of the known domains can be recovered. This is not surprising. Although there is a statistically significant correlation between protein domain boundaries and exon boundaries [LWWG05], about two thirds of the annotated protein domains domains are interrupted by at least one introns, and on average a domain contains 3 or 4 introns [BPMS09]. Thus most domains are undetectable in conceptual translations of the genomic DNA.

In the human data, the majority of domains is observed more frequently in annotated transcripts than in `genscan` predictions (Fig. 1a). This effect is less pronounced in chimpanzee (Fig. 1b). In yeast, on the other hand, the correspondence between transcript-based domain annotation and the `genscan`-based results is excellent. We can understand these differences because of dramatic differences in the quality and coverage of the transcript annotation. In the human genome, for example, a large number of annotated isoforms and alternative transcripts are annotated as a result of extensive cataloging efforts. Thus, multiple transcripts may incorporate the same genomic domain. A comparable density of data is not available for any other species, which results in an inevitable underestimation of annotated transcripts (as in the two ape genomes). Transcript annotation and `genscan` predictions agree extremely well in yeast, however. The data in Table 2 show a good overall correlation between the domain counts as reported by the SUPERFAMILY database and those computed from the `genscan` predictions, although counts can deviate largely in some species. For instance, in *Trypanosoma brucei* we detect 146 zinkfingers using gene predictions compared to only 7 annotated in SUPERFAMILY.

To investigate the suitability of gene predictions for the assessment of domain co-occurrences, we selected two very abundant classes of DNA binding domains: zink-finger domains (ZNF) and winged-helix domains. If the two domain types were distributed randomly, we would expect about 17.8 co-occurrences, estimated from the data in the SUPERFAMILY (30712 transcripts, of which 1324 contain a ZNF domain and 414 have a winged-helix domain). Surprisingly, not a single co-occurrence between these two domains is observed in the SUPERFAMILY data in any species, even though both domains are conserved throughout the Eukarya, Table 2.

In the `genscan`-based analysis, we detected co-occurrences of ZNF and winged-helix domains only in the clades Kinetoplastida (*Leishmania* and *Trypansoma*) and in *Phytophtora*. Upon closer inspection, these can can be identified as artifacts. In Kinetoplastida, the problem is caused by the unusual structure of the transcriptome of Kinetoplastida, which consists of long, polycistronic mRNAs that are processed by transsplicing [MCVdRFM$^+$10]. Our hits fall into a highly conserved polycistron of more than 10kb length, for which `genscan` predicts a "polyprotein". Interestingly, no spurious co-occurrences are found in the nematode *C. elegans*, whose polycistronic messages contain much fewer proteins. The second artifact are two hits in *Phytophtora*: one is again a putative artifact `genscan`, which here predicts a chimera of RNA polymerase III subunit C34 and a hypothetical zinkfinger protein. The second hit covers a protein annotated as homolog of the EAP30 subunit of the ELL complex containg two winged-helix domains. In the latter case, the zink-finger domain is most likely located in an additional downstream exon that is conserved between *Phytophtora sojae* and *Phytophtora ramorum*.

The exclusive usage of one of the two types of DNA binding domains is statistically highly significant. In human, for instance, we expect 11.7 co-occurrences (5090 ZNF and 274 winged-helix domains in 118894 `genscan` predictions) while none is observed

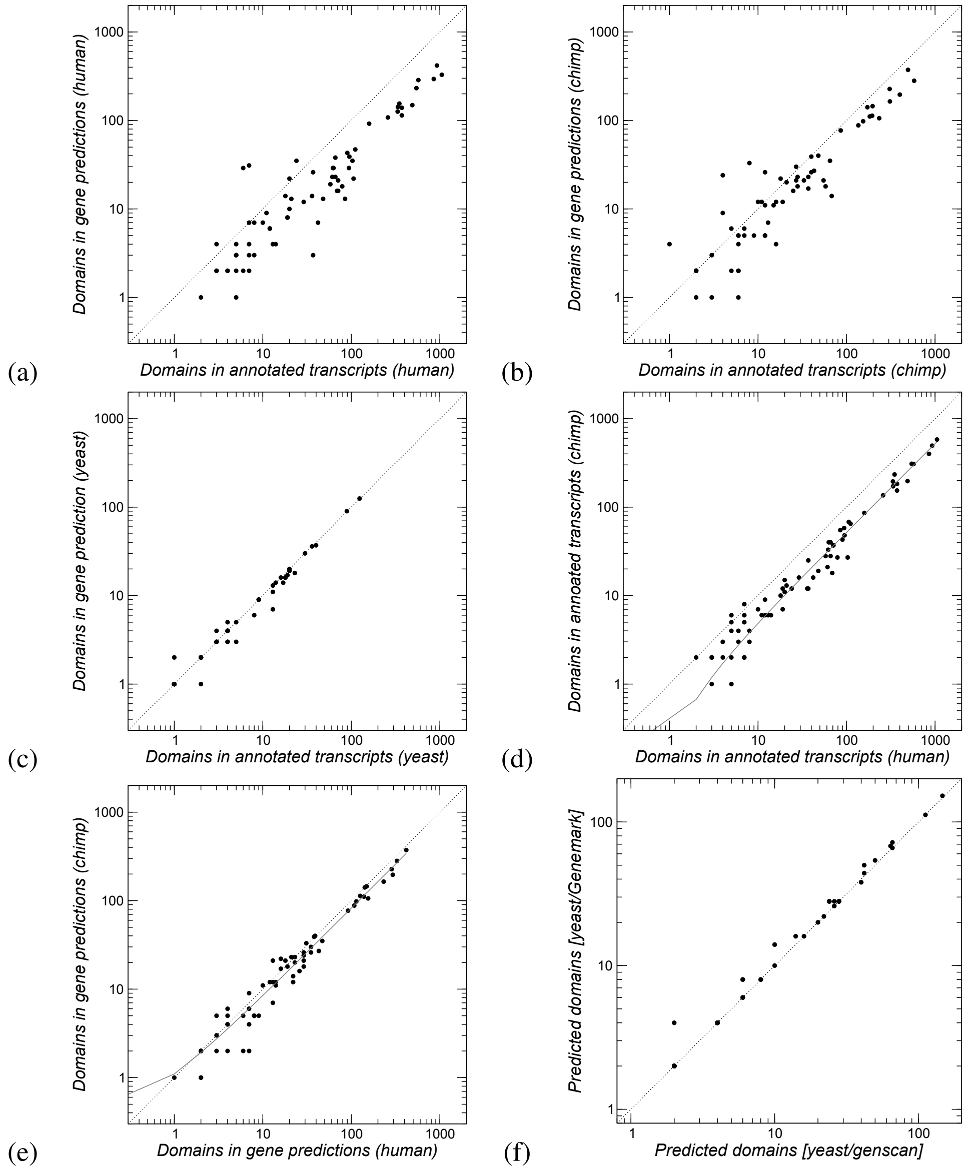


Figure 1: Correlation of the number of protein domains. Top row: Annotated transcripts compared to *de novo* predicted "genes" for (a) human, (b) chimp, and (c) yeast. Below: While domain prediction based on existing annotation yield systematic differences between human and chimp (d), congruent abundances are obtained from `genscan` predictions (e). Linear regression is shown as red line in panels (e) and (f). Different gene predictors (`genscan` and `GeneMark`) yield comparable results (f), shown here for yeast.

($p < 10^{-5}$). This indicates a selective pressure against their co-occurrences. We therefore also investigated two additional families of DNA binding domains, namely the leucine

Table 2: Domain occurrences and co-occurrences of zink-finger and winged-helix domains. The table shows the number of domains (Dom.), the number of “genes”, i.e., `genscan` predictions that contain the domain (Genes), and for comparison the number of genes that contain the domain in SUPERFAMILY (SF). For species marked with *, multiple entries from different strains or variants in the SUPERFAMILY database exist, and SF values tend to over-count in these cases.

	ZNF [57667]			winged helix [46785]			co-occurrence		
Species	Dom.	Genes	SF	Dom.	Genes	SF	Dom.	Genes	SF
Giardia lamblia	7	6	4	16	13	11	0	0	0
Trichomonas vaginalis	23	14	9	100	98	89	0	0	0
Trypanosoma brucei	156	148	6	34	32	24	1	1	0
Leishmania major *	29	14	6	50	27	23	2	1	0
Naegleria gruberi	20	7	6	67	45	47	0	0	0
Plasmodium falciparum *	5	5	12	3	3	38	0	0	0
Tetrahymena	1	1	13	3	3	39	0	0	0
Thalassiosira pseudonana	15	11	8	145	138	130	0	0	0
Phytophthora ramorum	81	46	34	80	75	62	6	2	0
Clamydomonas	18	13	7	48	44	37	0	0	0
Arabidopsis thaliana *	151	115	74	186	168	241	0	0	0
Oryza sativa *	284	224	307	151	146	443	0	0	0
Dictyostelium	21	10	12	42	37	48	0	0	0
Aspergilus niger	64	51	34	68	65	47	0	0	0
Schizosaccaromyces pombe *	34	24	38	43	41	80	0	0	0
Caenoharbditis elegans *	58	27	144	15	14	165	0	0	0
Drosophila melanogaster *	853	301	322	126	122	152	0	0	0
Homo sapiens *	5090	1048	1324	274	256	414	0	0	0

zippers (SUPERFAMILY ID 57979) and the “high mobility group” (HMG) domains (SUPERFAMILY ID 47095). We again observe only very few candidate co-occurrences with other DNA binding domains in the species listed in Table 2 (our co-occurences between leucine-zipper and winged-helix and one between HMG and winged-helix). Inspection of these five cases revealed that four of them are clear artifacts of `genscan`, which predicts a fusion protein. The last candidate, human LARP1B, is predicted by `genscan` to have an additional internal exon containing a leucine-zipper domain. More likely, however, `genscan` stumbled across a retro-pseudogene deriving from FOSL1 located in an intron of LARP1B. Conversely, SUPERFAMILY, reports the co-occurrence of leucine-zipper and zink-finger in some isoforms of the paralogous human ATF2 and ATF7 genes, which are not found in our `genscan`-based approach.

We therefore conclude that the major types of DNA binding domains, and possibly other evolutionarily unrelated domains of similar function, strongly avoid each other in Eukarya. In contrast, domains with complementary functions readily co-occur with each other. A good example are zink-fingers and the “Küppel associated box” (KRAB) domain. The KRAB domain is a small (75 AA) protein domain [SUPERFAMILY ID 57667] that functions as a transcriptional repressor and is predicted to act via protein-protein interactions. It appears in a highly prolific family of evolutionarily very young transcription factors. Among the species listed in Table 2, it appears only in human. We detected 446 domains in 421 “genes”, in agreement with the literature [NHZS10]. In contrast to the winged-helix domain, however, it readily combines with zink-finger domains: 351 `genscan` predictions (i.e., a third) of the 1048 ZNF proteins and 5/6 of the KRAB domain proteins belong to the KRAB-ZNF family, again in good agreement with the literature.

4 Discussion

Although a plethora of annotation data are available in publicly accessible databases for most of the published genomes, quantitative comparisons remain difficult due to dramatic differences in annotation methodology and data coverage. Consequently, comparative studies typically resort to testing for relative enrichment rather than considering absolute numbers of domains. In studies focusing on the evolution of regulatory mechanisms and regulatory complexity, however, absolute gene counts play an important role. For example, the fraction of transcription factors increases approximately quadratically with the total number of genes in eubacteria [vN03]. A result like this requires an estimate of the total number of genes with reasonable reliability and accuracy. Similarly, investigations into lineage-specific variations of regulatory schemes require plausible statistics of protein domains and their combinations [PSK10]. For prokaryotes, this task is more or less solved by the common practice of annotating all open reading frames. The HMM models of protein domains are easily searched against the (translation of) these ORFs and included e.g. in the SUPERFAMILY database. False positives in the ORF annotation pose little problem since they are very unlikely to contain recognizable protein domains.

In Eukarya, however, the situation is different. Direct annotation of ORFs on the genome level does not work for most organisms since introns interrupt many domains. On the other hand, databases of experimentally determined transcripts are often subject to massive sampling biases. Here, we show that protein domains can be annotated with acceptable accuracy using *de novo* gene predictors such as `genscan`. This strategy also avoids methodological biases such as the enrichment of 3'-exons in poly-A ESTs.

We emphasize that it is impossible in practice to devise a fair benchmark for domain co-occurrence counts since the ground truth depends on the complete knowledge of all transcripts, even if one settles for the definition that two particular protein domains co-occur if they appear together in at least one protein-coding transcript. Therefore, we have to resort to comparing counts between closely related species for which we can plausibly expect to obtain similar numbers.

In easy cases, such as yeast, where the transcript structure is simple and data coverage is excellent, gene prediction and transcript annotation yield nearly identical results. For large mammalian genomes, on the other hand, estimates of domain numbers depend strongly on transcript coverage, while gene predictions yield numbers that are consistent among closely related species. Our investigation suggests that the biases and artifacts in the `genscan` are small compared to the numerous problems of annotation-based approaches. In particular, we observe very a small number of false positive co-occurrences arising from the incorporation of additional introns and the erroneous prediction of fusion proteins.

As an application of genome-wide domain counts, we investigated the co-occurrences of four major types of DNA binding domains (zink-fingers, leucine-zipper, HMG-box domains, and winged-helix domains). We found a strong and statistically highly significant anti-correlation of the four different domains. In constrast, evolutionarily related DNA binding domains readily co-occur in DNA binding proteins. It will be interesting to investigate whether a similar avoidance can be observed among other evolutionarily unrelated protein domains that share a common molecular function.

Acknowledgment. AAP is supported by a DAAD fellowship. We thank Christian Arnold for helpful comments and for proofreading the manuscript.

References

[AHC$^+$08] Antonina Andreeva, Dave Howorth, John-Marc Chandonia, Steven E. Brenner, Tim J. P. Hubbard, Cyrus Chothia und Alexey G. Murzin. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, 36:D419–D425, 2008.

[BBHS10] E Bornberg-Bauer, A K Huylmans und T. Sikosek. How do new proteins arise? *Curr Opin Struct Biol.*, 20:390â396, 2010.

[BK97] C. Burge und S. Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268:78–94, 1997.

[BK98] C. B. Burge und S. Karlin. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.*, 8:346–354, 1998.

[BPMS09] A Bhasi, P Philip, V Manikandan und P. Senapathy. `ExDom`: an integrated database for comparative analysis of the exon-intron structures of protein domains in eukaryotes. *Nucleic Acids Res.*, 37:D703–D711, 2009.

[CR07] Kevin Chen und Nikolaus Rajewsky. The Evolution of gene regulation by Transcription Factors and microRNAs. *Nature Genetics*, 8:93–103, 2007.

[DEKM98] R. Durbin, Sean Eddy, Anders Krogh und G. Mitchison. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[Edd96] S R Eddy. Hidden Markov models. *Curr Opin Struct Biol*, 6:361–365, 1996.

[Edd98] S R Eddy. Profile hidden Markov models. *Bioinformatics*, 14:755–763, 1998.

[FMSB$^+$06] R D Finn, J Mistry, B Schuster-Böckler, S Griffiths-Jones, V Hollich, T Lassmann, S Moxon, M Marshall, A Khanna, R Durbin, S R Eddy, E L Sonnhammer und A. Bateman. `Pfam`: clans, web tools and services. *Nucleic Acids Res.*, 34:D247–D251, 2006.

[GN08] Y Go und Y. Niimura. Similar numbers but different repertoires of olfactory receptor genes in humans and chimpanzees. *Mol Biol Evol.*, 25:1897–1907, 2008.

[INK$^+$07] Masumi Itoh, Jose C Nacher, Kei-ichi Kuma, Susumu Goto und Minoru Kanehisa. Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. *Genome Biol.*, 8:R121, 2007.

[KAK00] E Koonin, L Aravind und A Kondrashov. The impact of comparative genomics on our understanding of evolution. *Cell*, 101:573–576, 2000.

[Kor04] I. Korf. Gene finding in novel genomes. *BMC Bioinformatics*, 5:59, 2004.

[LTHCM05] A. Lomsadze, V. Ter-Hovhannisyan, Y. Chernoff und Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.*, 33:6494–6506, 2005.

[LWWG05] Mingyi Liu, Heiko Walch, Shaoping Wu und Andrei Grigoriev. Significant expansion of exon-bordering protein domains during animal proteome evolution. *Nucleic Acids Res.*, 33:95–105, 2005.

[MBE$^+$08] Andrew D. Moore, Åsa K. Björklund, Diana Ekman, Erich Bornberg-Bauer und Arne Elofsson. Arrangements in the modular evolution of proteins. *Trends Biochem. Sci.*, 33:444–451, 2008.

[MCVdRFM$^+$10] Santiago Martínez-Calvillo, Juan C. Vizuet-de Rueda, Luis E. Florencio-Martínez, Rebeca G. Manning-Cela und Elisa E. Figueroa-Angulo. Gene Expression in Trypanosomatid Parasites. *J. Biomed. Biotech.*, 2010. doi:10.1155/2010/525241.

[MDB08] Dennis Murphy, Barry Dancis und James R. Brown. The evolution of core proteins involved in microRNA biogenesis. *BMC Evolutionary Biology*, 8:92, 2008.

[MMNH04] Webb Miller, Kateryna D Makova, A Nekrutenko und Ross C Hardison. Comparative Genomics. *Ann. Rev. Genomics Hum. Genet.*, 5:15–56, 2004.

[NHZS10] K Nowick, A T Hamilton, H Zhang und L Stubbs. Rapid sequence and expression divergence suggests selection for novel function in primate-specific KRAB-ZNF genes. *Mol Biol Evol.*, 2010. doi:10.1093/molbev/msq157.

[Nii09] Y. Niimura. Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents. *Hum. Genomics*, 4:107–118, 2009.

[PSK10] Sonja J Prohaska, Peter F. Stadler und David C. Krakauer. Innovation in Gene Regulation: The Case of Chromatin Computation. *J. Theor. Biol.*, 265:27–44, 2010.

[SeoopstfDbd08] Structures und evolutionary origins of plant-specific transcription factor DNA-binding domains. Yamasaki, Kazuhiko and Kigawa, Takanori and Inoue, Makoto and Watanabe, Satoru and Tateno, Masaru and Seki, Motoaki and Shinozaki, Kazuo and Yokoyama, Shigeyuki. *Plant Physiol. Biochem.*, 46:394–401, 2008.

[TGMB09] Nicolas Terrapon, Olivier Gascuel Gascuel, Éric Maréchal und Laurent Bréhélin. Detection of new protein domains using co-occurrence: application to *Plasmodium falciparum*. *Bioinformatics*, 25:3077–3083, 2009.

[vN03] Erik van Nimwegen. Scaling laws in the functional content of genomes. *Trends Genetics*, 19:479–484, 2003.

[WA05] Stefan Wuchty und Eivind Almaas. Evolutionary cores of domain co-occurence networks. *BMC Evol. Biol.*, 5:24, 2005.

[WPZ$^+$09] D Wilson, R Pethica, Y Zhou, C Talbot, C Vogel, M Madera, C Chothia und J. Gough. SUPERFAMILY — Comparative Genomics, Datamining and Sophisticated Visualisation. *Nucleic Acids Res.*, 37:D380–D386, 2009.

METAtarget – extracting key enzymes of metabolic regulation from high-throughput metabolomics data using KEGG REACTION information

Jan Budczies[1*], Carsten Denkert[1], Berit M Müller[1], Scarlet F Brockmöller[1],
Manfred Dietel[1], Jules L Griffin[2], Matej Oresic[3], Oliver Fiehn[4]

[1]Institute of Pathology, Charité – Universitätsmedizin Berlin, 10117 Berlin, Germany
[2]Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom
[3]VTT Technical Research Centre of Finnland, Espoo, Finnland
[4]Genome Center, University of California Davis, Davis, CA, USA
[*]Email: jan.budczies@charite.de

Abstract: METAtarget is a new method for reverse engineering of metabolic networks and the detection of targets enzymes from high-throughput metabolomics data. Using KEGG REACTION, reactant partners are identified and the ratio of product to substrate metabolite concentrations is employed as surrogate for the reaction activity. A test statistics is introduced to assess changes in the activity of reactions between different disease states. In an application of METAtarget to breast cancer, we investigate the dependence of tumor metabolism on hormone receptor status. To this end, we analyze metabolomics data that were generated within the METAcancer project and compare the identified reactions with data on enzyme expression that are obtained from publicly available breast cancer gene expression series. As result, deregulation of key enzymes and reactions of glycolysis, glutaminolysis and other metabolic pathways are detected.

1 Introduction

In recent years, techniques for metabolic profiling based on mass spectrometry (MS) and nuclear magnetic resonance spectrometry (NMR) advanced and now allow the simultaneous monitoring of hundreds of metabolites [Fi01, GS04]. Metabolomics emerged as an additional high-throughput technology complementary to other -omics approaches like genomics, transcriptomics and proteomics. In cancer research, liquid and gas chromatography-based MS have been successfully applied to the analysis of body fluids and tissues [De08, De09, Sr09].

Uncovering of the biochemical pathways that constitute the human metabolism represents one of the major achievements of biochemical research over the past 100 years [GHW05]. This knowledge has been an invaluable information source to improve human health by providing new insights into nutrition, disease mechanisms and the effect of drugs. In the postgenomic era metabolic pathway knowledge has been integrated with information from sequencing of genomes and is publicly available from databases like KEGG [Ka10], Reactome [Ma09] and BioCarta (www.biocarta.com).

Consequently, suitable tools for the integration of metabolomics data together with pathway knowledge are urgently needed and will facilitate and accelerate the interpretation of experimental data. Currently, several tools are available for the visualization of metabolomics data in context of biochemical networks. Metscape [Ga10] and MetaNetter [Jo08] are plugins for cytoscape, a powerful and widely-used software environment for models of biomolecular interaction networks [Sh03]. Web-based metabolic network explorers include the KEGG atlas [Ok08] and iPath [Le08]. Going beyond a network visualization, PROFILE clustering orders metabolites according to their distance in KEGG pathways and visualizes metabolic changes in context of the functional clustering [De08]. TICL is a tool for network generation from metabolite lists that includes a significance assessment for the relevance of the generated networks [An09].

Here we present METAtarget, a method for quantitative analysis of metabolomics data in context of biochemical pathways. METAtarget employs the ratio of product to substrate concentrations as surrogate for the activity of metabolic reactions. A suitable test statistic is defined in order to measure changes of reaction activities between two disease states. METAtarget delivers a list significantly changed reactions and the associated enzymes that are possible targets for a therapeutic intervention.

As an application of METAtarget, we analyze GC-MS data that were generated in the framework of METAcancer, a European collaboration on the metabolism of breast cancer. Comparing metabolic profiles of estrogen receptor positive (ER+) and receptor negative (ER-) breast cancer, METAtarget delivers a list of metabolic reactions that are regulated depending on hormone receptor status. Using an independent gene expression data set on breast cancer, we evaluate the hypothesis that the detected changes in metabolism are associated with transcriptional regulation of enzymes.

2 Material and Methods

2.1 Assessing the activity of metabolic reactions

The simplest design of a metabolomics experiment deals with the comparison of diseased and healthy tissue or of different tissues types in disease states *a* and *b*. Let us denote the concentration of a metabolite *Z* in two tissue types by Z_a and Z_b. As it is commonpraxis for the analysis of –omics data, we assume that the variables *Z* are transformed to the log-scale. For the log-scale concentrations of the product *X* and the substrate *Y* of a metabolic reaction we define the statistics

$$t = \frac{\mathrm{E}(X_a)-\mathrm{E}(X_b)-\mathrm{E}(Y_a)+\mathrm{E}(Y_b)}{\sqrt{\mathrm{VAR}(X_a-Y_a)/\,N_a+\mathrm{VAR}(X_b-Y_b)/\,N_b}},$$

wherein N_a and N_b are the numbers of tissues in disease states a and b. The numerator of t can be interpreted in two ways: (i) as a different effect of the disease state on the product compared to the substrate and (ii) as a measure how the product-substrate ratio changes when comparing the two disease states a and b. The statistics defined above can be read as Welch's t-statistics in the difference variable $X - Y$. Significance is assessed by Welch's t-test.

METAtarget extracts information about reactions and enzymes from the KEGG database (www.genome.jp/kegg). Only metabolites annotated as "main" reaction partners in KEGG RPAIR are considered as pair of substrate and product. For each pair of substrate and product, the regulation of the reaction is assessed by the statistics t and Welch's t-test. P-values < 0.05 after Bonferroni correction for the number of tested reactions are considered statistically significant.

2.2 Breast cancer metabolomics data

METAcancer (www.metacancer-fp7.eu) is an EU-funded project aiming at the analysis of the breast cancer metabolome and the discovery of new molecular markers. A series of more than 200 breast cancers was investigated using three different metabolic platforms, GC-MS, LC-MS and NMC. Here, we analyze the METAcancer GC-MS data that include measurements of 124 KEGG annotated metabolites in 188 ER+ and 58 ER- breast cancers. Estrogene receptor status of the METAcancer samples was determined immunohistologically, tumors with ≥ 10% ER positive cells were considered as ER positive. Prior to analysis, metabolomics data were transformed to the log2-scale.

2.3 Breast cancer transcriptomics data

Three publicly available breast cancer gene expression series GSE2034, GSE7390 and GSE11121 were downloaded from the GEO repository (www.ncbi.nlm.nih.gov/geo). All data sets were generated using the same kind of microarrays (Affymetrix HG-U133A GeneChips). The expression series were merged to a large expression data set of 684 nodal negative breast cancers. Microarray data were preprocessed with the standard mas5 method and transformed to the log2-scale. As immunohistological data for estrogen receptor (ER) status were not available for all samples, ER status was derived from gene expression data. Samples with ESR1 absolute expression ≥ 10 (measured by probe set 205225_at) were considered as ER positive, samples with ESR1 expression < 10 as ER negative. 176 of the 684 breast cancers were ER negative, 508 ER positive. Significance of differential expression between ER+ and ER- tumors was assessed by Welch's t-test. P-values < 0.05 after Bonferroni correction for the number of genes were considered statistically significant.

3 Results

GC-MS profiling of 246 breast cancers within the METAcancer project led to the identification of 468 metabolites. 162 out of these could be mapped to known chemical structures and metabolite names, 124 could be found in the KEGG database. Using KEGG RPAIR we identified 91 substrate-product pairs that were main reactants in metabolic reactions.

Next, we analyzed the substrate-product pairs for differential regulation between ER+ and ER- breast cancers. Using the statistics *t* (cf. material and methods section) we detected 13 differentially regulated pairs of reactants. Bar plots show the differential expression between ER+ and ER- breast cancer for substrates and products (Fig. 1A) and the substrate / product ratio for ER+ and ER- breast cancer (Fig. 1B).

Using KEGG REACTION the 13 reactant pairs could be mapped to 51 metabolic reactions (Tab. 1). In KEGG REACTION, the reactant pairs are stored with the EC numbers of the catalysing reactions. This information, together with the information on the human genome was used to map the reactant pairs to 29 human genes.

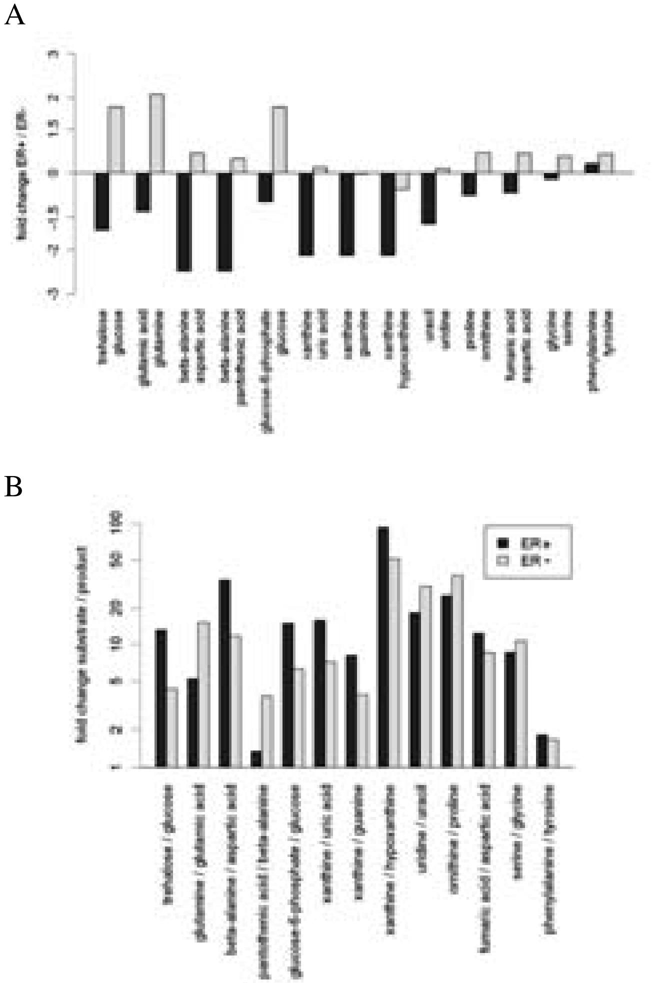


Fig. 1: Substrate-product pairs that are differently regulated in ER+ compared to ER- breast cancer. Significance was assessed by the statistics t (cf. material and methods). **A** Fold change between ER+ and ER- breast cancer in pairs of substrates and products. **B** Fold change between substrates and products in ER+ and ER- breast cancer.

Tab. 1: Differently regulated metabolic reactions between ER+ and ER- breast cancer. After mapping of metabolites to reactions using KEGG RPAIR, differently regulated substrate-product pairs were detected by the analysis of metabolomics data. For each substrate-product pair the numerator of the statistics *t*, a difference of differences (dd), is reported. Catalyzing enzymes were identified using KEGG ENZYME. Enzymes were investigated for differential expression between ER+ and ER- tumors using an independent breast cancer genes expression data set. Meaning of the check marks behind the gene symbols: “+” = significant up-regulation, “-“ = significant down-regulation, “~” = no differential regulation, no check mark = not represented by the microarray.

substrate	product	dd	reaction	enzymes
glucose	trehalose	-3.1	alpha,alpha-Trehalose + H2O <=> 2 D-Glucose	TREH~
glutamic acid	glutamine	2.9	ATP + L-Glutamate + NH3 <=> ADP + Orthophosphate + L-Glutamine	GLUL+
			L-Glutamine + H2O <=> L-Glutamate + NH3	ASNS-, CAD-,GLS-, GLS2+
			ATP + Deamino-NAD+ + L-Glutamine + H2O <=> AMP + Diphosphate + NAD+ + L-Glutamate	NADSYN1+
			ATP + UTP + L-Glutamine + H2O <=> ADP + Orthophosphate + CTP + L-Glutamate	CTPS-, CTPS2~
			2 ATP + L-Glutamine + HCO3- + H2O <=> 2 ADP + Orthophosphate + L-Glutamate + Carbamoyl phosphate	CAD-
			ATP + L-Aspartate + L-Glutamine + H2O <=> AMP + Diphosphate + L-Asparagine + L-Glutamate	ASNS-
			L-Glutamine + D-Fructose 6-phosphate <=> L-Glutamate + D-Glucosamine 6-phosphate	GFPT1-, GFPT2~
			5-Phosphoribosylamine + Diphosphate + L-Glutamate <=> L-Glutamine + 5-Phospho-alpha-D-ribose 1-diphosphate + H2O	PPAT-
			ATP + Xanthosine 5'-phosphate + L-Glutamine + H2O <=> AMP + Diphosphate + GMP + L-Glutamate	GMPS-
			ATP + 5'-Phosphoribosyl-N-formylglycinamide + L-Glutamine + H2O <=> ADP + Orthophosphate + 2-(Formamido)-N1-(5'-phosphoribosyl)acetamidine + L-Glutamate	PFAS+
aspartic acid	beta-alanine	-2.9	L-Aspartate <=> beta-Alanine + CO2	GAD1+, GAD2~
beta-alanine	pantothenic acid	2.8	ATP + (R)-Pantoate + beta-Alanine <=> AMP + Diphosphate + Pantothenate	
			Pantothenate + H2O <=> (R)-Pantoate + beta-Alanine	
glucose	glucose-6-phosphate	-2.4	ATP + D-Glucose <=> ADP + D-Glucose 6-phosphate	GCK~, HK1~, HK2~, HK3-
			D-Glucose 6-phosphate + H2O <=> D-Glucose + Orthophosphate	G6PC~, G6PC2~
			ITP + D-Glucose <=> IDP + D-Glucose 6-phosphate	HK1~, HK2~, HK3-
			dATP + D-Glucose <=> dADP + D-Glucose 6-phosphate	HK1~, HK2~, HK3-
uric acid	xanthine	-2.2	Xanthine + NAD+ + H2O <=> Urate + NADH + H+	XDH~
			Xanthine + H2O + Oxygen <=> Urate + H2O2	XDH~
guanine	xanthine	-2.1	Guanine + H2O <=> Xanthine + NH3	GDA
hypoxanthine	xanthine	-1.8	Hypoxanthine + NAD+ + H2O <=> Xanthine + NADH + H+	XDH~
			Hypoxanthine + Oxygen + H2O <=> Xanthine + H2O2	XDH~
uracil	uridine	1.7	Uridine + H2O <=> Uracil + D-Ribose	
			Uridine + Orthophosphate <=> Uracil + alpha-D-Ribose 1-phosphate	UPP1-, UPP2
ornithine	proline	-1.5	L-Ornithine <=> L-Proline + NH3	
aspartic acid	fumaric acid	-1.4	L-Aspartate <=> Fumarate + NH3	
glycine	serine	1.2	5,10-Methylenetetrahydrofolate + Glycine + H2O <=> Tetrahydrofolate + L-Serine	SHMT1~, SHMT2-
phenylalanine	tyrosine	1.1	Tetrahydrobiopterin + L-Phenylalanine + Oxygen <=> Dihydrobiopterin + L-Tyrosine + H2O	PAH+
			L-Phenylalanine + Tetrahydrobiopterin + Oxygen <=> L-Tyrosine + 4a-Hydroxytetrahydrobiopterin	PAH+

To evaluate the hypothesis that metabolic changes are associated with transcriptional regulation of enzymes, we analyzed an independent gene expression data set of 684 breast cancers. 27 of the 29 enzymes identified before were represented on the microarray, 16 were differentially expressed between ER+ and ER- breast cancer. As shown in Tab. 1, six enzymes were up-regulated in ER+ breast cancer, while ten enzymes were down-regulated.

Examples for the detected reactions are conversion of glucose to glucose-6-phosphate being the first step of glycolysis and conversion of glutamine to glutamate being the first step of glutaminolysis. Both reactions belong to catabolic pathways that can be used for the production of energy and have been described as up-regulated in cancer cells [VCT09]. Fig. 2 shows the differential expression of the human enzymes catalysing these reactions.

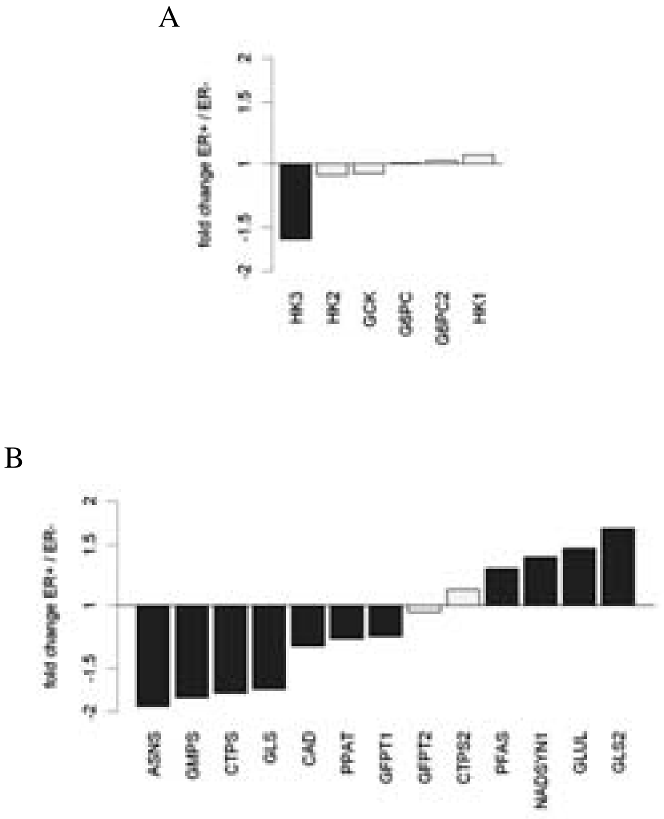


Fig. 2: Differentially expressed enzymes between ER+ and ER- breast cancer. **A** Enzymes catalyzing the first step of glycolysis, glucose -> glucose-6-phosphate. **B** Enzymes catalyzing the first step of glutaminolysis, glutamine -> glutamate. Green bars indicate significant differential expression (after Bonforroni correction for testing 27 genes).

For glycolysis, only hexokinase 3 (HK3) turns out to be differentially expressed between ER+ and ER- breast cancer (down-regulated in ER+ tumors with fold change 1.6). For glutaminolysis, a number of enzymes are differentially expressed between ER+ and ER- tumors (4 up-regulated, 7 down-regulated). Many of these enzymes catalyse several reactions, for example the asparagine synthetase (ASNS) and the tri-functional carbamoyl-phosphate synthetase 2, aspartate tracarbamylase, and dihydroorotase (CAD) that use glutamine as amide-N-donor.

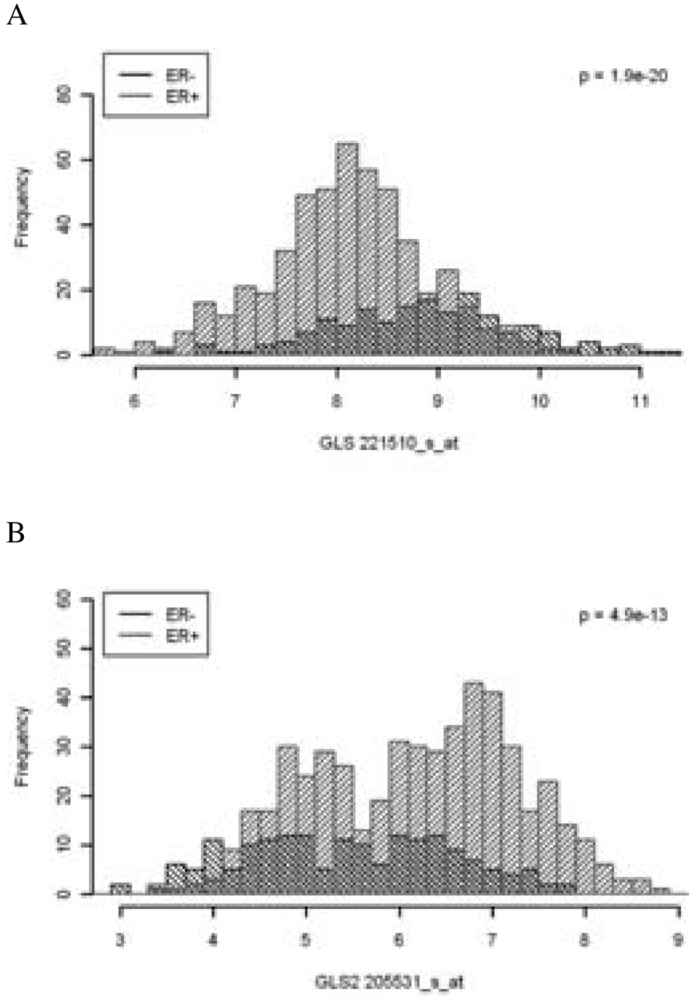


Fig. 3: Differential expression of glutaminase (GLS) and glutaminase 2 (GLS2) between ER+ and ER- breast cancer. Histograms show the distribution of GLS and GLS2 expression as they are measured by the microarrays (log-2 scale).

Among human genes, only the two isoenzymes glutaminase (GLS, also termed kidney type glutaminase) and glutaminase 2 (GLS2, also termed liver type glutaminase) exclusively convert glutamine to glutamate. Fig. 3 shows the expression of GLS (down-regulated in ER+ tumors, fold change 1.7) and the expression of GLS2 (up-regulated in ER+ tumors, fold change 1.7) in ER+ and ER- breast cancer.

4 Discussion

Changes in metabolite concentrations are the final response of a cell to genetic or environmental changes. Metabolite concentrations reflect the outcome of regulation at many molecular levels that takes place in living cells. Indeed, to monitor the final outcome of many regulatory layers is one of the strengths of the metabolomics approach. Regulation of metabolite concentrations can take place at the following levels:

- DNA level: loss of function mutations of enzymes
- RNA level: regulation of enzyme expression (epigenetic, transcriptional or post-transcriptional)
- Protein level: phosphorylation or other post-translational modifications of enzymes
- Interaction level: allosteric regulation of enzymes

A difficulty in the analysis of metabolomics data is connected with backward analysis of the causal chain, in order to understand the mechanism of metabolic regulation and to detect targets for a possible therapeutic intervention. METAtarget implements a reverse engineering step for metabolic networks by using the ratio of product to substrate concentrations as surrogate for the activity of enzymes. METAtarget is based on information on reactant pairs and enzymes that is obtained from the KEGG REACTION database.

In this paper we have assessed the significance metabolic changes by Welch's t-test. Validity of this approach depends on at least approximate normal distribution of the difference variables $X - Y$. As a more conservative alternative, but with costs of losing power, a rank statistics based approach (Mann-Whitney test) can be applied. As unimportant for significance assessment, we did not take into account reversibility or the direction of reactions. Denotation of a reactant as substrate or product is arbitrary.

Using METAtarget, we have analyzed the metabolism of breast cancer cells in dependence of hormone receptor status. Worldwide, immunohistological determination of estrogene receptor status is part of the breast cancer routine diagnostics. Patients with ER+ tumors are known to benefit from hormone therapy (for example treatment with tamoxifen), while ER- breast cancer is known to be a more aggressive breast cancer subtype. Most of the ER- tumors are highly proliferating and have tumor grades 2 or 3 [DRL07]. Furthermore, these tumors include the triple-negative subtype that is difficult to treat and has a poor prognosis.

Analyzing metabolomics data generated within the METAcancer project, we have identified 13 reactant pairs with a shifted equilibrium depending on ER status. The expression pattern of the corresponding 29 enzymes was analyzed in three publicly available gene expression series. 7 enzymes turned out to be significantly up-regulated, 10 enzymes significantly down-regulated in ER+ tumors compared to ER- tumors, while 12 enzymes remained unchanged. All 29 enzymes are interesting as targets, because manipulation of enzyme activity could restore the metabolism towards a less aggressive type. On the other side, differential expression of the 17 regulated enzymes is expected to contribute to the regulation of breast cancer metabolism in dependence of hormone receptor status.

In particular, we detected an up-regulation of glycolysis and glutaminolysis in ER- tumors, compatible with a higher demand on energy of a more aggressive cancer. Targeting these pathways could be an opportunity for treatment. More inside in the regulation processes has been provided by analyzing the gene expression of enzymes that catalyze the entry reaction of glycolysis and of glutaminolysis (Fig. 2).

The expression of glutaminases has been extensively studied before und shown to exhibit tissue-specific expression profiles [SO09]. Co-expression of GLS and GLS2 is a frequent event in human cancer [Pé05]. Furthermore, high GLS expression has been described as being associated with high proliferation rates, whereas repression of GLS and prevalence of GLS2 has been described to be related to quiescent or resting states [LKM69]. This observation is compatible with high expression of GLS in the strong proliferation ER- tumors and higher average expression GLS2 in ER+ tumors, an entity that also contains weak proliferating G1 tumors.

In summary, METAtarget is a new method for reverse engineering of metabolic networks and the detection of targets enzymes from high-throughput metabolomics data. In an application to METAcancer, deregulation of key reactions of glycolysis, glutaminolysis and other pathways could be detected between ER+ and ER- breast cancer.

Acknowledgements

The work was funded by the European Commission, FP7 grant #200327 (METAcancer).

Bibliography

[An09] Antonov, A.V.; Dietmann, S.; Wong, P. et. al.: TICL - a web tool for network-based interpretation of compound lists inferred by high-throughput metabolomics. FEBS J. (2009) 276; pp. 2084-2094.

[De08] Denkert, C.; Budczies, J.; Weichert, W. et. al.: Metabolite profiling of human colon carcinoma - deregulation of tca cycle and amino acid turnover. Mol. Cancer (2008) 7; p. 72.

[De09] Denkert, C.; Budczies, J.; Darb-Esfahani, S. et. al.: A prognostic gene expression index in ovarian cancer - validation across different independent data sets. J. Pathol. (2009) 218; pp. 273-280

[DRL07] Dunnwald, L.K.; Rossing, M.A.; Li, C.I.: Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. Breast Cancer Res. (2007) 9; p. R6.

[Fi01] Fiehn, O.: Combining genomics, metabolome analysis, and biochemical modelling to understand metabolic networks. Comp. Funct. Genomics (2001) 2; pp. 155-168.

[Ga10] Gao, J.; Tarcea, V.G.; Karnovsky, A. et. al.: Metscape: a cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. Bioinformatics (2010) 26; pp. 971-973.

[GHW05] German, J.B.; Hammock, B.D.; Watkins, S.M.: Metabolomics: building on a century of biochemistry to guide human health. Metabolomics (2005) 1; pp. 3-9.

[GS04] Griffin, J.L.; Shockcor, J.P.: Metabolic profiles of cancer cells. Nat. Rev. Cancer (2004) 4; pp. 551-561.

[Jo08] Jourdan, F.; Breitling, R.; Barrett, M.P. et. al.: MetaNetter: inference and visualization of high-resolution metabolomic networks. Bioinformatics (2008) 24; pp. 143-145.

[Ka10] Kanehisa, M.; Goto, S.; Furumichi, M. et. al.: KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. (2010) 38; p. D355-60.

[Le08] Letunic, I.; Yamada, T.; Kanehisa, M. et. al.: iPath: interactive exploration of biochemical pathways and networks. Trends Biochem. Sci. (2008) 33; pp. 101-103.

[LKM69] Linder-Horowitz, M.; Knox, W.E.; Morris, H.P.: Glutaminase activities and growth rates of rat hepatomas. Cancer Res. (1969) 29; pp. 1195-1199.

[Ma09] Matthews, L.; Gopinath, G.; Gillespie, M. et. al.: Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res. (2009) 37; p. D619-22.

[Ok08] Okuda, S.; Yamada, T.; Hamajima, M. et. al.: KEGG atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res. (2008) 36; p. W423-6.

[Pé05] Pérez-Gómez, C.; Campos-Sandoval, J.A.; Alonso, F.J. et. al.: Co-expression of glutaminase K and L isoenzymes in human tumour cells. Biochem. J. (2005) 386; pp. 535-542.

[Sh03] Shannon, P.; Markiel, A.; Ozier, O. et. al.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. (2003) 13; pp. 2498-2504.

[Sr09] Sreekumar, A.; Poisson, L.M.; Rajendiran, T.M. et. al.: Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature (2009) 457; pp. 910-914.

[SO09] Szeliga, M.; Obara-Michlewska, M.: Glutamine in neoplastic cells: focus on the expression and roles of glutaminases. Neurochem. Int. (2009) 55; pp. 71-75.

[VCT09] Vander Heiden, M.G.; Cantley, L.C.; Thompson, C.B.: Understanding the warburg effect: the metabolic requirements of cell proliferation. Science (2009) 324; pp. 1029-1033.

Finding Optimal Sets of Enriched Regions in ChIP-Seq Data

Andreas Gogol-Döring* and Wei Chen
Berlin Institute for Medical Systems Biology,
Max Delbrück Center for Molecular Medicine, Berlin, Germany

andreas.doering@mdc-berlin.de

Abstract: The main challenge when analyzing ChIP-Seq data is the identification of DNA-protein binding sites by finding genomic regions that are enriched with sequencing reads. We present a new tool called *qips* especially suited for processing ChIP-Seq data containing broader enriched regions. Our tool certainly finds all enriched regions that are not exceeded by higher significant alternatives.

1 Introduction

Chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-Seq) [JMMW07] is a common method for genome-wide profiling protein-DNA interactions. In ChIP-Seq antibodies specifically select the proteins of interest together with any piece of randomly fragmented DNA bound to them, and the origins of the selected DNA fragments are then determined by sequencing and mapping to a reference genome. Protein binded regions feature an increased number of mapped sequencing reads. Several software packages were recently published for finding enriched regions in ChIP-Seq sequencing data; a good survey can be found in [PWM09]. Most tools, for example SISSRs [JCB$^+$08], F-Seq [BGCF08], or QuEST [V$^+$08], concentrate on finding short *peaks* indicating nearly punctate protein bindings as it would be typical for transcription factors, whereas nucleosomes or polymerases bind to much broader regions. Some other tools like MACS [Z$^+$08] and CisGenome [JJM$^+$08] can also find longer enriched regions by merging overlapping short regions found in fixed-length sliding windows, but long regions are likely to be cut when a short sliding window is applied, and increasing the windows size would make it impossible to estimate the region boundaries precisely. SICER [ZSZ$^+$09] tries to overcome this problem by partitioning the genome into non-overlapping windows and searching for sequences of succeeding enriched windows which may be interrupt by a limited number of non-enriched windows. However, the precision of this approach is limited the granularity of the applied window grid.

In this paper, we present a new algorithm that indentifies enriched regions of arbitrary length and boundaries in ChIP-Seq data. Our method finds an *optimal set of enriched regions*, which means that it reports an enriched region if there is no better, i.e. more significant, alternative overlapping region. Note that prior approaches for analyzing ChIP-

Seq data do not guarantee to find optimal region sets.

Our tool *qips* (*q*uantification of *IP*-*S*eq) also features a new way for estimating the average DNA fragment length from single-end sequencing data; this is discussed in Section 2.1. In Section 2.2 we describe how *qips* estimates the statistical background from mappability information or, if available, a control data set. A quick approximation formula for scoring candidate regions is presented in Section 2.3. The algorithm for finding enriched regions is described in Section 2.4. We discuss our results in Section 3.1.

2 Methods

The analysis of ChIP-Seq data starts with the mapping of the sequenced reads to a reference genome G using a tool like Bowtie [LTPS09] or RazerS [WER$^+$09]. The *position* of a read in G is the center of the subsequence of G to which the read matches. A read is only used for the following analysis if there is a single ‘best’ match of it in G, because otherwise we cannot infer its true origin. However, this also means that it is hard to detect protein bindings in repetitive genomic regions. This is a general limitation of the ChIP-Seq technology, and our statistical model explicitely take it into account; see Section 2.2 For avoiding biases due to PCR artifacts, we retained only one read at the same position and the same strand orientation. To identify the centers of the ChIP-Seq fragments, which are usually much longer than the sequenced reads, we shift each read downstream by $s = (f - q)/2$, where f is the average fragment length and q the read length. Our method for accurately estimating f from single-end reads is described in Section 2.1.

In the final set of uniquely mapped, non-redundant, and shifted reads we then search for *enriched regions*, namely for intervals $[a, b] \subset G$ containing significantly more read positions than expected by chance, which means that the p-value $p[a, b]$ relative to a given background model is below a certain user-defined threshold α. Our background model assumes an uniform distribution of the reads over all (mappable) positions $i \in G$ or, if available, it accounts for a second control data set obtained, e.g., from a ChIP-Seq using Immunoglobulin G (IgG); see Section 2.2. As *qips* was designed to find enriched regions of arbitrary size, we must also decide whether two neighboring enriched regions $[a, b]$ and $[a', b']$ could in fact be a single enriched region $[a, b']$. We prefer $[a, b']$ instead of $[a, b]$ and $[a', b']$, if and only if the combined region is more significant than the two sub-regions, i.e. if $p[a, b'] < \min(p[a, b], p[a', b']) \leq \alpha$. *qips* computes an *optimal set of enriched regions*, which is defined as follows:

Definition: Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of intervals in a genome G. Two intervals $[a, b]$ and $[a', b']$ *overlap* if $a \leq b'$ and $a' \leq b$. A subset $R \subseteq I$ is a *set of enriched regions* if the intervals in R are pairwise disjoint and if $p(r) \leq \alpha$ for all $r \in R$. R is *optimal* if for each interval $i \in I \setminus R$ exists an interval $r \in R$ overlapping with i and $p(r) \leq p(i)$.

R is unique if $p(i) \neq p(i')$ for any two overlapping intervals $i \neq i' \in I$, otherwise there are several optimal sets. We assume that the differences between those alternative sets have only minor practical relevance and can therefore be ignored. Our tool finds an optimal set of enriched regions for n reads in time O(n^2) and linear space; see Section 2.4.

2.1 Estimating the Fragment Length

ChIP-Seq DNA fragments are usually sequenced only from a single end, so it is not possible to deduce their lengths directly from the data. On the other hand, measuring the fragment length using laboratory equipment does not account for a length bias introduced in the sequencing procedure. Several authors therefore described methods for estimating average fragment lengths from single-end data, either using shift distances between peaks with different strand orientation [Z$^+$08] [V$^+$08] [JJM$^+$08], or the distances between forward reads and their closest reverse read [JCB$^+$08]. All these approaches assume (nearly) punctate peaks in the data, so they are less appropriate for ChIP-Seq of proteins binding to broader regions. Some methods are also susceptible for noise in the data or could be affected by a *locality bias*, which is the preference for reads being mapped to genomic positions where they overlap to other reads; see Figure 1A. We found locality biases of varying intensities in numerous public available data sets from different labs, and since it is partly caused by the limited mappability of short reads to large genomes, it can hardly be avoided completely.

Our method for estimating the average fragment length f relies on the shift between read distributions of different strand orientation: For each forward read at position i we compute the frequency $F_i(d)$ of forward reads and the frequency $R_i(d)$ of reverse reads at position $i + d$. We define the total read frequencies $F(d) = \sum_i F_i(d)/k$ and $R(d) = \sum_i R_i(d)/k$, where the normalization factor k is the total number of forward strand reads. Figure 1A illustrates that R typically resembles F shifted downstream. The mode of R added to the read length q would be a simple estimate for the fragment length f. Here we use a different approach that is less prone to noise and yields more accurate results for skewed fragment length distributions. We computed for shift widths $d > 0$ the average $A(d)$ of the squared difference $(F(j) - R(j + d))^2$ over all $j \notin [-q, q] \cup [d - q, d + q]$, i.e. we exclude all j where $F(j)$ or $R(j)$ could be affected by a locality bias. Then the average fragment length is estimated by $f = q + \text{argmin}_d A(d)$.

We tested our method on various data sets, and it yielded reasonable results even for data containing very broad or unspecific binding, like ChIP-Seq targeting H3K36me3 histone marks or using IgG. For proving the accuracy of our method, we also sampled single-end reads from a paired-end ChIP-Seq data set [WXZ$^+$10] and compared the estimates from different tools to the actual fragment lengths; see Figure 1B.

2.2 Modeling the Statistical Background

A q-mer is *unique* if it occurs only once in a genome G, and its position in G is called a *(uniquely) mappable* position. Repetitive regions of G are characterized by a lower density of mappable positions and therefore contain less reads than regions with higher mappability. However, only few software tools for ChIP-Seq data analysis take mappability variations in genomes into account [REA$^+$09]. Let $mp[a, b]$ be the number of mappable positions in $[a, b] \subseteq G$, then the maximal number of reads in $[a, b]$ after shifting the reads

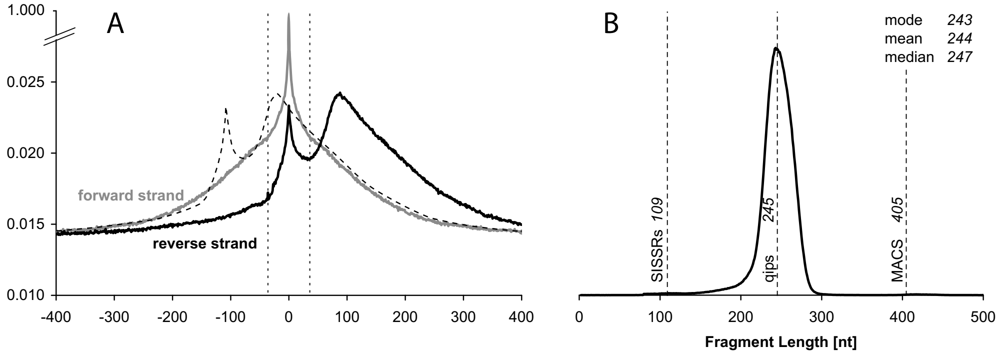


Figure 1: A: Typical normalized read frequencies F and R for a ChIP-Seq data with read length $q = 36$. The peaks between the dotted vertical lines at $-q$ and $+q$ reflect a locality bias. The dashed line is R shifted to the left by $f - q$, where f is the estimated fragment length. B: The fragment length distribution of a paired-end ChIP-Seq data set [WXZ$^+$10]. *qips* estimated the average fragment length very accurately, whereas the results of the two other tools considerably diverged from the actual lengths.

downstream by s is given by:

$$maplen[a, b] = mp[a - s, b - s] + mp[a + s, b + s].$$

Our background model assumes that reads in the data set S are spread independently over the genome G by a *Poisson process*, which means that, given an interval $[a, b] \subset G$, the reads may occur at any position $i \in [a, b]$ with the same rate $\mu = \lambda / \, maplen[a, b]$, where $\lambda = E(count_S[a, b])$ is the expected number of reads in $[a, b]$. These assumptions may be questionable, especially the independence between the reads; nevertheless, this kind of model is very common because there is a lack of better alternatives.

We apply two ways for estimating the λ in a given interval $[a, b]$:

1. The expected number λ_M of reads in $[a, b]$ assuming an uniform distribution of all $count_S(G)$ reads in S to the $maplen(G)$ mappable positions in G is:

$$\lambda_M = \frac{maplen[a, b]}{maplen(G)} \, count(G).$$

   In order to get a more local estimation of λ_M, one could also use a region $G_{part} \subset G$ containing $[a, b]$ instead of the whole genome G.

2. A second estimation λ_C is done if a control data set C is available. We calculate the density of reads in C by:

$$\mu_C = \frac{count_C[a - s, b + s]}{maplen[a - s, b + s]},$$

   i.e. we enlarge $[a, b]$ in both sides by the shift width s to avoid clipping effects due to variations in the fragment lengths. The read density μ_S in S can be estimated given μ_C after normalizing for the different read quantities in both data sets, so we get:

$$\lambda_C = \mu_S\, maplen[a,b] = \mu_C \frac{count_S(G)}{count_C(G)}\, maplen[a,b].$$

If both estimates are available, then we use the maximum $\lambda = \max(\lambda_M, \lambda_C)$.

2.3 Computing p-Values

Let $T = \{t_1, t_2, \ldots, t_n\}$ be s set of different read positions, $t_i < t_j$ for $i < j$. Regarding all intervals containing exactly the reads at the positions $t_i, \ldots, t_j$ is $[t_i, t_j]$, obviously $[t_i, t_j]$ is the interval with maximum read density, so we can restrict the search for enriched regions on intervals starting and ending at read positions. This is a great saving of time, because the typical number of reads in a ChIP-Seq data set is two to three orders of magnitude smaller than the genome length.

Starting with a fixed read position t_i, the probability for finding the next $k = count[t_i + 1, t_j]$ reads within the interval $[t_i + 1, t_j]$ is given by an *Erlang distribution*:

$$f(x; k, \mu) = \frac{\mu^k x^{k-1} e^{-\mu x}}{(k-1)!},$$

where $x = maplen[t_i + 1, t_j]$. The p-value is defined by the cumulative density function:

$$p[t_i, t_j] = \sum_{x \leq maplen[t_i+1, t_j]} f(x; k, \mu) = \frac{\gamma(k, \lambda)}{(k-1)!},$$

where γ is the *lower incomplete gamma function*. Note that $p[t_i, t_j]$ only depends on the actual number k and expected number λ of reads in $[t_i + 1, t_j]$.

In practice, it is often more convenient to deal with logarithmic scores than with p-values, so we further define $score[t_i, t_j] = -\log(p[t_i, t_j])$. Since Algorithm 1 has to calculate a huge amount of scores, we substituted the time consuming computation of the function γ by the following approximation formula:

$$score[t_i, t_j] \approx \lambda - k \log(\lambda) + \log(k!) - 0.08 \log(k)^{1.6}$$

This way we speed up our program by more than 50 times compared to a direct computation of γ using the GNU Scientific Library (GSL) [GDT$^+$10]. For scores ≥ 10, the approximations diverge by less then 5% from the exact values; see Figure 2A.

2.4 Finding Optimal Sets of Enriched Regions

Let $T = \{t_1, \ldots, t_n\}$ be a sorted set of interval boundaries. FINDOPTIMALSET (see Algorithm 1) calculates an optimal set R of enriched regions in two steps: First, it determines for each start position $t_i \in T$ the optimal end position $t_{E[i]}$, where $[t_i, t_{E[i]}]$ must

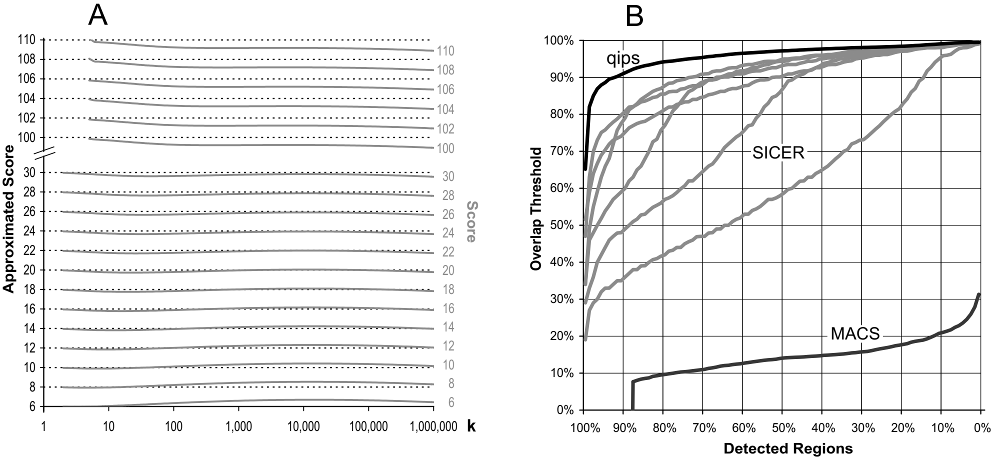


Figure 2: A: Comparison between exact and approximated scores. B: Fraction of detected regions depending on the required minimum overlap between actual and predicted regions for *qips* (this paper), MACS, and SICER with different parameter settings.

not overlap with any higher scoring region starting at $t_k > t_i$. Second, the algorithm selects intervals $[t_i, t_{E[i]}]$ with increasing starting positions t_i. Obviously, the resulting set R is a set of non-overlapping enriched regions. We show that R is optimal as follows: Let $\mathcal{M}$ be the set of intervals with maximum score in $I = \{[t_i, t_j] \mid t_i, t_j \in T\}$, let $\mathcal{M}' \subset \mathcal{M}$ be the intervals in $\mathcal{M}$ with maximum start position, and $[t_i, t_j] \in \mathcal{M}'$ the interval with minimum t_j. The array E is constructed such that $E[t_i] = j$ and $E[t_k] < i$ for all $k < i$, hence i is not skipped in line 18 of Algorithm 1, and therefore $[t_i, t_j] \in R$. The optimality of R follows by applying structural induction to the remaining sets of boundaries $\{t_1, \ldots, t_{i-1}\}$ and $\{t_{j+1}, \ldots, t_n\}$.

The algorithm can easily be modified such that it restricts the search to a subset of I. For example, *qips* allows to set the minimum and maximum length as well as the minimum number of reads in a candidate region. Moreover, it is possible to exclude any interval containing a *drop*, which we define here as an interval $[t_i, t_j]$ having a certain minimum length and either contains less reads than expected by chance, or has a mappability $map[t_i, t_j]/(t_j - t_i)$ below a minimum threshold. A drop cuts the search space into two parts, hence the run time of the algorithm gets linear after choosing appropriate drop parameters.

```
▷ FINDOPTIMALSET $(T = \{t_1 \dots t_n\})$
$S[j] \leftarrow 0$ for all $j \in \{1, \dots, n\}$
for $i \leftarrow n$ down to $1$ do
    $s_{min} \leftarrow -\log(\alpha)$
    $s_{opt} \leftarrow 0$
    $j_{opt} \leftarrow$ nil
    for $j \leftarrow i$ to $n$ do                                Find optimal end
        if $score[t_i, t_j] > \max(s_{min}, s_{opt})$ then        position $t_{E[i]}$ for start
            $s_{opt} \leftarrow score[t_i, t_j]$                  position $t_i$
            $j_{opt} \leftarrow j$
        $s_{min} \leftarrow \max(s_{min}, S[j])$
    $S[i] \leftarrow s_{opt}$
    $E[i] \leftarrow j_{opt}$
$R \leftarrow \{\}$
$i \leftarrow 1$
while $i < n$ do
    if $E[i] \neq$ nil then
        $R \leftarrow R \cup \{[t_i, t_{E[i]}]\}$                 Select optimal
        $i \leftarrow E[i] + 1$                                    regions $[t_i, t_{E[i]}]$
    else
        $i \leftarrow i + 1$
return $R$
```

Algorithm 1: Finding an optimal set of enriched regions. T is a sorted set of interval boundaries, i.e. read positions, and α the p-value threshold. $score[t_i, t_j] = -\log(p[t_i, t_j])$.

3 Results and Discussion

3.1 Results

We simulated threefold enriched regions, each of length $10kb$, on a $\mu = 1\%$ read density background. This data set was used to compare *qips* with SICER, which is a tool especially designed for searching long enriched regions, and with the popular peak finder MACS. The output of SICER depends very much on the input parameters, so we tried several settings. MACS was started with the `--nolambda` command line option for finding longer enriched regions. *qips* detected all enriched regions in the data set, whereas MACS totally missed about 12% of them. MACS and SICER (for some settings) also splitted some of the enriched regions into smaller parts.

We measured the *overlap* between two regions by the number of common bases divided by the length of the longer region. An enriched region was counted among the detected regions, if its overlap to one region in the tool output file was above a certain threshold. Figure 2B shows the sensitivity of the three tools depending on this overlap threshold. It can be seen that *qips* detects enriched regions more precisely than the competitors.

3.2 Discussion

Our approach performs an exhaustive search of all possibly enriched regions and, consequently, should have better chances to detect enriched regions than a heuristical approach that limits the search space. The results presented above illustrate that, at least in some cases, *qips* indeed has some advantages compared to previously published tools like SICER or MACS. On the other hand, our algorithm takes quadratic run time and is therefore significantly more time-consuming than other tools. Applying a relaxed drop condition can significantly improve the run time, but this also increases the risk for missing high scoring enriched regions. A thorough test of our tool both for simulated and real ChIP-Seq data would help to find a good balance between the sensitivity and the performance of our software. This is future work.

3.3 Implementation

We implemented *qips* in C++ and Python, using the GNU Scientific Library [GDT$^+$10] and SeqAn [DWRR08]. The program is controlled by a make file, so it can simply be parallelized by specifying the GNU make `-f` command line option, or distributed to a computer cluster using the Sun Grid Engine `qmake` tool.

Our software will be published free and open source.

References

[BGCF08] A. P. Boyle, J. Guinney, G. E. Crawford, and T. S. Furey. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, 24:2537–2538, 2008.

[DWRR08] Andreas Döring, David Weese, Tobias Rausch, and Knut Reinert. SeqAn an efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, 9:11, 2008.

[GDT$^+$10] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, and F. Rossi. *GNU Scientific Library Reference Manual, Edition 1.14*. Network Theory Ltd, 2010. www.gnu.org/software/gsl/.

[JCB$^+$08] R. Jothi, S. Cuddapah, A. Barski, K. Cui, and K. Zhao. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-seq data. *Nucleic Acids Res.*, 36:5221–5231, 2008.

[JJM$^+$08] H. Ji, H Jiang, W. Ma, D. S. Johnson, R. M. Myers, and W. H. Wong. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotechnol.*, 26:1293–1300, 2008.

[JMMW07] D. S. Johnson, A. Mortazavi, R. M. Myers, and B. Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–1502, 2007.

[LTPS09] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25, 2009.

[PWM09] S. Pepke, B. Wold, and A. Mortazavi. Computation for ChIP-seq and RNA-seq studies. *Nat. Methods*, 6:S22–S32, 2009.

[REA$^+$09] J. Rozowsky, G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder, and M. B. Gerstein. PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, 27:66–75, 2009.

[V$^+$08] A. Valouev et al. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, 5:829–834, 2008.

[WER$^+$09] D. Weese, A.-K. Emde, T. Rausch, A. Döring, and K. Reinert. RazerS – fast read mapping with sensitivity control. *Genome Res.*, 19(9):1646–1654, 2009.

[WXZ$^+$10] C. Wang, J. Xu, D. Zhang, Z. A. Wilson, and D. Zhang. An effective approach for identification of in vivo protein-DNA binding sites from paired-end ChIP-Seq data. *BMC Bioinformatics*, 11(81), 2010.

[Z$^+$08] Y. Zhang et al. Model-based analysis of ChIP-seq (MACS). *Genome Biol.*, 9:R137.9, 2008.

[ZSZ$^+$09] C. Zang, D. E. Schones, C. Zeng, K. Cui, K. Zhao, and W. Peng. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. *Bioinformatics*, 25:1952–1958, 2009.

Learning pathway-based decision rules to classify microarray cancer samples

*Enrico Glaab, Jonathan M. Garibaldi and Natalio Krasnogor
School of Computer Science, University of Nottingham, United Kingdom

enrico.glaab@cs.nott.ac.uk

Abstract:

Despite recent advances in DNA chip technology current microarray gene expression studies are still affected by high noise levels, small sample sizes and large numbers of uninformative genes. Combining microarray data with cellular pathway data by using new integrative analysis methods could help to alleviate some of these problems and provide new biological insights.

We present a method for learning simple decision rules for class prediction from pairwise comparisons of cellular pathways in terms of gene set expression levels representing the up- and down- regulation of pathway members. The procedure generates compact and comprehensible sets of rules, describing changes in the relative ranks of gene expression levels in pairs of pathways across different biological conditions. Results for two large-scale microarray studies, containing samples from prostate cancer and B-cell lymphoma patients, show that the method provides robust and accurate rule sets and new insights on differentially regulated pathway pairs. However, the main benefit of these predictive models in comparison to other classification methods like support vector machines lies not in the attained accuracy levels but in the ease of interpretation and the insights they provide on the relative regulation of cellular pathways in the biological conditions under consideration.

1 Introduction

Classification of microarray gene expression samples often suffers from several limitations resulting from the high dimensionality of the data, a typically small number of available samples, and from various sources of technical and biological noise. In recent years, several methods have extended or replaced classical machine learning methods to provide more compact, robust and easily interpretable classification models. These approaches reduce the prediction model complexity and increase its robustness by using regularization and shrinkage techniques [AMD$^+$05, GHT07], by generating more human-interpretable machine learning models, which are based on simple decision rules [A$^+$06, BK08], or by using more robust data representations and model formulations, e.g. computing discretized expression values or rank scores [LGGV08, WEB05] or only considering relative expression values by comparing pairs of genes [G$^+$04a, TNX$^+$05].

In this paper, we address the problem of low model robustness due to noise by combining ideas from the techniques mentioned above with an approach to analyse the data at the level of pathways instead of at the single-gene level. Briefly, we map the genes in a microarray study onto cellular pathways and processes from public databases and learn simple decision rules for sample classification by comparing gene expression levels in pairs of pathways. Rules describing single pathway-pairs are then weighted and combined into a unified classification model by applying a boosting algorithm. The approach can be understood as a methodological extension of the "top-scoring pairs" (TSP) algorithm [G$^+$04a, TNX$^+$05], which identifies discriminative pairs of genes in microarray data, and has therefore been named "top-scoring pathway pairs" (TSPP) algorithm. Moreover, we draw inspiration from other pathway-based microarray analysis approaches, which use summarized expression values for genes in cellular pathways and processes for enrichment analysis (e.g. the methods GSEA [S$^+$05], MaxMean [ET07] and the global test [G$^+$04b]) or as features for sample classification [G$^+$05].

In contrast to previous methods comparing single gene expression values or summarized expression values for single pathways against fitted threshold values, TSPP provides increased robustness by at the same time combining expression levels of multiple genes into "pathway expression fingerprints" and making pairwise, relative comparisons between pathways. In summary, the TSPP approach is not designed to compete with existing microarray sample classification and data mining methods, but to complement them with the following added benefits:

- New biological insights can be gained from easily interpretable decision rules on the relative up- and down-regulation of cellular pathways.
- The prediction models are applicable to data from other microarray platforms without requiring that all platforms contain the same genetic probes and that cross-study normalization is applied (the integration takes place at the level of pathways, and the gene expression values are replaced by rank scores).
- By summarizing the expression values of multiple genes belonging to the same pathway, the dimensionality of the data is reduced (from about 50.000 genes to a few hundred pathways) and the summarized "pathway expression fingerprints" have a higher robustness than single gene expression vectors (however, at the expense of losing detail; therefore single-gene based methods should be applied additionally).

2 Methods

The TSPP algorithm identifies, scores and combines decision rules based on pathway-pairs according to the following five-step procedure:

1. **Rank score transformation**:

   A gene expression matrix X with dimension $n \times p$ (n: number of samples, p: number of genes) and class labels y for the samples is read as input and transformed into a "rank matrix" R by sorting the expression values for each gene across the n samples and replacing them with their position index in the sorted vector (ties are handled by replacing equal values by the mean of the corresponding position indices).

2. **Pathway mapping**:

   Gene sets representing cellular pathways and processes are extracted from a public database (e.g. KEGG, Gene Ontology, BioCarta or Reactome). Pathway assignments are computed for the p genes in the microarray input data by testing whether they occur in these gene sets. For genes which cannot be assigned to a pathway the corresponding rows are removed from matrix R.

3. **Scoring of pathway pairs**:

   To score a pair of pathways as being useful for discriminating between two sample class labels 1 and 2, e.g. “tumour (1) vs. normal (2)” or “drug treatment (1) vs. no treatment (2)”), the pathway-submatrices R_1 and R_2, corresponding to these two samples classes, are extracted from matrix R based on the mappings from step 2. The matrices R_1 and R_2 are then reduced to vectors r_1 and r_2 by replacing each column of expression level ranks by its median value. For a two-class problem, the score for a pathway-pair is then obtained by comparing the median ranks in pathway 1 to those in pathway 2 and computing the maximum of two relative frequencies: The relative frequency of samples which are up-regulated for class 1 and down-regulated for class 2, and vice versa, the relative frequency of cases which are down-regulated for class 1 and up-regulated for class 2 (i.e. there are two possibilities for the relation of sample ranks in two pathways to differ across the sample classes). Given the sets of column indices for two sample classes S_1 and S_2, the final score can thus be computed as follows:

$$partial\_score_1 = \sum_{i \in S1} I(r_{1i} >= r_{2i}) + \sum_{i \in S2} I(r_{1i} < r_{2i}) \tag{1}$$

$$partial\_score_2 = \sum_{i \in S1} I(r_{1i} < r_{2i}) + \sum_{i \in S2} I(r_{1i} >= r_{2i}) \tag{2}$$

$$score = \frac{max(partial\_score_1, partial\_score_2)}{|S1| + |S2|} \tag{3}$$

   where I is the indicator function. For a multi-class problem, a similar score can be obtained by computing the mean of the scores obtained for all pairs of sample classes.

4. **Identification of top-scoring pairs**:

   By default top-scoring pathway pairs (TSPPs) are identified by performing an exhaustive search across all pairs of pathways. This should be feasible in most practical applications, because the number of pathways is typically much smaller than the number of genes, and the scoring method is kept simple. Moreover, the method does not assume that all genes in a pathway are either up- or down-regulated, but searches for pairs of pathways for which many genes occurring in the first pathway change their relation of expression level ranks across the sample classes to genes in the second pathway. Nevertheless, it might be beneficial to investigate whether alterations in the pathway definitions can provide improved results. Therefore, the user can alternatively let the algorithm introduce “mutations” into the pathway gene sets, by randomly adding or deleting genes up to a small user-defined maximum number of mutations, and replacing the exhaustive search by a previously published evolutionary search algorithm [JUA05]. Only one modification is applied to this algorithm: A genome contains two bit-vectors representing two pathways and mutations are only applied to one of these bit-vectors, selected randomly. The scoring function in the evolutionary algorithm is the same as for the exhaustive search.

5. **Classification model generation**:

   Each TSPP provides a simple decision rule for classifying microarray samples depending on the relative median expression value ranks of their genes in a pair of pathways. To combine multiple TSPPs into a unified classification model, we use the TSPP decision rules as “base classifiers” in the Adaboost.M1 algorithm [FS96], adding one decision rule at a time to the boosting model based on the order of the TSPP-scores computed in step 3. This boosting scheme assigns weights to each decision rule in the combined ensemble model, accounting for a rule’s prediction accuracy and capacity to correctly classify samples that were misclassified by decision rules added in previous iterations of the algorithm. Previous experiments with boosting and ensemble techniques applied to microarray data [GGK09] have shown that improvements can be obtained both in terms of robustness and accuracy.

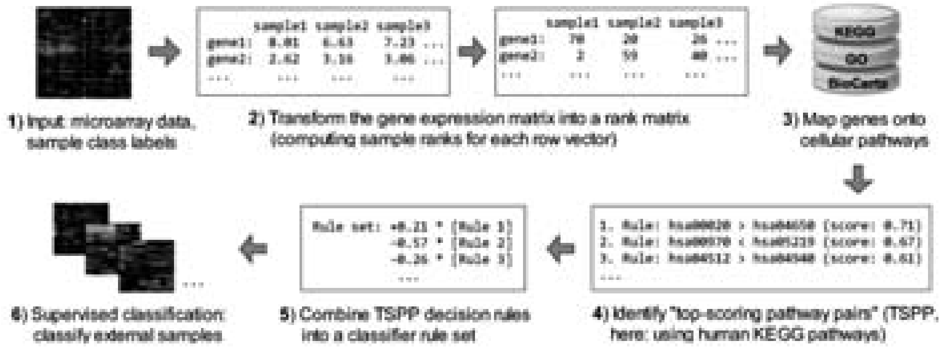

sample1 sample2 sample3
gene1: 8.01 6.63 7.23 ...
gene2: 2.62 3.16 3.06 ...
... ...
sample1 sample2 sample3
gene1: 78 20 26 ...
gene2: 2 59 40 ...
... ...
KEGG
GO
BioCarta
1) Input: microarray data, sample class labels
2) Transform the gene expression matrix into a rank matrix (computing sample ranks for each row vector)
3) Map genes onto cellular pathways
1. Rule: hsa00020 > hsa04650 (score: 0.71)
2. Rule: hsa00970 < hsa05219 (score: 0.67)
3. Rule: hsa04512 > hsa04540 (score: 0.61)
...
Rule set: +0.21 * [Rule 1]
-0.57 * [Rule 2]
-0.26 * [Rule 3]
...
...
6) Supervised classification: classify external samples
5) Combine TSPP decision rules into a classifier rule set
4) Identify "top-scoring pathway pairs" (TSPP, here: using human KEGG pathways)

Figure 1: An overview of the workflow in the TSPP algorithm (example data is derived from a human prostate cancer microarray dataset [S^{+}02b])

3 Results

The TSPP algorithm was applied to the gene expression matrices from two public microarray studies covering different types of cancer: B-cell lymphoma [S^+02a] (7129 genes and 77 samples) and prostate cancer [S^+02b] (12600 genes and 102 samples). Both datasets contain samples from two biological classes: In the B-cell lymphoma dataset 58 samples were obtained from patients suffering from diffuse large B-cell lymphoma (class D), while the remaining samples derive from a related follicular B-cell lymphoma (class F). The prostate cancer expression measurements were obtained from 50 healthy control tissues (class C) and 52 tumour tissues (class T) (for details on the normalization and preprocessing of the datasets, see the Data Sets section).

To evaluate the predictive accuracy for TSPP-models generated for these datasets, we applied an external leave-one-out cross-validation (LOOCV) procedure using different numbers of top-scoring pairs k (for k = 1, 3, 5, 10 and 15) and including all modelling steps in the cross-validation procedure. The parameter k can be regarded as a bias/variance trade-off, enabling the user to control the complexity of the generated classifiers. The cross-validation results, computed both for mappings of genes to KEGG pathways and to Gene Ontology (GO) terms, include the average accuracy, sensitivity and specificity for each LOOCV run and are shown in Tables 1 and 2.

Table 1: **Leave-one-out cross-validation results (TSPP on KEGG database)**

Dataset	No. of top-scoring pairs	Sensitivity (%)	Specificity (%)	Avg. Accuracy (%)
Prostate cancer	1	83.7	71.7	77.5
	3	87.8	73.6	80.4
	5	85.7	77.4	81.4
	10	77.6	73.6	75.5
	15	79.6	64.2	71.6
Lymphoma	1	64.9	85.0	70.1
	3	68.4	90.0	74.0
	5	78.9	90.0	81.8
	10	77.2	90.0	80.5
	15	75.4	90.0	79.2

In summary, average classification accuracies above 70% were obtained in all cases, and for both datasets the best accuracies (prostate cancer: 81.4%, DLBCL: 81.8%) were achieved when using 5 top-scoring pairs, suggesting that k = 5 represents a reasonable bias/variance trade-off. The sensitivity and specificity scores were in a roughly similar percentage range.

Apart from using the decision rules for class prediction, their simplicity also makes them suitable for direct human interpretation. The ten top-scoring pathway pairs for each dataset are shown in Tables 4 and 5. Interestingly, the top-ranked rule for the prostate cancer dataset contains the KEGG-pathways “Prostate cancer” and “Insulin signaling”, which are both known to be de-regulated in the disease [SK03, H^+01]. However, the results

Table 2: **Leave-one-out cross-validation results (TSPP on GO database)**

Dataset	No. of top-scoring pairs	Sensitivity (%)	Specificity (%)	Avg. Accuracy (%)
Prostate cancer	1	83.7	67.9	75.5
	3	89.8	67.9	78.4
	5	89.8	69.8	79.4
	10	91.8	66.0	78.4
	15	85.7	67.9	76.5
Lymphoma	1	68.4	80.0	71.4
	3	57.9	90.0	66.2
	5	71.9	90.0	76.6
	10	52.6	90.0	62.3
	15	71.9	85.0	75.3

Table 3: **Leave-one-out cross-validation results (Gene-based: eBayes & SVM)**

Dataset	No. of features (genes)	Sensitivity	Specificity	Avg. Accuracy (%)
Prostate cancer	2	88.0	84.6	86.3
	6	96.0	88.5	92.2
	10	96.0	86.5	91.2
	20	90.0	88.5	89.2
	30	90.0	90.4	90.2
Lymphoma	2	91.4	68.4	85.7
	6	93.1	78.9	89.6
	10	94.8	94.7	94.8
	20	96.6	84.2	93.5
	30	98.3	100.0	98.7

also point to relative de-regulations in other pathways with less obvious associations to the cancer disease, e.g. “Pyrimidine metabolism” and “Glycerolipid metabolism”, with a score close to the best-ranked pair. Similarly, for the B-cell dataset the top-ranked pathway pairs contain pathways known to be associated with B-cell neoplasia, e.g. the “Wnt signaling pathway” [QERR03, LB03], whereas for other pathways no direct and specific associations with the disease are known. In spite of the class-imbalance in this dataset, the prediction models did not display a preference to assign samples to the majority class; however, similar to other statistical methods for microarray data analysis, problems with robustness can occure whe the sample size per condition is very small. Thus, when planning a microarray study, the experimenter might first want to study the literature on sample size estimation [LHC10], microarray study design [Chu02] and sampling techniques to alleviate these problems [VHKNW09].

It is also important to note that in a top-scoring pathway pair (TSPP) not necessarily both pathways are differentially regulated across the sample classes, but one pathway might have a constant expression, while the other pathway is highly de-regulated in one of the sample classes. The main benefit of comparing pairs of pathways lies in the possibility to avoid comparing single pathways against fitted thresholds, which would more likely be affected by experimental bias and thus provide prediction models with higher generalization error. However, if a user's main goal is not to obtain a prediction model from the TSPP-algorithm, but to identify pathway associations, then TSPPs in which one of the pathways is not differentially regulated across the sample classes can easily be identified and filtered out by computing the variance for the corresponding gene expression vectors and removing TSPPs containing a pathway with low variance.

When using the evolutionary search methodology and allowing the algorithm to introduce small numbers of random gene deletions and insertions into the pathways (up to five genes), in spite of the higher flexibility of this method, in all experiments the prediction accuracies are either similar or lower than those obtained for the original pathways using an exhaustive search (data not shown). The weaker performance might result from an entrapment in local minima due to the expansion of the search space, but could also suggest that the original pathways and processes are already well defined and therefore hard to optimize based on an evolutionary search procedure.

Overall, the results from the cross-validation analysis and the lists of top-scoring pathways show that the method can generate compact predictive models with both high interpretability and high accuracy in comparison with a random model predictor (when measuring this using the "proportional chance criterion" by Huberty [Hub94], we obtain p-values < 0.01 in all cases). To put these results into relation with existing machine learning methods based on single genes as predictors, we applied a C-SVM from the e1071 R software package [DHL$^+$05], a wrapper for the well-known LibSVM library [CL01], with different kernel functions, including the radial basis function and polynomial kernels with a degree up to 3 (the results for the best kernel, a linear SVM, are reported in Table 3). The gene-based SVM-models achieve higher average accuracies than pathway-based models, with the best models reaching more than 90% accuracy on both datasets; however, these models only contain information on the relevance of single genes for the prediction and do not enable an interpretation of the data on the level of cellular pathways and processes. Although the simple decision rules generated by the TSPP algorithm do not reach the highest accuracies obtained by the support vector machine on single genes, their high interpretability and significant predictive information content allow the user to quickly identify cases, in which the relative gene expression in pathway pairs is differentially regulated across different biological conditions.

To investigate the utility of top-scored pathway pairs (TSPPs) in more detail, we have mapped the genes in these pathways onto their corresponding proteins in a large-scale protein-protein interaction network, consisting of 38857 interactions between 9392 proteins assembled from direct binary interactions in a previous study [GBKV10]. Figure 2 a) shows the largest connected component of an example mapping for the TSPP with the highest score on the Prostate cancer dataset, "hsa05215 Prostate cancer" vs. "hsa04910 Insulin signaling pathway" (see also Figure 1), revealing a strong network of interactions between these pathways, which also share a significantly large set of overlapping genes/proteins (q-value = 5.1E-17, when testing the hsa04910 pathway against all other KEGG pathways using the one-sided Fisher exact test and adjusting for multiple testing with the Benjamini-Hochberg method [BH95]). However, the TSPP-method also points the user to differentially regulated pathway pairs which would not be detected as signif-

Table 4: **Top-ranked pathway pairs (Prostate cancer data)**

Rank	Pathway 1	Pathway 2	Direction	Score
1	hsa05215 Prostate cancer	hsa04910 Insulin signaling pathway	down	0.81
2	hsa00240 Pyrimidine metabolism	hsa00561 Glycerolipid metabolism	up	0.80
3	hsa04540 Gap junction	hsa05210 Colorectal cancer	up	0.78
4	hsa04115 p53 signaling pathway	hsa00230 Purine metabolism	down	0.75
5	hsa04510 Focal adhesion	hsa00071 Fatty acid metabolism	down	0.75
6	hsa04514 Cell adhesion molecules (CAMs)	hsa04610 Complement and coagulation cascades	up	0.72
7	hsa03050 Proteasome	hsa01430 Cell Communication	up	0.69
8	hsa04920 Adipocytokine signaling pathway	hsa04730 Long-term depression	up	0.69
9	hsa04810 Regulation of actin cytoskeleton	hsa04530 Tight junction	down	0.65
10	hsa04512 ECM-receptor interaction	hsa04110 Cell cycle	down	0.63

The 10 top-ranked pathways for the prostate cancer dataset based on the TSPP-score (Direction “down” means that in the healthy control samples, pathway 1 is down-regulated in relation to pathway 2, whereas in the prostate cancer samples, pathway 1 is up-regulated in relation to pathway 2, and respectively, “up” means the pathways have opposite relations in the two sample classes).

icantly associated based on an overlap-based significance test, e.g. Figure 2 b) shows the largest connected component for the TSPP “hsa04115 p53 signaling pathway” vs. “hsa00230 Purine metabolism”, with only two overlapping proteins, but a multitude of direct binary protein-protein interactions between the two pathways. Further experimental evidence for an association between these pathways is provided by a study showing that the inhibition of de novo purine synthesis by the drug “AG2034”, which also inhibits prostate cancer cell growth, increases the expression levels of p53 [OKM09]. Thus, although the up- and down-regulation of top-scoring pathway pairs does not necessarily result from a regulatory relationship between the pathways, the analysis of the TSPPs can help to point the user to associations between pathways, which would remain unnoticed by other methods, such as an overlap-based Fisher test.

.

3.1 Data sets

3.1.1 Diffuse large B-cell lymphoma (DLBCL)

The DLBCL data set [S$^+$02a] contains expression values for 7,129 genes and 77 microarray samples, 58 of which were obtained from patients suffering from diffuse large B-cell lymphoma (D), while the remaining samples derive from a related B-cell lymphoma, called follicular lymphoma (F). The experiments in this microarray study had been carried out on an Affymetrix HU6800 oligonucleotide platform [Aff01].

To pre-process the raw data, we applied the “Variance stabilizing normalization” [HvHS$^+$02]

Table 5: **Top-ranked pathway pairs (B-Cell lymphoma data)**

Rank	Pathway 1	Pathway 2	Direction	Score
1	hsa00020 Citrate cycle (TCA cycle)	hsa04310 Wnt signaling pathway	down	0.88
2	hsa00052 Galactose metabolism	hsa04664 Fc epsilon RI signaling pathway	down	0.87
3	hsa04670 Leukocyte transendothelial migration	hsa03050 Proteasome	up	0.87
4	hsa04514 Cell adhesion molecules (CAMs)	hsa00030 Pentose phosphate pathway	up	0.86
5	hsa04730 Long-term depression	hsa00240 Pyrimidine metabolism	up	0.85
6	hsa00562 Inositol phosphate metabolism	hsa00051 Fructose an mannose metabolism	up	0.84
7	hsa00220 Urea cycle and metabolism of amino groups	hsa00980 Metabolism of xenobiotics by cytochrome P450	down	0.84
8	hsa04540 Gap junction	hsa00330 Arginine and proline metabolism	up	0.84
9	hsa00252 Alanine and aspartate metabolism	hsa04630 Jak-STAT signaling pathway	down	0.84
10	hsa00970 Aminoacyl-tRNA biosynthesis	hsa04912 GnRH signaling pathway	down	0.81

The 10 top-ranked pathways for the B-Cell lymphoma dataset based on the TSPP-score (Direction “down” means that in the DLBCL samples, pathway 1 is down-regulated in relation to pathway 2, whereas in the follicular B-cell lymphoma samples, pathway 1 is up-regulated in relation to pathway 2, and respectively, “up” means the pathways have opposite relations in the two sample classes).

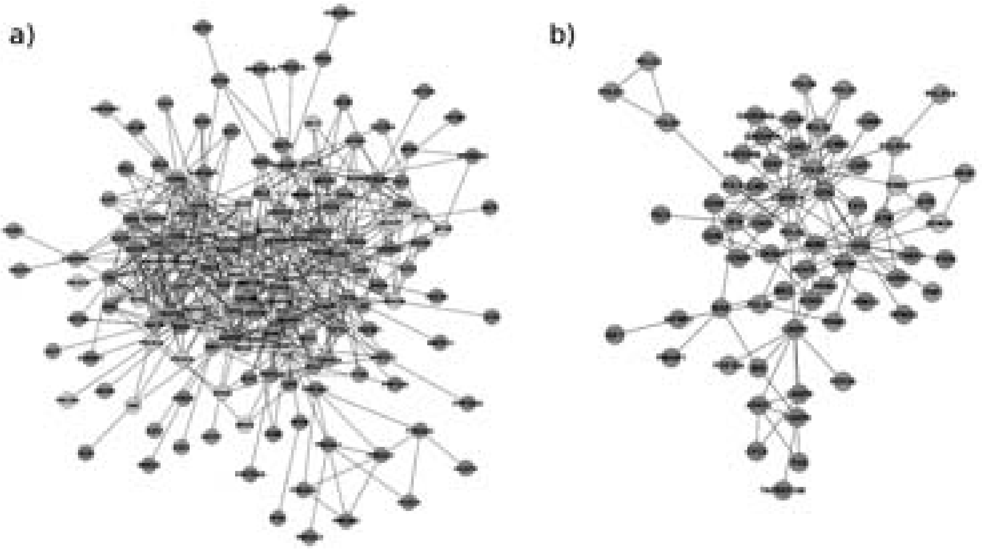


Figure 2: Analysing TSPPs in a protein-protein interaction network: a) Largest connected component for KEGG pathways: “Prostate cancer” and “Insulin signaling” (blue: Prostate cancer, red: Insulin signaling, green: members in both pathways); b) Largest connected component for KEGG pathways “P53 signaling” and “Purine metabolism” (blue: P53 signaling, red: Purine metabolism, green: members in both pathways)

to filter out intensity-dependent variance (this was done using the vsn-library and the expresso-package in the R statistical learning environment [Tea10]). Moreover, we applied thresholding based on the suggestions in the supplementary material of the original publication [S$^+$02a] and a "fold change"-filter to remove all genes with less than a 3-fold change between the maximum and minimum expression value.

3.1.2 Prostate cancer

The prostate cancer data set [S$^+$02b] consists of expression measurements for 12,600 genetic probes across 50 healthy control tissues (C) and 52 prostate cancer tissues (C). All experiments have been carried out on Affymetrix Hum95Av2 arrays [Aff01]. Due to the large number of samples and memory limitations of the expresso-package (used to normalize the other two data sets), we applied the fast GeneChip RMA (GCRMA) normalization algorithm [WI05]. Moreover, we employed thresholding based on the suggestions in the original publication of the dataset [S$^+$02b] and a fold change filter to remove all probes with less than a 2-fold change between the maximum and minimum expression value.

Table 6: **Data sets used in this paper**

Data set	Platform	No. of genes	No. of samples class 1; class 2	references
B-cell lymphoma	Affymetrix	7,129	58 (D) ; 19 (F)	[S$^+$02a]
Prostate cancer	Affymetrix	12,600	52 (T) ; 50 (C)	[S$^+$02b]

4 Conclusion

We present a new method for extracting pathway-based decision rules from combined gene expression data and gene sets representing cellular pathways and processes. When applying prediction models derived from these decision rules for sample classification on two public microarray cancer datasets, we obtain compact and easily interpretable models with significant predictive information content. The generated decision rules are robust against monotonic transformations of the data, and the algorithm is easy to implement and has a comparatively short run-time due to the reduction of the data dimensionality when considering summarized pathway expression values instead of gene expression values. Moreover, these models also enable a different interpretation of microarray data by analysing the data at the level of pathways. Specifically, the top-scoring pathway pairs can point the user to regulatory relationships or other functional associations between the corresponding pathways. In summary, the TSPP algorithm provides both a novel method to generate compact and accurate classification models and a new exploratory tool to analyse microarray data at the level of pairwise pathway-relations.

References

[A$^+$06] G. Alexe et al. Breast cancer prognosis by combinatorial analysis of gene expression data. *Breast Cancer Res*, 8(4):R41, 2006.

[Aff01] Affymetrix. *Affymetrix Microarray Suite User Guide, Version 5*, 2001.

[AMD$^+$05] N. Ancona, R. Maglietta, A. D'Addabbo, S. Liuni, and G. Pesole. Regularized least squares cancer classifiers from DNA microarray data. *BMC Bioinformatics*, 6(Suppl 4):S2, 2005.

[BH95] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B (Methodological)*, 57:289–300, 1995.

[BK08] J. Bacardit and N. Krasnogor. Fast rule representation for continuous attributes in genetics-based machine learning. In *Genet Evol Comput Conf*, pages 1421–1422. ACM, 2008.

[Chu02] G.A. Churchill. Fundamentals of experimental design for cDNA microarrays. *Nat Genet*, 32:490–495, 2002.

[CL01] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001.

[DHL$^+$05] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, A. Weingessel, and M.F. Leisch. Misc functions of the department of statistics (e1071), TU Wien, 2005. R-Package e1071 version 1.5-19.

[ET07] B. Efron and R. Tibshirani. On testing the significance of sets of genes. *Ann Appl Stat*, 1(1):107–129, 2007.

[FS96] Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *Proc Int Conf Mach Learn*, pages 148–156. ACM, 1996.

[G$^+$04a] D. Geman et al. Classifying gene expression profiles from pairwise mRNA comparisons. *Stat Appl Genet Mol Biol*, 3(19), 2004.

[G$^+$04b] J.J. Goeman et al. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

[G$^+$05] Z. Guo et al. Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, 6(1):58, 2005.

[GBKV10] E. Glaab, A. Baudot, N. Krasnogor, and A. Valencia. TopoGSA: network topological gene set analysis. *Bioinformatics*, 26(9):1271–1272, 2010.

[GGK09] E. Glaab, J.M. Garibaldi, and N. Krasnogor. ArrayMining: a modular web-application for microarray analysis combining ensemble and consensus methods with cross-study normalization. *BMC Bioinformatics*, 10(1):358, 2009.

[GHT07] Y. Guo, T. Hastie, and R. Tibshirani. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, 8(1):86–100, 2007.

[H$^+$01] A.W. Hsing et al. Prostate cancer risk and serum levels of insulin and leptin: a population-based study. *J Natl Cancer Inst*, 93(10):783–789, 2001.

[Hub94] C. J. Huberty. *Applied Discriminant Analysis*. John Wiley, New York, 1994.

[HvHS^{+}02] W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(1):96–104, 2002.

[JUA05] T. Jirapech-Umpai and S. Aitken. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics*, 6(1):148, 2005.

[LB03] B. Lustig and J. Behrens. The Wnt signaling pathway and its role in tumor development. *J Cancer Res Clin Oncol*, 129(4):199–221, 2003.

[LGGV08] J.L. Lustgarten, V. Gopalakrishnan, H. Grover, and S. Visweswaran. Improving Classification Performance with Discretization on Biomedical Datasets. In *AMIA Annu Symp Proc*, volume 2008, pages 445–449. AMIA, 2008.

[LHC10] W.J. Lin, H.M. Hsueh, and J.J. Chen. Power and sample size estimation in microarray studies. *BMC Bioinformatics*, 11(1):48, 2010.

[OKM09] O. Obajimi, J.C. Keen, and P.W. Melera. Inhibition of de novo purine synthesis in human prostate cells results in ATP depletion, AMPK activation and induces senescence. *The Prostate*, 69(11):1206–1221, 2009.

[QERR03] Y.W. Qiang, Y. Endo, J.S. Rubin, and S. Rudikoff. Wnt signaling in B-cell neoplasia. *Oncogene*, 22(10):1536–1545, 2003.

[S^{+}02a] M.A. Shipp et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat Med*, 8(1):68–74, 2002.

[S^{+}02b] D. Singh et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, 2002.

[S^{+}05] A. Subramanian et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci*, 102(43):15545–15550, 2005.

[SK03] P. Stattin and R. Kaaks. Prostate cancer, insulin, and androgen deprivation therapy. *Br J Cancer*, 89(9):1814–1815, 2003.

[Tea10] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010.

[TNX^{+}05] A.C. Tan, D.Q. Naiman, L. Xu, R.L. Winslow, and D. Geman. Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21(20):3896–3904, 2005.

[VHKNW09] J. Van Hulse, T.M. Khoshgoftaar, A. Napolitano, and R. Wald. Feature Selection with High-Dimensional Imbalanced Data. In *2009 IEEE International Conference on Data Mining Workshops*, pages 507–514. IEEE, 2009.

[WEB05] P. Warnat, R. Eils, and B. Brors. Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. *BMC Bioinformatics*, 6(1):265, 2005.

[WI05] Z. Wu and R.A. Irizarry. Stochastic Models Inspired by Hybridization Theory for Short Oligonucleotide Arrays. *J Comput Biol*, 12(6):882–893, 2005.

Index of authors

GI-Edition Lecture Notes in Informatics

P-1 Gregor Engels, Andreas Oberweis, Albert Zündorf (Hrsg.): Modellierung 2001.

P-2 Mikhail Godlevsky, Heinrich C. Mayr (Hrsg.): Information Systems Technology and its Applications, ISTA'2001.

P-3 Ana M. Moreno, Reind P. van de Riet (Hrsg.): Applications of Natural Language to Information Systems, NLDB'2001.

P-4 H. Wörn, J. Mühling, C. Vahl, H.-P. Meinzer (Hrsg.): Rechner- und sensorgestützte Chirurgie; Workshop des SFB 414.

P-5 Andy Schürr (Hg.): OMER – Object-Oriented Modeling of Embedded Real-Time Systems.

P-6 Hans-Jürgen Appelrath, Rolf Beyer, Uwe Marquardt, Heinrich C. Mayr, Claudia Steinberger (Hrsg.): Unternehmen Hochschule, UH'2001.

P-7 Andy Evans, Robert France, Ana Moreira, Bernhard Rumpe (Hrsg.): Practical UML-Based Rigorous Development Methods – Countering or Integrating the extremists, pUML'2001.

P-8 Reinhard Keil-Slawik, Johannes Magenheim (Hrsg.): Informatikunterricht und Medienbildung, INFOS'2001.

P-9 Jan von Knop, Wilhelm Haverkamp (Hrsg.): Innovative Anwendungen in Kommunikationsnetzen, 15. DFN Arbeitstagung.

P-10 Mirjam Minor, Steffen Staab (Hrsg.): 1st German Workshop on Experience Management: Sharing Experiences about the Sharing Experience.

P-11 Michael Weber, Frank Kargl (Hrsg.): Mobile Ad-Hoc Netzwerke, WMAN 2002.

P-12 Martin Glinz, Günther Müller-Luschnat (Hrsg.): Modellierung 2002.

P-13 Jan von Knop, Peter Schirmbacher and Viljan Mahni_ (Hrsg.): The Changing Universities – The Role of Technology.

P-14 Robert Tolksdorf, Rainer Eckstein (Hrsg.): XML-Technologien für das Semantic Web – XSW 2002.

P-15 Hans-Bernd Bludau, Andreas Koop (Hrsg.): Mobile Computing in Medicine.

P-16 J. Felix Hampe, Gerhard Schwabe (Hrsg.): Mobile and Collaborative Busi-ness 2002.

P-17 Jan von Knop, Wilhelm Haverkamp (Hrsg.): Zukunft der Netze –Die Verletzbarkeit meistern, 16. DFN Arbeitstagung.

P-18 Elmar J. Sinz, Markus Plaha (Hrsg.): Modellierung betrieblicher Informationssysteme – MobIS 2002.

P-19 Sigrid Schubert, Bernd Reusch, Norbert Jesse (Hrsg.): Informatik bewegt – Informatik 2002 – 32. Jahrestagung der Gesellschaft für Informatik e.V. (GI) 30.Sept.-3.Okt. 2002 in Dortmund.

P-20 Sigrid Schubert, Bernd Reusch, Norbert Jesse (Hrsg.): Informatik bewegt – Informatik 2002 – 32. Jahrestagung der Gesellschaft für Informatik e.V. (GI) 30.Sept.-3.Okt. 2002 in Dortmund (Ergänzungsband).

P-21 Jörg Desel, Mathias Weske (Hrsg.): Promise 2002: Prozessorientierte Methoden und Werkzeuge für die Entwicklung von Informationssystemen.

P-22 Sigrid Schubert, Johannes Magenheim, Peter Hubwieser, Torsten Brinda (Hrsg.): Forschungsbeiträge zur "Didaktik der Informatik" – Theorie, Praxis, Evaluation.

P-23 Thorsten Spitta, Jens Borchers, Harry M. Sneed (Hrsg.): Software Management 2002 – Fortschritt durch Beständigkeit

P-24 Rainer Eckstein, Robert Tolksdorf (Hrsg.): XMIDX 2003 – XML-Technologien für Middleware – Middleware für XML-Anwendungen

P-25 Key Pousttchi, Klaus Turowski (Hrsg.): Mobile Commerce – Anwendungen und Perspektiven – 3. Workshop Mobile Commerce, Universität Augsburg, 04.02.2003

P-26 Gerhard Weikum, Harald Schöning, Erhard Rahm (Hrsg.): BTW 2003: Datenbanksysteme für Business, Technologie und Web

P-27 Michael Kroll, Hans-Gerd Lipinski, Kay Melzer (Hrsg.): Mobiles Computing in der Medizin

P-28 Ulrich Reimer, Andreas Abecker, Steffen Staab, Gerd Stumme (Hrsg.): WM 2003: Professionelles Wissensmanagement – Erfahrungen und Visionen

P-29 Antje Düsterhöft, Bernhard Thalheim (Eds.): NLDB'2003: Natural Language Processing and Information Systems

P-30 Mikhail Godlevsky, Stephen Liddle, Heinrich C. Mayr (Eds.): Information Systems Technology and its Applications

P-31 Arslan Brömme, Christoph Busch (Eds.): BIOSIG 2003: Biometrics and Electronic Signatures

P-32 Peter Hubwieser (Hrsg.): Informatische Fachkonzepte im Unterricht – INFOS 2003

P-33 Andreas Geyer-Schulz, Alfred Taudes (Hrsg.): Informationswirtschaft: Ein Sektor mit Zukunft

P-34 Klaus Dittrich, Wolfgang König, Andreas Oberweis, Kai Rannenberg, Wolfgang Wahlster (Hrsg.): Informatik 2003 – Innovative Informatikanwendungen (Band 1)

P-35 Klaus Dittrich, Wolfgang König, Andreas Oberweis, Kai Rannenberg, Wolfgang Wahlster (Hrsg.): Informatik 2003 – Innovative Informatikanwendungen (Band 2)

P-36 Rüdiger Grimm, Hubert B. Keller, Kai Rannenberg (Hrsg.): Informatik 2003 – Mit Sicherheit Informatik

P-37 Arndt Bode, Jörg Desel, Sabine Rathmayer, Martin Wessner (Hrsg.): DeLFI 2003: e-Learning Fachtagung Informatik

P-38 E.J. Sinz, M. Plaha, P. Neckel (Hrsg.): Modellierung betrieblicher Informationssysteme – MobIS 2003

P-39 Jens Nedon, Sandra Frings, Oliver Göbel (Hrsg.): IT-Incident Management & IT-Forensics – IMF 2003

P-40 Michael Rebstock (Hrsg.): Modellierung betrieblicher Informationssysteme – MobIS 2004

P-41 Uwe Brinkschulte, Jürgen Becker, Dietmar Fey, Karl-Erwin Großpietsch, Christian Hochberger, Erik Maehle, Thomas Runkler (Edts.): ARCS 2004 – Organic and Pervasive Computing

P-42 Key Pousttchi, Klaus Turowski (Hrsg.): Mobile Economy – Transaktionen und Prozesse, Anwendungen und Dienste

P-43 Birgitta König-Ries, Michael Klein, Philipp Obreiter (Hrsg.): Persistance, Scalability, Transactions – Database Mechanisms for Mobile Applications

P-44 Jan von Knop, Wilhelm Haverkamp, Eike Jessen (Hrsg.): Security, E-Learning. E-Services

P-45 Bernhard Rumpe, Wofgang Hesse (Hrsg.): Modellierung 2004

P-46 Ulrich Flegel, Michael Meier (Hrsg.): Detection of Intrusions of Malware & Vulnerability Assessment

P-47 Alexander Prosser, Robert Krimmer (Hrsg.): Electronic Voting in Europe – Technology, Law, Politics and Society

P-48 Anatoly Doroshenko, Terry Halpin, Stephen W. Liddle, Heinrich C. Mayr (Hrsg.): Information Systems Technology and its Applications

P-49 G. Schiefer, P. Wagner, M. Morgenstern, U. Rickert (Hrsg.): Integration und Datensicherheit – Anforderungen, Konflikte und Perspektiven

P-50 Peter Dadam, Manfred Reichert (Hrsg.): INFORMATIK 2004 – Informatik verbindet (Band 1) Beiträge der 34. Jahrestagung der Gesellschaft für Informatik e.V. (GI), 20.-24. September 2004 in Ulm

P-51 Peter Dadam, Manfred Reichert (Hrsg.): INFORMATIK 2004 – Informatik verbindet (Band 2) Beiträge der 34. Jahrestagung der Gesellschaft für Informatik e.V. (GI), 20.-24. September 2004 in Ulm

P-52 Gregor Engels, Silke Seehusen (Hrsg.): DELFI 2004 – Tagungsband der 2. e-Learning Fachtagung Informatik

P-53 Robert Giegerich, Jens Stoye (Hrsg.): German Conference on Bioinformatics – GCB 2004

P-54 Jens Borchers, Ralf Kneuper (Hrsg.): Softwaremanagement 2004 – Outsourcing und Integration

P-55 Jan von Knop, Wilhelm Haverkamp, Eike Jessen (Hrsg.): E-Science und Grid Ad-hoc-Netze Medienintegration

P-56 Fernand Feltz, Andreas Oberweis, Benoit Otjacques (Hrsg.): EMISA 2004 – Informationssysteme im E-Business und E-Government

P-57 Klaus Turowski (Hrsg.): Architekturen, Komponenten, Anwendungen

P-58 Sami Beydeda, Volker Gruhn, Johannes Mayer, Ralf Reussner, Franz Schweiggert (Hrsg.): Testing of Component-Based Systems and Software Quality

P-59 J. Felix Hampe, Franz Lehner, Key Pousttchi, Kai Ranneberg, Klaus Turowski (Hrsg.): Mobile Business – Processes, Platforms, Payments

P-60 Steffen Friedrich (Hrsg.): Unterrichtskonzepte für inforrmatische Bildung

P-61 Paul Müller, Reinhard Gotzhein, Jens B. Schmitt (Hrsg.): Kommunikation in verteilten Systemen

P-62 Federrath, Hannes (Hrsg.): „Sicherheit 2005“ – Sicherheit – Schutz und Zuverlässigkeit

P-63 Roland Kaschek, Heinrich C. Mayr, Stephen Liddle (Hrsg.): Information Systems – Technology and ist Applications

P-64 Peter Liggesmeyer, Klaus Pohl, Michael Goedicke (Hrsg.): Software Engineering 2005

P-65 Gottfried Vossen, Frank Leymann, Peter Lockemann, Wolffried Stucky (Hrsg.): Datenbanksysteme in Business, Technologie und Web

P-66 Jörg M. Haake, Ulrike Lucke, Djamshid Tavangarian (Hrsg.): DeLFI 2005: 3. deutsche e-Learning Fachtagung Informatik

P-67 Armin B. Cremers, Rainer Manthey, Peter Martini, Volker Steinhage (Hrsg.): INFORMATIK 2005 – Informatik LIVE (Band 1)

P-68 Armin B. Cremers, Rainer Manthey, Peter Martini, Volker Steinhage (Hrsg.): INFORMATIK 2005 – Informatik LIVE (Band 2)

P-69 Robert Hirschfeld, Ryszard Kowalcyk, Andreas Polze, Matthias Weske (Hrsg.): NODe 2005, GSEM 2005

P-70 Klaus Turowski, Johannes-Maria Zaha (Hrsg.): Component-oriented Enterprise Application (COAE 2005)

P-71 Andrew Torda, Stefan Kurz, Matthias Rarey (Hrsg.): German Conference on Bioinformatics 2005

P-72 Klaus P. Jantke, Klaus-Peter Fähnrich, Wolfgang S. Wittig (Hrsg.): Marktplatz Internet: Von e-Learning bis e-Payment

P-73 Jan von Knop, Wilhelm Haverkamp, Eike Jessen (Hrsg.): “Heute schon das Morgen sehen“

P-74 Christopher Wolf, Stefan Lucks, Po-Wah Yau (Hrsg.): WEWoRC 2005 – Western European Workshop on Research in Cryptology

P-75 Jörg Desel, Ulrich Frank (Hrsg.): Enterprise Modelling and Information Systems Architecture

P-76 Thomas Kirste, Birgitta König-Riess, Key Pousttchi, Klaus Turowski (Hrsg.): Mobile Informationssysteme – Potentiale, Hindernisse, Einsatz

P-77 Jana Dittmann (Hrsg.): SICHERHEIT 2006

P-78 K.-O. Wenkel, P. Wagner, M. Morgenstern, K. Luzi, P. Eisermann (Hrsg.): Land- und Ernährungswirtschaft im Wandel

P-79 Bettina Biel, Matthias Book, Volker Gruhn (Hrsg.): Softwareengineering 2006

P-80 Mareike Schoop, Christian Huemer, Michael Rebstock, Martin Bichler (Hrsg.): Service-Oriented Electronic Commerce

P-81 Wolfgang Karl, Jürgen Becker, Karl-Erwin Großpietsch, Christian Hochberger, Erik Maehle (Hrsg.): ARCS´06

P-82 Heinrich C. Mayr, Ruth Breu (Hrsg.): Modellierung 2006

P-83 Daniel Huson, Oliver Kohlbacher, Andrei Lupas, Kay Nieselt and Andreas Zell (eds.): German Conference on Bioinformatics

P-84 Dimitris Karagiannis, Heinrich C. Mayr, (Hrsg.): Information Systems Technology and its Applications

P-85 Witold Abramowicz, Heinrich C. Mayr, (Hrsg.): Business Information Systems

P-86 Robert Krimmer (Ed.): Electronic Voting 2006

P-87 Max Mühlhäuser, Guido Rößling, Ralf Steinmetz (Hrsg.): DELFI 2006: 4. e-Learning Fachtagung Informatik

P-88 Robert Hirschfeld, Andreas Polze, Ryszard Kowalczyk (Hrsg.): NODe 2006, GSEM 2006

P-90 Joachim Schelp, Robert Winter, Ulrich Frank, Bodo Rieger, Klaus Turowski (Hrsg.): Integration, Informationslogistik und Architektur

P-91 Henrik Stormer, Andreas Meier, Michael Schumacher (Eds.): European Conference on eHealth 2006

P-92 Fernand Feltz, Benoît Otjacques, Andreas Oberweis, Nicolas Poussing (Eds.): AIM 2006

P-93 Christian Hochberger, Rüdiger Liskowsky (Eds.): INFORMATIK 2006 – Informatik für Menschen, Band 1

P-94 Christian Hochberger, Rüdiger Liskowsky (Eds.): INFORMATIK 2006 – Informatik für Menschen, Band 2

P-95 Matthias Weske, Markus Nüttgens (Eds.): EMISA 2005: Methoden, Konzepte und Technologien für die Entwicklung von dienstbasierten Informationssystemen

P-96 Saartje Brockmans, Jürgen Jung, York Sure (Eds.): Meta-Modelling and Ontologies

P-97 Oliver Göbel, Dirk Schadt, Sandra Frings, Hardo Hase, Detlef Günther, Jens Nedon (Eds.): IT-Incident Mangament & IT-Forensics – IMF 2006

P-98 Hans Brandt-Pook, Werner Simonsmeier und Thorsten Spitta (Hrsg.): Beratung in der Softwareentwicklung – Modelle, Methoden, Best Practices

P-99 Andreas Schwill, Carsten Schulte, Marco Thomas (Hrsg.): Didaktik der Informatik

P-100 Peter Forbrig, Günter Siegel, Markus Schneider (Hrsg.): HDI 2006: Hochschuldidaktik der Informatik

P-101 Stefan Böttinger, Ludwig Theuvsen, Susanne Rank, Marlies Morgenstern (Hrsg.): Agrarinformatik im Spannungsfeld zwischen Regionalisierung und globalen Wertschöpfungsketten

P-102 Otto Spaniol (Eds.): Mobile Services and Personalized Environments

P-103 Alfons Kemper, Harald Schöning, Thomas Rose, Matthias Jarke, Thomas Seidl, Christoph Quix, Christoph Brochhaus (Hrsg.): Datenbanksysteme in Business, Technologie und Web (BTW 2007)

P-104 Birgitta König-Ries, Franz Lehner, Rainer Malaka, Can Türker (Hrsg.)
MMS 2007: Mobilität und mobile Informationssysteme

P-105 Wolf-Gideon Bleek, Jörg Raasch, Heinz Züllighoven (Hrsg.)
Software Engineering 2007

P-106 Wolf-Gideon Bleek, Henning Schwentner, Heinz Züllighoven (Hrsg.)
Software Engineering 2007 – Beiträge zu den Workshops

P-107 Heinrich C. Mayr, Dimitris Karagiannis (eds.)
Information Systems
Technology and its Applications

P-108 Arslan Brömme, Christoph Busch, Detlef Hühnlein (eds.)
BIOSIG 2007:
Biometrics and Electronic Signatures

P-109 Rainer Koschke, Otthein Herzog, Karl-Heinz Rödiger, Marc Ronthaler (Hrsg.)
INFORMATIK 2007
Informatik trifft Logistik
Band 1

P-110 Rainer Koschke, Otthein Herzog, Karl-Heinz Rödiger, Marc Ronthaler (Hrsg.)
INFORMATIK 2007
Informatik trifft Logistik
Band 2

P-111 Christian Eibl, Johannes Magenheim, Sigrid Schubert, Martin Wessner (Hrsg.)
DeLFI 2007:
5. e-Learning Fachtagung Informatik

P-112 Sigrid Schubert (Hrsg.)
Didaktik der Informatik in Theorie und Praxis

P-113 Sören Auer, Christian Bizer, Claudia Müller, Anna V. Zhdanova (Eds.)
The Social Semantic Web 2007
Proceedings of the 1st Conference on Social Semantic Web (CSSW)

P-114 Sandra Frings, Oliver Göbel, Detlef Günther, Hardo G. Hase, Jens Nedon, Dirk Schadt, Arslan Brömme (Eds.)
IMF2007 IT-incident management & IT-forensics
Proceedings of the 3rd International Conference on IT-Incident Management & IT-Forensics

P-115 Claudia Falter, Alexander Schliep, Joachim Selbig, Martin Vingron and Dirk Walther (Eds.)
German conference on bioinformatics
GCB 2007

P-116 Witold Abramowicz, Leszek Maciszek (Eds.)
Business Process and Services Computing
1st International Working Conference on Business Process and Services Computing
BPSC 2007

P-117 Ryszard Kowalczyk (Ed.)
Grid service engineering and manegement
The 4th International Conference on Grid Service Engineering and Management
GSEM 2007

P-118 Andreas Hein, Wilfried Thoben, Hans-Jürgen Appelrath, Peter Jensch (Eds.)
European Conference on ehealth 2007

P-119 Manfred Reichert, Stefan Strecker, Klaus Turowski (Eds.)
Enterprise Modelling and Information Systems Architectures
Concepts and Applications

P-120 Adam Pawlak, Kurt Sandkuhl, Wojciech Cholewa, Leandro Soares Indrusiak (Eds.)
Coordination of Collaborative Engineering - State of the Art and Future Challenges

P-121 Korbinian Herrmann, Bernd Bruegge (Hrsg.)
Software Engineering 2008
Fachtagung des GI-Fachbereichs Softwaretechnik

P-122 Walid Maalej, Bernd Bruegge (Hrsg.)
Software Engineering 2008 - Workshopband
Fachtagung des GI-Fachbereichs Softwaretechnik

P-123 Michael H. Breitner, Martin Breunig, Elgar Fleisch, Ley Pousttchi, Klaus Turowski (Hrsg.)
Mobile und Ubiquitäre Informationssysteme – Technologien, Prozesse, Marktfähigkeit
Proceedings zur 3. Konferenz Mobile und Ubiquitäre Informationssysteme (MMS 2008)

P-124 Wolfgang E. Nagel, Rolf Hoffmann, Andreas Koch (Eds.)
9th Workshop on Parallel Systems and Algorithms (PASA)
Workshop of the GI/ITG Speciel Interest Groups PARS and PARVA

P-125 Rolf A.E. Müller, Hans-H. Sundermeier, Ludwig Theuvsen, Stephanie Schütze, Marlies Morgenstern (Hrsg.)
Unternehmens-IT:
Führungsinstrument oder Verwaltungsbürde
Referate der 28. GIL Jahrestagung

P-126 Rainer Gimnich, Uwe Kaiser, Jochen Quante, Andreas Winter (Hrsg.)
10th Workshop Software Reengineering (WSR 2008)

P-127 Thomas Kühne, Wolfgang Reisig, Friedrich Steimann (Hrsg.)
Modellierung 2008

P-128 Ammar Alkassar, Jörg Siekmann (Hrsg.)
Sicherheit 2008
Sicherheit, Schutz und Zuverlässigkeit
Beiträge der 4. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI)
2.-4. April 2008
Saarbrücken, Germany

P-129 Wolfgang Hesse, Andreas Oberweis (Eds.)
Sigsand-Europe 2008
Proceedings of the Third AIS SIGSAND European Symposium on Analysis, Design, Use and Societal Impact of Information Systems

P-130 Paul Müller, Bernhard Neumair, Gabi Dreo Rodosek (Hrsg.)
1. DFN-Forum Kommunikationstechnologien Beiträge der Fachtagung

P-131 Robert Krimmer, Rüdiger Grimm (Eds.)
3rd International Conference on Electronic Voting 2008
Co-organized by Council of Europe, Gesellschaft für Informatik and E-Voting.CC

P-132 Silke Seehusen, Ulrike Lucke, Stefan Fischer (Hrsg.)
DeLFI 2008:
Die 6. e-Learning Fachtagung Informatik

P-133 Heinz-Gerd Hegering, Axel Lehmann, Hans Jürgen Ohlbach, Christian Scheideler (Hrsg.)
INFORMATIK 2008
Beherrschbare Systeme – dank Informatik
Band 1

P-134 Heinz-Gerd Hegering, Axel Lehmann, Hans Jürgen Ohlbach, Christian Scheideler (Hrsg.)
INFORMATIK 2008
Beherrschbare Systeme – dank Informatik
Band 2

P-135 Torsten Brinda, Michael Fothe, Peter Hubwieser, Kirsten Schlüter (Hrsg.)
Didaktik der Informatik –
Aktuelle Forschungsergebnisse

P-136 Andreas Beyer, Michael Schroeder (Eds.)
German Conference on Bioinformatics
GCB 2008

P-137 Arslan Brömme, Christoph Busch, Detlef Hühnlein (Eds.)
BIOSIG 2008: Biometrics and Electronic Signatures

P-138 Barbara Dinter, Robert Winter, Peter Chamoni, Norbert Gronau, Klaus Turowski (Hrsg.)
Synergien durch Integration und Informationslogistik
Proceedings zur DW2008

P-139 Georg Herzwurm, Martin Mikusz (Hrsg.)
Industrialisierung des Software-Managements
Fachtagung des GI-Fachausschusses Management der Anwendungsentwicklung und -wartung im Fachbereich Wirtschaftsinformatik

P-140 Oliver Göbel, Sandra Frings, Detlef Günther, Jens Nedon, Dirk Schadt (Eds.)
IMF 2008 - IT Incident Management & IT Forensics

P-141 Peter Loos, Markus Nüttgens, Klaus Turowski, Dirk Werth (Hrsg.)
Modellierung betrieblicher Informationssysteme (MobIS 2008)
Modellierung zwischen SOA und Compliance Management

P-142 R. Bill, P. Korduan, L. Theuvsen, M. Morgenstern (Hrsg.)
Anforderungen an die Agrarinformatik durch Globalisierung und Klimaveränderung

P-143 Peter Liggesmeyer, Gregor Engels, Jürgen Münch, Jörg Dörr, Norman Riegel (Hrsg.)
Software Engineering 2009
Fachtagung des GI-Fachbereichs Softwaretechnik

P-144 Johann-Christoph Freytag, Thomas Ruf,
Wolfgang Lehner, Gottfried Vossen
(Hrsg.)
Datenbanksysteme in Business,
Technologie und Web (BTW)

P-145 Knut Hinkelmann, Holger Wache (Eds.)
WM2009: 5th Conference on Professional
Knowledge Management

P-146 Markus Bick, Martin Breunig,
Hagen Höpfner (Hrsg.)
Mobile und Ubiquitäre
Informationssysteme – Entwicklung,
Implementierung und Anwendung
4. Konferenz Mobile und Ubiquitäre
Informationssysteme (MMS 2009)

P-147 Witold Abramowicz, Leszek Maciaszek,
Ryszard Kowalczyk, Andreas Speck (Eds.)
Business Process, Services Computing
and Intelligent Service Management
BPSC 2009 · ISM 2009 · YRW-MBP 2009

P-148 Christian Erfurth, Gerald Eichler,
Volkmar Schau (Eds.)
9th International Conference on Innovative
Internet Community Systems
I^{2}CS 2009

P-149 Paul Müller, Bernhard Neumair,
Gabi Dreo Rodosek (Hrsg.)
2. DFN-Forum
Kommunikationstechnologien
Beiträge der Fachtagung

P-150 Jürgen Münch, Peter Liggesmeyer (Hrsg.)
Software Engineering
2009 - Workshopband

P-151 Armin Heinzl, Peter Dadam, Stefan Kirn,
Peter Lockemann (Eds.)
PRIMIUM
Process Innovation for
Enterprise Software

P-152 Jan Mendling, Stefanie Rinderle-Ma,
Werner Esswein (Eds.)
Enterprise Modelling and Information
Systems Architectures
Proceedings of the 3rd Int‘l Workshop
EMISA 2009

P-153 Andreas Schwill,
Nicolas Apostolopoulos (Hrsg.)
Lernen im Digitalen Zeitalter
DeLFI 2009 – Die 7. E-Learning
Fachtagung Informatik

P-154 Stefan Fischer, Erik Maehle
Rüdiger Reischuk (Hrsg.)
INFORMATIK 2009
Im Focus das Leben

P-155 Arslan Brömme, Christoph Busch,
Detlef Hühnlein (Eds.)
BIOSIG 2009:
Biometrics and Electronic Signatures
Proceedings of the Special Interest Group
on Biometrics and Electronic Signatures

P-156 Bernhard Koerber (Hrsg.)
Zukunft braucht Herkunft
25 Jahre »INFOS – Informatik und
Schule«

P-157 Ivo Grosse, Steffen Neumann,
Stefan Posch, Falk Schreiber,
Peter Stadler (Eds.)
German Conference on Bioinformatics
2009

P-158 W. Claupein, L. Theuvsen, A. Kämpf,
M. Morgenstern (Hrsg.)
Precision Agriculture
Reloaded – Informationsgestützte
Landwirtschaft

P-159 Gregor Engels, Markus Luckey,
Wilhelm Schäfer (Hrsg.)
Software Engineering 2010

P-160 Gregor Engels, Markus Luckey,
Alexander Pretschner, Ralf Reussner
(Hrsg.)
Software Engineering 2010 –
Workshopband
(inkl. Doktorandensymposium)

P-161 Gregor Engels, Dimitris Karagiannis
Heinrich C. Mayr (Hrsg.)
Modellierung 2010

P-162 Maria A. Wimmer, Uwe Brinkhoff,
Siegfried Kaiser, Dagmar Lück-Schneider, Erich Schweighofer,
Andreas Wiebe (Hrsg.)
Vernetzte IT für einen effektiven Staat
Gemeinsame Fachtagung
Verwaltungsinformatik (FTVI) und
Fachtagung Rechtsinformatik (FTRI) 2010

P-163 Markus Bick, Stefan Eulgem,
Elgar Fleisch, J. Felix Hampe,
Birgitta König-Ries, Franz Lehner,
Key Pousttchi, Kai Rannenberg (Hrsg.)
Mobile und Ubiquitäre
Informationssysteme
Technologien, Anwendungen und Dienste
zur Unterstützung von mobiler
Kollaboration

P-164 Arslan Brömme, Christoph Busch (Eds.)
BIOSIG 2010: Biometrics and Electronic
Signatures Proceedings of the Special
Interest Group on Biometrics and
Electronic Signatures

P-165 Gerald Eichler, Peter Kropf,
Ulrike Lechner, Phayung Meesad,
Herwig Unger (Eds.)
10th International Conference on
Innovative Internet Community Systems
(I^{2}CS) – Jubilee Edition 2010 –

P-166 Paul Müller, Bernhard Neumair,
Gabi Dreo Rodosek (Hrsg.)
3. DFN-Forum Kommunikationstechnologien
Beiträge der Fachtagung

P-167 Robert Krimmer, Rüdiger Grimm (Eds.)
4th International Conference on
Electronic Voting 2010
co-organized by the Council of Europe,
Gesellschaft für Informatik and
E-Voting.CC

P-168 Ira Diethelm, Christina Dörge,
Claudia Hildebrandt,
Carsten Schulte (Hrsg.)
Didaktik der Informatik
Möglichkeiten empirischer
Forschungsmethoden und Perspektiven
der Fachdidaktik

P-169 Michael Kerres, Nadine Ojstersek
Ulrik Schroeder, Ulrich Hoppe (Hrsg.)
DeLFI 2010 - 8. Tagung
der Fachgruppe E-Learning
der Gesellschaft für Informatik e.V.

P-170 Felix C. Freiling (Hrsg.)
Sicherheit 2010
Sicherheit, Schutz und Zuverlässigkeit

P-171 Werner Esswein, Klaus Turowski,
Martin Juhrisch (Hrsg.)
Modellierung betrieblicher
Informationssysteme (MobIS 2010)
Modellgestütztes Management

P-173 Dietmar Schomburg,
Andreas Grote (Eds.)
German Conference on Bioinformatics
2010

P-174 Arslan Brömme, Torsten Eymann,
Detlef Hühnlein, Heiko Roßnagel,
Paul Schmücker (Hrsg.)
perspeGKtive 2010
Workshop „Innovative und sichere
Informationstechnologie für das
Gesundheitswesen von morgen“

P-175 Klaus-Peter Fähnrich,
Bogdan Franczyk (Hrsg.)
INFORMATIK 2010
Service Science – Neue Perspektiven für
die Informatik
Band 1

P-176 Klaus-Peter Fähnrich,
Bogdan Franczyk (Hrsg.)
INFORMATIK 2010
Service Science – Neue Perspektiven für
die Informatik
Band 2

P-177 Witold Abramowicz, Rainer Alt,
Klaus-Peter Fähnrich, Bogdan Franczyk,
Leszek A. Maciaszek (Eds.)
INFORMATIK 2010
Business Process and Service Science –
Proceedings of ISSS and BPSC

The titles can be purchased at:

Köllen Druck + Verlag GmbH
Ernst-Robert-Curtius-Str. 14 · D-53117 Bonn
Fax: +49 (0)228/9898222
E-Mail: druckverlag@koellen.de