GCB 2013 Göttingen - Highlight Papers

Contents

Exploiting structural information for target assessment

Andrea Volkamer and Matthias Rarey
Center for Bioinformatics, University of Hamburg, Bundesstrae 43, 20146 Hamburg, Germany

volkamer@zbh.uni-hamburg.de

Abstract: The amount of solved protein structures is continuously growing. Pharmaceutical research recently recognized the potential of computationally extracting information from this large data pool and using them for homology-based knowledge transfer to new structures. This article focuses on computational approaches for structure-based target assessment. Highlighted are novel approaches for target classification, i.e., druggability or function prediction, and target comparison together with the underlying methods for active site detection and description. Protein function predictions, e.g., yielded accuracies between 54% and 81% for predicting the correct main class, subclass and substrate-specific subsubclass on a test set of 26632 pockets. Besides the presentation of successful retrospective application studies of these methods, challenging tasks in the individual computational steps are discussed in this article.

1 Introduction

Drug discovery is a cost and time consuming venture, thus, computational approaches have long entered the early drug development pipeline. While the classical computer-aided application is screening of large compound data sets for new lead compounds, recent advances in structure elucidation and structural genomic projects, enabled high-throughput approaches for target screening, e.g., target prioritization, characterization and comparison. Learning from what is known, extracting patterns and transferring them to novel targets is one major goal in modern structure-based computational approaches. In this article, we present our recent developments addressing target assessment. Comparing proteins on a functional level means comparing their centers of action. Starting from the protein structure, our in house software DoGSite [VGGR10] can be applied for automated active site detection and representation by numerical descriptors. For classification scenarios, the descriptors are incorporated

into a support vector machine (SVM) to learn from the features of known member of specific classes. In this context, SVMs have been trained for target prioritization [VKG+12] and function annotation [VKRR13]. Another important task is the direct comparison of proteins by specific active site features. In the novel approach TrixP [BVH+13], active sites are represented by pharmacophoric triangles and closest homologous can be identified with high-throughput by screening a pre-calculated index consisting of active sites with known classification.

2 Methods

DoGSite [VGGR10] detects potential pockets solely based on the 3D structure of a protein. For this purpose, a grid representation of the protein is used. Grid points are labeled as free or occupied, dependent on their coverage by a protein atom. Subsequently, a Difference of Gaussian (DoG) filter is applied to find positions on the protein surface where the location of small sphere-like objects is favorable. Grid points are clustered into subpockets based on the calculated DoG value and neighboring subpockets are assigned to pockets. The **DoGSiteScorer** [VKG+12, VKRR13] is a generic approach for structure-based classification scenarios. The program automatically calculates numerical descriptors for self-predicted pockets. These global descriptors include size, shape and physicochemical properties of the pockets, e.g. volume, surface, ellipsoidal shape, enclosure, hydrophilic surface fraction, functional groups, element and amino acid compositions. Given a training data set with annotated classes, a discriminant analysis is used to select those descriptors which separate best between the different input target classes. Eventually, a support vector machine is trained on a property of interest related to binding or function. **TrixP** [BVH+13] is a novel method enabling fast index-based binding site comparisons. Recognition features are encoded via a triangular descriptor, holding physicochemical and shape information of the binding site [SR09]. Triangles are spanned between all present hydrogen bond donor, acceptor, and apolar point triplets and the shape of the binding site is captured by an 80-ray bulk, placed at the respective triangle's center. For efficient screening applications, an index with known binding site descriptors can be built and unlimitedly queried. For a new query protein, descriptors are calculated, the database is screened, and targets with matching descriptors are returned, superimposed and ranked by their estimated similarity to the query.

3 Results and Discussion

In the following, the results of the presented methods and their contributions to overcome the difficulties within the processed tasks are summarized. The pocket prediction method, **DoGSite**, was evaluated on the PDBBind and scPDB data set, containing 828 and 6754 structures, respectively, and detected the true ligand binding site in over 92% of the test cases. The major challenge in correct pocket and boundary detections arises from the nature of pockets. In contrast to the crystallized representatives, protein structures are flexible which produces a magnitude of potential pocket shapes from being small to large, shallow to deep, and homogeneous to highly branched. In this context, DoGSite especially convinced by its novel granular subpocket detection and a globular pocket ceiling definition. **DoGSiteScorer** was trained and evaluated for two different classification scenarios, namely druggability [VKG$^+$12] and function prediction [VKRR13]. The one prerequisite, when working with machine learning techniques, is a large and reliable data set to train the method on. Unfortunately, most available data sets are either too small or suffer from wrong or miss-annotated structures. For target prioritization, the method was trained on a subset of the recently published DD data set containing 1069 druggable, difficult and undruggable protein pockets and yielded accuracies of 88% in the testing phase. Next, to allow for enzymatic function predictions on different granularity levels, we created a data set of over 26000 enzymatic pockets and classified them with respect to the enzymatic classification (EC) scheme. Subsequently, models for predicting enzyme class, subclass or substrate-specificity based on structural features were build. Cross-validation studies showed accuracies of 68% for correct main class prediction and accuracies between 63% and 81% for the six subclasses. Substrate-specific recall rates for a kinase subset were 54%. Finally, our active site comparison tool **TrixP** was evaluated on several screening scenarios based on the scPDB data set. Using a subset of 769 similar and dissimilar protein pairs, a similarity cut-off was introduced with which similar pairs could be recovered in 82% of the cases, while dissimilar pairs were discarded with 99.5%. Screening the complete data set with four query proteins, 84-100% of the index-contained family members could be identified. Even correct subfamilies could be assigned for a small kinase data set. Again, flexibility upon ligand binding challenges structure-based comparison methods. The consideration of partial similarities based on matching triangle descriptors in combination with introduced tolerance values, allows TrixP to recover similarities between partially different conformations.

4 Conclusion

Due to the continuous growth in elucidated structures, learning from known structures has become more and more important, increasing the demand of fast and reliable computational approaches for target assessment. A set of software approaches has been presented which can be used for active site detection, target druggability and function prediction as well as target comparison. The methods have been evaluated on large data sets and showed good results in retrospective applications. A major drawback of automated methods is generally the dependence on the quality of the structures, the size of the available training data, the reliability of the annotated classes of their members as well as the necessary homology to securely transfer a specific property. Nevertheless, such methods allow to perform high-throughput target screening and, thus, to think outside the box, and detect similarities and cross-links between structures that would not have been found by hand.

References

[BVH+13] M. v. Behren, A. Volkamer, A. M. Henzler, K. T. Schomburg, S. Urbaczek, and R. Rarey. Fast binding site comparison via an index-based screening technology. *Journal of Chemical Information and Modeling*, 53:411–422, 2013.

[SR09] J. Schlosser and M. Rarey. Beyond the Virtual Screening Paradigm: Structure-Based Searching for New Lead Compounds. *Journal of Chemical Information and Modeling*, 49(4):800–809, 2009.

[VGGR10] A. Volkamer, A. Griewel, T. Grombacher, and M. Rarey. Analyzing the topology of active sites: on the prediction of pockets and subpockets. *Journal of Chemical Information and Modeling*, 50(11):2041–2052, 2010.

[VKG+12] A. Volkamer, D. Kuhn, T. Grombacher, F. Rippmann, and M. Rarey. Combining global and local measures for structure-based druggability predictions. *Journal of Chemical Information and Modeling*, 52(2):360–372, 2012.

[VKRR13] A. Volkamer, D. Kuhn, F. Rippmann, and M. Rarey. Predicting enzymatic function from global binsing site descriptors. *Proteins: Structure, Function and Bioinformatics*, 81(3):479–489, 2013.

Versatile prioritization of candidate disease genes or other molecules with NetworkPrioritizer

Tim Kacprowski[1,2], Nadezhda T. Doncheva[1], Mario Albrecht[1,2]
[1]*Max Planck Institute for Informatics, Saarbrücken and*
[2]*University Medicine Greifswald, Greifswald, Germany*

tim.kacprowski@uni-greifswald.de

Abstract: The prioritization of candidate disease genes or other molecules is often based on heterogeneous data. Nevertheless, most prioritization methods do not allow for a straight-forward integration of the user's own input data. Therefore, we developed NetworkPrioritizer, a Cytoscape plugin that enables the integrative network-based prioritization of bio-molecules. Our versatile software tool computes a number of important centrality measures to rank nodes based on their relevance for the connectivity in weighted and unweighted networks. As further novelty, it provides different methods to aggregate and compare rankings. NetworkPrioritizer and its documentation are freely available at http://www.networkprioritizer.de

1 Introduction

To elucidate the genetic foundations of human diseases, it is crucial to identify genes that might predispose to or cause specific diseases. Computational prioritization methods exploit the available biomedical knowledge to rank candidate genes according to their disease relevance. Many methods integrate multiple data sources, e.g., gene expression, protein interactions, and overlapping disease characteristics [DoKA12]. Integrated information of biological relationships and interactions is often represented as network. The connections between known disease genes and the remaining genes in a network are of particular interest as they can point to new disease genes.

The majority of prioritization methods are available only as web services [TCNM10]. Since the latter require the upload of the user's input data, they are not well suited to analyze confidential data. Furthermore, most web services rely on pre-defined background data and do not allow the user to include own data, to control the data integration, or to modify the aggregation of multiple rankings [DoKA12, TCNM10]. Existing Cytoscape plugins for

prioritization are also subject to a number of limitations. For instance, cytoHubba [LCWC08] and GPEC [LeKw12] rank network nodes focused on their close neighborhood or the steady-state probability of a random walk with restart, respectively. Neither supports further analysis or aggregation of rankings. The plugin NetworkAnalyzer [ARSL08] and the Java application CentiBiN [JuKS06] feature a large set of centrality measures, but cannot compute them for a user-defined set of seed nodes or for weighted networks.

Here, we present NetworkPrioritizer [KaDA13], a novel Cytoscape plugin for the integrative network-based prioritization of candidate genes or other molecules. It comprises two main functionalities. First, it facilitates the estimation of the relevance of network nodes, e.g., candidate genes, with regard to a set of seed nodes, e.g., known disease genes. Second, the plugin allows for the user-guided aggregation and comparison of multiple rankings.

2 Software Features

NetworkPrioritizer ranks nodes based on their relevance for the network connectivity, estimated by a number of centrality measures based on shortest paths and random walks (see web site). Closeness quantifies the distance of a node to the rest of the network. Betweenness measures the influence of a node on the paths connecting other nodes. These measures interpret all networks as undirected. NetworkPrioritizer can handle unweighted and weighted networks with user-adjustable effect of the edge weights on the computed centralities [OpAS10]. A particular feature of NetworkPrioritizer is the computation of the centrality measures with regard to a set of seed nodes.

In contrast to other prioritization tools, NetworkPrioritizer offers multiple methods to aggregate and compare multiple primary rankings. Weighted Borda Fuse (WBF) is a generalization of the popular Borda count [Saar99] and essentially ranks nodes according to their weighted mean rank in the primary rankings. Weighted AddScore Fuse (WASF) calculates the weighted sum of scores for each node in the primary rankings and awards a higher rank the larger this sum is. WBF and WASF can be used to identify candidate genes that attain high ranks in all primary rankings. If the primary rankings are based on scores on the same scale, WASF is more distinctive and thus more accurate than WBF. MaxRank Fuse (MRF) assigns each node the highest rank achieved in any primary ranking. MRF can quickly identify candidates with a high rank in at least one primary ranking. Furthermore, NetworkPrioritizer provides two widely used measures of ranking distance, the Spearman footrule and the Kendall tau [DKNS01]. The Spearman footrule is the sum of the rank differences of all nodes in two compared rankings. The Kendall tau is the number of nodes with different ranks.

3 Benchmark and Case Study

We evaluated the performance of NetworkPrioritizer in a comprehensive benchmark on data described in [ScLA10]. Briefly, this data contains artificial quantitative trait loci for 99 diseases extracted from OMIM [HSAB05] that have at least three known associated disease proteins. Disease-specific protein-protein interaction networks were extracted from BioMyn [RaLA12] and complemented by functional similarity links inferred from FunSimMat [ScLA10]. The receiver operating characteristics (ROCs) computed in a leave-one-out cross-validation are shown in Fig. 1. The shortest path-based centrality measures performed best, reaching areas under the ROC curve (AUCs) of up to 0.90. Notably, aggregating the primary rankings further improved the performance and resulted in AUCs from 0.87 to 0.92 (superior to the AUC of 0.85 achieved by MedSim on the same data).

As a case study, we applied NetworkPrioritizer to a Crohn's Disease (CD) related protein network. Protein-protein interactions and functional similarity links were compiled from BioMyn and FunSimMat, respectively, for proteins encoded by genes in CD-associated loci [FMBW10]. Proteins associated with inflammatory bowel disease (IBD), or CD as a subtype of IBD, were used as seed nodes (see web site). The 10 top-ranked proteins (HLA-B, SMAD3, CCL2, NOTCH1, STAT5A/B, HLA-A2/26/66, BECN1) function in the 'immune system process', 'response to stress', 'signal transduction', and 'homeostatic process' according to their Gene Ontology annotation. Since these processes are closely related to IBD [ZhLi12], the proteins are promising candidates for further experimental studies.

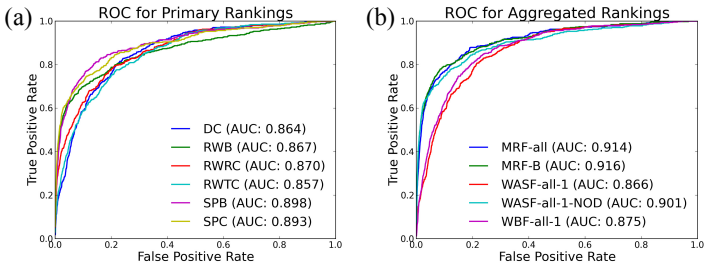


Fig. 1: (a) ROC for degree centrality (DC), random walk betweenness (RWB), random walk receiver closeness (RWRC), random walk transmitter closeness (RWRC), shortest path betweenness (SPB), and shortest path closeness (SPB). **(b)** ROC for the aggregation of all primary rankings using MRF (MRF-all), the aggregation of SPB and RWB (MRF-B), the aggregation of all rankings (equally weighted) using WASF (WASF-all-1), the aggregation of all rankings but DC (equally weighted) using WASF (WAS-all-1-NOD), and the aggregation of all rankings (equally weighted) using WBF (WBF-all-1).

4 Conclusions

NetworkPrioritizer enables the ranking of individual network nodes based on their relevance for connecting a set of seed nodes to the rest of the network. The Cytoscape plugin computes centrality measures for unweighted and weighted networks and, as a particular novelty, provides rank aggregation methods and ranking distance calculations. Its versatility makes NetworkPrioritizer a very useful tool for integrative network-based prioritization of candidate disease genes and proteins or other molecules.

References

[ARSL08] Yassen Assenov, Fidel Ramírez, Sven-Eric Schelhorn, et al. Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–4, 2008

[DKNS01] Cynthia Dwork, Ravi Kumar, Moni Naor, et al. Rank aggregation methods for the web. *Proceedings of the 10th International Conference on World Wide Web - WWW*, 613–22, 2001

[DoKA12] Nadezhda T Doncheva, Tim Kacprowski and Mario Albrecht. Recent approaches to the prioritization of candidate disease genes. *Wiley Interdiscip Rev Syst Biol Med*, 4(5):429–42, 2012

[FMBW10] Andre Franke, Dermot PB McGovern, Jeffrey C Barrett, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat Genet*, 42(12):1118–25, 2010

[HSAB05] Ada Hamosh, Alan F Scott, Joanna S Amberger, et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33:D514–7, 2005

[JuKS06] Björn H Junker, Dirk Koschützki and Falk Schreiber. Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics*, 7:1–7, 2006

[KaDA13] Tim Kacprowski, Nadezhda T Doncheva and Mario Albrecht. NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. *Bioinformatics*, 29(11):1471–3, 2013

[LCWC08] Chung-Yen Lin, Chia-Hao Chin, Hsin-Hung Wu, et al. Hubba: hub objects analyzer – a framework of interactome hubs identification for network biology. *Nucleic Acids Res*, 36:W438–43, 2008

[LeKw12] Duc-Hau Le and Yung-keun Kwon. GPEC: a Cytoscape plug-in for random walk-based gene prioritization and biomedical evidence collection. *Comput Biol Chem*, 37:17–23, 2012

[OpAS10] Tore Opsahl, Filip Agneessens and John Skvoretz. Node centrality in weighted networks: generalizing degree and shortest paths. *Soc Networks*, 32(3):245–51, 2010

[RaLA12] Fidel Ramírez, Glenn Lawyer and Mario Albrecht. Novel search method for the discovery of functional relationships. *Bioinformatics*, 28(2):269–76, 2012

[Saar99] Donald G Saari. Explaining all three-alternative voting outcomes. *J Econ Theory*, 87(2):313–55, 1999

[ScLA10] Andreas Schlicker, Thomas Lengauer and Mario Albrecht. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. *Bioinformatics*, 26(18):i561–7, 2010

[TCNM10] Léon-Charles Tranchevent, Francisco Bonachela Capdevila, Daniela Nitsch, et al. A guide to web tools to prioritize candidate genes. *Brief Bioinform*, 12(1):22–32, 2010

[ZhLi12] Hong Zhu and Y Robert Li. Oxidative stress and redox signaling mechanisms of inflammatory bowel disease: updated experimental and clinical evidence. *Exp Biol Med*, 237(5):474–80, 2012

Learning Gene Network Structure from Time Laps Cell Imaging in RNAi Knock-Downs

Henrik Failmezger [1+] and Paurush Praveen [2+] and Achim Tresch [1] and Holger Fröhlich [2]

[1] *Computational Biology and Regulatory Networks, Max-Planck Institute for Plant Breeding Research, Carl-von-Linne-Weg 10, 50829 Cologne, Germany*

[2] *Algorithmic Bioinformatics, Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2, 53113 Bonn, Germany*

frohlich@bit.uni-bonn.de

[+] *These authors contributed equally.*

Abstract: As RNA interference is becoming a standard method for targeted gene perturbation, computational approaches to reverse engineer parts of biological networks based on measure-able effects of RNAi become increasingly relevant. The vast majority of these methods use gene expression data, but little attention has been paid so far to other data types. Here we present a method, which can infer gene networks from high-dimensional phenotypic perturbation effects on single cells recorded by time-lapse microscopy. We use data from the Mitocheck project to extract multiple shape, intensity and texture features at each frame. Features from different cells and movies are then aligned along the cell cycle time. Subsequently we employ Dynamic Nested Effects Models (dynoNEMs) to estimate parts of the network structure between perturbed genes via a Markov Chain Monte Carlo approach. Our simulation results indicate a high reconstruction quality of this method. A reconstruction based on a 22 gene knock-downs yielded a network, where all edges could be explained via the biological literature. The implementation of dynoNEMs is part of the Bioconductor R-package *nem*.

1 Introduction

The availability of large RNAi screens has raised the interest in computational approaches for reverse engineering parts of biological networks from measure-able effects of targeted gene perturbations. Most existing methods make use of gene expression data. The few attempts to reverse engineer gene networks from phenotypic data include [BACP07], who rely on hierarchical clustering of

static images, and [KDZ+09] who use a probabilistic graphical model for only one binary phenotypic variable in static images. To our knowledge, there is yet no method for the inference of networks from time lapse microscopy based on large numbers of statistical image features.

Nested Effects Models (NEMs) are a class of probabilistic graphical models that have been introduced originally by [MBS05] and extended substantially later on by several other authors. In NEMs indirect, high-dimensional down-stream effects of multiple single-gene knock-downs are studied. NEMs allow for inferring the signaling flow between these perturbed genes on a transcriptional as well as non-transcriptional level based on the measured intervention effects. [ASJ+09] and [FPT11] extended the theory of NEMs to time series data, and applied it to infer parts of a transcriptional network involved in murine stem cell development. Originally NEMs assumed downstream effects to be measured via gene expression profiling, but here we use phenotypic image features from movies instead. Our movies were taken from the Mitocheck database [NWH+10], in which ~20,000 human genes were silenced via RNAi and subsequently screened for cell cycle defects. We use Dynamic Nested Effects Models [FPT11] to estimate the network between perturbed genes based on the dynamic response of the phenotype along the cell cycle. The inference is based on a Markov Chain Monte Carlo (MCMC) algorithm. The whole approach consisting of image feature extraction, estimation of perturbation effects and network estimation via dynoNEMs is called MovieNEM and described in detail in our paper [FPTF13]. A schematic overview about our method is shown in Figure 1.

2 Main Results

2.1 Simulations

In our paper we conducted extensive simulations on reconstruction of randomly selected sub-graphs of KEGG signaling pathways using simulated phenotypic features. In general we observed a very high accuracy of our method, which was notably dependent on the network size and network topology. Very densely connected networks with many loops appeared to be harder to learn than acyclic graphs. This can be explained by the fact that with many loops it is more unlikely to observe a time delayed nested effects structure, which is exploited by our method. We also investigated the influence of uninformative features on our method. Here we observed a highly robust behavior of our method, which underlines the success of the automated feature selection mechanism, which is inbuilt in MovieNEM.

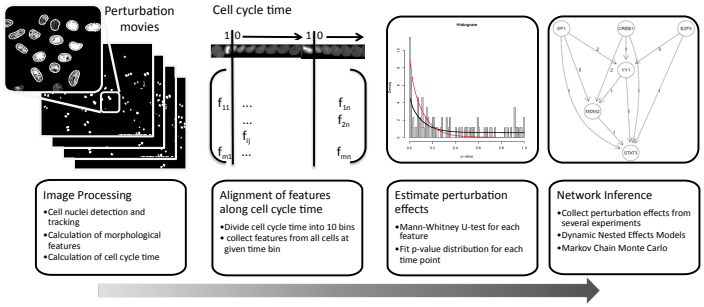


Figure 1: Overview about MovieNEM: Individual movies are first fed into an image processing pipeline consisting of four steps: (1) cell nuclei detection in the individual movie frames; (2) tracking of the nuclei over time; (3) calculation of morphological features and (4) calculation of cell cycle time. After image processing features are grouped according to the binned cell cycle time. This allows for estimating time-wise perturbation effects. Several movies, each showing one perturbation, are processed in this way and the perturbation likelihoods collected along the binned cell cycle time axis. This allows for applying Dynamic Nested Effects Models to infer the network between perturbed genes via Markov Chain Monte Carlo.

2.2 Application to Movie Data

In our paper we applied MovieNEM to infer a network between 22 genes with significant phenotype. These 22 genes are mainly involved into cell cycle, transcriptional regulation and cell differentiation. We looked at the network of edge-wise posterior expectations, which scored better than 1000 random S-gene permutations. All 122 edges could be mapped to known literature pathways. On the other hand and not very surprisingly the literature mentions some additional interactions, which could not be observed in our estimated network. This can have two reasons: Either the additional literature known interactions exist in reality, but MovieNEM could not infer them or they do not exist in HeLa cells and are hence not inferred. Notably, we can only infer interactions between genes that show a clear phenotypic knock-down effect.

3 Conclusion

In our paper we have shown that it is possible to learn pathway structures from phenotypic perturbation effects recorded in time-lapse movies. At the heart of

the method lies the extraction of morphological features yielding measurable differences in cell phenotypes. We have developed a method to quantify these differences such that an extended version of the dynoNEM method [FPT11] is applicable. We have developed a novel Markov Chain Monte Carlo sampler for network structure learning in order estimate the posterior likelihood of each interaction. Our method allows for the inclusion of prior knowledge in a Bayesian fashion. In summary MovieNEM offers an approach to exploit the rich information that is present in phenotypic RNAi screening data using image based techniques.

References

[ASJ+09] Benedict Anchang, Mohammad J Sadeh, Juby Jacob, Achim Tresch, Marcel O Vlad, Peter J Oefner, and Rainer Spang. Modeling the temporal interplay of molecular signaling and gene expression by using dynamic nested effects models. *Proc Natl Acad Sci U S A*, 106(16):6447–6452, Apr 2009.

[BACP07] Chris Bakal, John Aach, George Church, and Norbert Perrimon. Quantitative morphological signatures define local signaling networks regulating cell morphology. *Science*, 316(5832):1753–1756, Jun 2007.

[FPT11] H. Fröhlich, P. Praveen, and A. Tresch. Fast and Efficient Dynamic Nested Effects Models. *Bioinformatics*, 27(2):238–244, 2011.

[FPTF13] Henrik Failmezger, Paurush Praveen, Achim Tresch, and Holger Fröhlich. Learning gene network structure from time laps cell imaging in RNAi Knock downs. *Bioinformatics*, pages 1534 – 1540, May 2013.

[KDZ+09] Lars Kaderali, Eva Dazert, Ulf Zeuge, Michael Frese, and Ralf Bartenschlager. Reconstructing signaling pathways from RNAi data using probabilistic Boolean threshold networks. *Bioinformatics*, 25(17):2229–2235, Sep 2009.

[MBS05] Florian Markowetz, Jacques Bloch, and Rainer Spang. Non-transcriptional pathway features reconstructed from secondary effects of RNA interference. *Bioinformatics*, 21(21):4026–4032, Nov 2005.

[NWH+10] Beate Neumann, Thomas Walter, Jean-Karim Heriche, Jutta Bulkescher, Holger Erfle, Christian Conrad, Phill Rogers, Ina Poser, Michael Held, Urban Liebel, Cihan Cetin, Frank Sieckmann, Gregoire Pau, Rolf Kabbe, Annelie Wünsche, Venkata Satagopam, Michael H A. Schmitz, Catherine Chapuis, Daniel W. Gerlich, Reinhard Schneider, Roland Eils, Wolfgang Huber, Jan-Michael Peters, Anthony A. Hyman, Richard Durbin, Rainer Pepperkok, and Jan Ellenberg. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. *Nature*, 464(7289):721–727, Apr 2010.

GenomeTools: a comprehensive software library for efficient processing of structured genome annotations

Gordon Gremme, Sascha Steinbiss and Stefan Kurtz
Center for Bioinformatics, University of Hamburg

kurtz@zbh.uni-hamburg.de

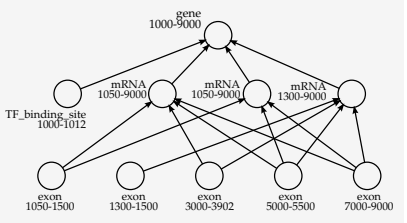
Abstract: Annotations of genomic features and their subcomponents can be conveniently and intuitively described by annotation graphs, a representation also serving as the basis for common text formats like GFF3. However, current bioinformatics toolkits do not make use of the full expressiveness of such a representation. We present the *GenomeTools*, an efficient software library allowing for convenient development of new software tools which create or process (e.g. augment) annotation graphs. The *GenomeTools* API is modeled around the annotation graph concept, making it easy to access the information contained in the annotation and to design graph-based algorithms based on them. The object-oriented *GenomeTools* library is optimized to keep a small memory footprint for large annotation sets (such as variation annotations like SNVs) by careful data structure design and by implementing an efficient pull-based approach for sequential processing of annotations. It also provides bindings to a variety of script programming languages (like Python, Lua and Ruby) sharing a common programming interface.

1 Introduction

Genomic annotations connect raw sequence information to the associated structural and functional properties, such as gene location, gene structure, and transcript variety. In the scope of bioinformatics software, they can act as both output or input. For instance, in a gene prediction tool, the locations of each detected gene, transcript, and their exons are typical annotation output. Moreover, repeat instances like transposon insertions, tRNA genes, and even regulatory regions like transcription factor binding sites are common constituents of genomic annotations output by specific bioinformatic software tools. As input, annotations are important – for example when integrated with experimental data – as the basis for hypothesis generation aided by software tools (e.g. custom genome browsers). In some cases, the more fine-grained structure of the genomic features' com-

```
##gff-version   3
##sequence-region    chr1 1 1497228
 chr1 . gene 1000 9000 . + . ID=gene1;Name=EDEN
 chr1 . TF_binding_site 1000 1012 . + . Parent=gene1
 chr1 . mRNA 1050 9000 . + . ID=mRNA1;Parent=gene1
 chr1 . mRNA 1050 9000 . + . ID=mRNA2;Parent=gene1
 chr1 . exon 1050 1500 . + . Parent=mRNA1,mRNA2
 chr1 . mRNA 1300 9000 . + . ID=mRNA3;Parent=gene1
 chr1 . exon 1300 1500 . + . Parent=mRNA3
 chr1 . exon 3000 3902 . + . Parent=mRNA1,mRNA3
 chr1 . exon 5000 5500 . + . Parent=mRNA1,mRNA2,mRNA3
 chr1 . exon 7000 9000 . + . Parent=mRNA1,mRNA2,mRNA3
###
```

(a)



(b)

Figure 1: (a) Part of a GFF3 file describing a gene with three possible alternatively spliced transcripts. (b) Corresponding annotation graph consisting of a single connected component.

ponents is important as well, for example when utilizing information about gene and transcript structure to perform a census of alternative splicing events [ELM+05]).

To establish a standard model for structured genomic annotations, Eilbeck et al. [ELM+05] introduced the concept of *annotation graphs* as a generic representation of genomic annotations for prokaryotic and eukaryotic genomes. Annotation graphs are directed acyclic graphs (DAGs) in which nodes represent genomic features and edges represent *part-of* relationships between them. Nodes are typed according to the Sequence Ontology (SO), a standardized set of terms and relationships describing genomic entities [ELM+05]. For example, in the gene example mentioned above, *mRNA* nodes can be connected to *gene* nodes to express that the transcripts are parts (or *subfeatures*) of the gene (which has no parent and hence is the *top-level* node). Nodes of the *exon* type are in turn connected to *mRNA* nodes as the exons are constituents of the mRNA feature (Fig. 1b). Due to the DAG structure, exons can also belong to multiple transcripts. *Gene* nodes do not have a parent. Hence each gene is represented as a connected component (CC, for short) in the annotation graph. SO-compliant annotations are typically given as plain text in the *Generic Feature Format*, Version 3 (GFF3), which basically describes the annotation graph by tagging nodes with plain text attributes specifying the *part-of* relations between parent and child nodes in the graph (Fig. 1a).

We have identified several key requirements to be satisfied by a software toolkit for annotation processing in order to make full use of the information contained within such structured annotation files. Besides (obviously) capturing the graph structure of annotation graphs using appropriate data

structures and access methods, the software should not restrict the user to a specific subset of features (e.g. genes) but support all SO terms instead. Due to the large size of annotations, a space efficient representation is also important. This encompasses efficient handling of common sequential processing operations and non-redundant storage of repetitive data such as sequence identifiers. Another requirement is a simple yet flexible and extensible application programming interface (API) for accessing annotations. It should also support accessing the sequence, a genomic feature refers to, a task complicated by the fact that the sequence is often stored separately from the annotation and not labeled with a unique or standardized identifier. Finally, flexible and validating parsers are required for the most common annotation formats.

Generic software satisfying these desired features is scarce. Previous popular programming environments for bioinformatics show widely varying levels of support for handling genomic annotations and none of them has all the required features.

2 Implementation

The *GenomeTools* toolkit, which in contrast has all the required features, uses an object-oriented approach to represent nodes in the annotation graph as individual implementations of different classes with a common *genome node* interface, modeled in accordance with the GFF3 specification. Each node contains the genomic location (position, chromosome, etc.) of the feature it represents and additional attributes, given as key-value pairs.

The annotation graph can be partitioned into weakly connected components based on the connectivity of its underlying undirected graph. Using the *GenomeTools* API, CCs can be traversed using iterators and modified by changing attributes or adding new child nodes. All actions performed on nodes are implemented as *node streams*. Streams are active program components which either create nodes, modify them or output them. The basic approach is to sequentially pass a set of CCs (accessed through their top-level nodes) through a stream of chains, possibly applying modifications before passing a CC to the successor stream. This approach makes use of lazy evaluation and is very memory-efficient if input data is appropriately sorted.

As a software development kit, the *GenomeTools* are available as a shared

library which provides interface headers to implement custom streams, allowing for interoperability between native *GenomeTools* streams and custom ones. To access the *GenomeTools* functionality from scripting languages, we have created bindings for the languages Java, Ruby, Python and Lua using foreign function interfaces acting as a thin wrapper layer around the library, allowing to write streams in other languages than C.

Input and output streams for various formats (GFF3, GVF, GTF, BED) are available. We have taken special care to handle boundary cases which may occur in GFF3 input to finally ensure that parsed graphs are always correct – even when the input does not fully comply to the GFF3 specification. For instance, GFF3 allows features with the same ID attribute spanning multiple feature lines. Such a *multi-feature* implicitly specifies parent nodes. For each part of a multi-feature, a separate node in the feature DAG is introduced and tagged with a special multi-feature flag. Furthermore, one of the nodes comprising the multi-feature is distinguished as a *representative*. Since no explicit top-level node is present for these multi-feature nodes, we introduce an artificial *pseudo-feature* as a new unique top-level feature node. All features comprising the multi-feature become the children of a new pseudo-feature, to guarantee that each CC has a top-level node.

The *GenomeTools* provide mechanisms for persistent storage of annotation and indexed random access to features overlapping query regions. These are useful to, for example, develop genome browsers or similar software visualizing genome annotations. Finally, the *GenomeTools* library provides techniques for an efficient sequence representation which can be combined with the annotations [SK12]. It also provides a large variety of useful sequence analysis functionality (index construction and access, annotation visualization, and much more) as well as a collection of *tools*, which make use of the library to solve real-world bioinformatics tasks. These tools have been published separately[1].

3 Results

We have applied the *GenomeTools*, BioPerl and SeqAn C++ toolkits to parse a variety of annotation examples (gene annotations up to several GB in size, SNV annotations, repeat annotations) into their own representa-

[1] See `http://genometools.org` or the full *GenomeTools* paper [GSK13] for a complete list of associated publications.

tion and measured the time and space requirement. The *GenomeTools* library was consistently both the fastest and most memory-efficient of the three toolkits, while also being the only one with the most complete support for annotation graphs. For example, for processing the TAIR *A. thaliana* annotation [RBB+03] the SeqAn (BioPerl) representation required up to 11 times (34 times) more space than the *GenomeTools* representation. Regarding running times, the *GenomeTools* library was able to create full annotation graphs from the input data in a matter of seconds, while the competitors required up to several minutes.

4 Conclusion

We have developed a library for efficient handling of structured genomic annotations which retains the expressiveness of the annotation graph approach, thus allowing a developer to implement new algorithms very close to the intuitive theoretical concept. A simple concept of defining a processing pipeline using the the stream and visitor patterns facilitates easy interoperability between individual processing components. Tools built using the *GenomeTools* library require less memory and are faster than previous toolkits. We expect the *GenomeTools* software to continue being a basis for new software tools for an ever increasing number of sequence analysis tasks. The full paper [GSK13] was recently published in its preliminary form.

References

[ELM+05] K. Eilbeck, S. Lewis, C. Mungall, M. Yandell, L. Stein, R. Durbin, and M. Ashburner. The Sequence Ontology: A Tool for the Unification of Genome Annotations. *Genome Biology*, 6(5):R44, 2005.

[GSK13] G. Gremme, S. Steinbiss, and S. Kurtz. *GenomeTools*: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, PrePrints, accepted on May 29, 2013.

[RBB+03] S.Y. Rhee, W. Beavis, T.Z. Berardini, G. Chen, D. Dixon, et al. The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Research*, 31(1):224–228, 2003.

[SK12] S. Steinbiss and S. Kurtz. A New Efficient Data Structure for Storage and Retrieval of Multiple Biosequences. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(2):345–357, 2012.

TALEs of virulence and biotechnology

Jan Grau[1], Annett Wolf[1], Maik Reschke[2], Ulla Bonas[2], Stefan Posch[1], and Jens Boch[2]
[1]*Institute of Computer Science, Martin Luther University Halle–Wittenberg*
[2]*Department of Genetics, Institute of Biology, Martin Luther University Halle–Wittenberg*

grau@informatik.uni-halle.de

Abstract: Transcription activator-like effectors (TALEs) are injected into host plant cells by *Xanthomonas* pathogens to function as transcriptional activators. Their DNA-binding domain is composed of conserved amino acid repeats containing repeat-variable diresidues (RVDs) that determine DNA binding specificity.

We present TALgetter, a new approach for predicting TALE target sites based on a statistical model. The predictions of TALgetter indicate a previously unreported positional preference of TALE target sites relative to the transcription start site. In addition, several TALEs are predicted to bind to the TATA-box, which might constitute one general mode of transcriptional activation by TALEs.

1 Introduction

Transcription activator-like effectors (TALEs) are injected into the cells of host plants by plant-pathogenic *Xanthomonas* bacteria, where they act as transcription factors for the benefit of the pathogen [B$^+$09]. The hosts of different *Xanthomonas* strains span a variety of important crop plants including rice, sweet orange, tomato, pepper, and cabbage.

The DNA-binding domain of TALEs is composed of highly conserved tandem repeats, where each repeat usually spans 34 amino acids. These repeats bind to the nucleotides of a DNA target site in a contiguous, non-overlapping fashion. The DNA-specificity of an individual repeat depends on the two amino acids at position 12 and 13, termed repeat-variable diresidues (RVDs) [B$^+$09, MB09].

The computational prediction of virulence targets of natural TALEs is a key step to provide candidates for subsequent experimental validation. However, less than 30 virulence targets have been validated, often only one for an individual TALE. This set is complemented by a few hundred

artificial target sites from reporter assays. Hence, novel approaches are required that i) make use of the available data in a holistic manner, and ii) allow for predicting target sites of TALEs without any known target site. In [GWR+13], we propose such an approach for predicting TALE target sites called *TALgetter*.

2 Computational predictions provide insights into the biology of TALE target sites

In [GWR+13], we propose TALgetter, which uses a new statistical model representing *importance* of RVDs and their *binding specificity* independently in a local mixture model. The concept of importance is related to the *efficiency* of RVDs [S+12], but additionally affects the penalty for non-matching nucleotides. In the proposed model, the importance and binding specificity of an RVD are independent of its context in the TALE. For details of the statistical model, we refer to [GWR+13].

In contrast to previous approaches, the parameters of this model are estimated computationally, where different TALEs and their known target sites can be combined in a common training set due to independence assumptions. TALgetter is part of version 2.1 of the open-source Java library Jstacs [G+12].

In [GWR+13], we show that TALgetter yields an improved prediction performance compared to the existing approach, *Target Finder* of the TALE-NT suite [D+12]. Using TALgetter, we predict target sites of *Xanthomonas* TALEs in the important crop plants rice and sweet orange. These predictions elucidate novel putative virulence targets of several TALEs (c.f. Tables 6 to 8 of [GWR+13]).

In addition, we demonstrate that computational approaches are able to gain new insights into the biology of TALE targeting. Specifically, we combine predictions of TALgetter with gene expression data to identify functional TALE target sites. We find that functional target sites are preferentially located in a region from 300 bp upstream to 200 bp downstream of the transcription start (c.f. Figure 7 of [GWR+13]). Our predictions also indicate that many TALEs bind to the TATA-box in the promoters of their target genes. Based on these observations, we propose four biological models (c.f. Figure S7 of [GWR+13]) that may explain the apparent target site preference of TALEs.

The modular architecture of TALEs allows for a rearrangement of repeats

to easily generate any desired DNA-specificity. Hence, TALEs have become a preferred biotechnology tool for targeted DNA binding. Although TALgetter has been created for predicting virulence targets of natural TALEs, it is readily applicable to biotechnology problems as, for instance, the off-target prediction of artificial TALE activators.

TALEs are also the basis of TALE nucleases (TALENs), which have been established as a second genome-editing technique besides zinc-finger nucleases [GGB13]. In TALENs, the DNA-binding domain of TALEs is fused with a Fok1 endonuclease domain, where homo- or hetero-dimers of TALENs specifically cut the DNA double strand. Although TALENs cut DNA highly specific, undesired *off-targets* in addition to the targeted genomic region remain an important issue [O$^+$13]. Recently [GBP], we developed a novel tool for the genome-wide prediction of TALEN off-targets, named *TALENoffer*. TALENoffer is based on the same statistical model as TALgetter, and features an optimized runtime to scan complete genomes for TALEN off-target sites within a few minutes.

3 Availability

Web-applications of TALgetter and TALENoffer are available at `http://galaxy.informatik.uni-halle.de`, and can also be installed in a local Galaxy [B$^+$10] server. In addition, we provide command line version of TALgetter and TALENoffer at `http://jstacs.de/index.php/TALgetter` and `http://jstacs.de/index.php/TALENoffer`, respectively. TALgetter also allows users to estimate new model parameters from custom training data. Hence, users can adapt the parameters of the TALgetter model to improved sets of validated TALE target sites, which are to be expected in the near future.

4 Talk outline

We start our talk with an introduction to TALEs and the specific bioinformatics problems that arise in the prediction of TALE target sites. After a brief description of the statistical model of TALgetter, we focus on the biological findings that have been discovered using our computational predictions, namely the previously unreported target site preferences of TALEs and biological models explaining these. We finally succinctly in-

troduce TALENoffer for predicting off-target sites of TALE nucleases.

References

[B+09] Jens Boch et al. Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. *Science*, 326(5959):1509–1512, 2009.

[B+10] Daniel Blankenberg et al. *Galaxy: A Web-Based Genome Analysis Tool for Experimentalists*. John Wiley & Sons, Inc., 2010.

[D+12] Erin L. Doyle et al. TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. *Nucleic Acids Research*, 40(W1):W117–W122, 2012.

[G+12] Jan Grau et al. Jstacs: A Java Framework for Statistical Analysis and Classification of Biological Sequences. *Journal of Machine Learning Research*, 13(Jun):1967–1971, 2012.

[GBP] Jan Grau, Jens Boch, and Stefan Posch. TALENoffer: genome-wide TALEN off-target prediction. *Under review*.

[GGB13] Thomas Gaj, Charles A. Gersbach, and Carlos F. Barbas. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends in biotechnology, doi: 10.1016/j.tibtech.2013.04.004*, 05 2013.

[GWR+13] Jan Grau, Annett Wolf, Maik Reschke, Ulla Bonas, Stefan Posch, and Jens Boch. Computational Predictions Provide Insights into the Biology of TAL Effector Target Sites. *PLOS Computational Biology*, 9(3):e1002962, 03 2013.

[MB09] Matthew J. Moscou and Adam J. Bogdanove. A Simple Cipher Governs DNA Recognition by TAL Effectors. *Science*, 326(5959):1501, 2009.

[O+13] Mark J Osborn et al. TALEN-based Gene Correction for Epidermolysis Bullosa. *Molecular Therapy, doi:10.1038/mt.2013.56*, 04 2013.

[S+12] Jana Streubel et al. TAL effector RVD specificities and efficiencies. *Nature Biotechnology*, 30(7):593–595, 07 2012.

Paving the Way for Automated Clinical Breath Analysis and Biomarker Detection

Anne-Christin Hauschild[1,3], Jörg Ingo Baumbach[4], Jan Baumbach[1,3,2]

[1] *Computational Systems Biology Group, Max Planck Institute for Informatics, Saarbrücken, Germany*
[2] *Computational Biology group, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark*
[3] *Cluster of Excellence for Multimodal Computing and Interaction. Saarland University, Saarbrücken, Germany*
[4] *B & S Analytik, BioMedizinZentrum Dortmund, Germany*

a.hauschild@mpi-inf.mpg.de

Abstract: It is common knowledge that the human breath contains metabolites allowing to infer a patient's health status, especially for diseases related to the respiratory system. This information is encoded in the "volatolom", a combination of volatile organic compounds (VOCs) produced by the human metabolism and environmental perturbations. Nevertheless, due to a lack of alternative analytical techniques most of the traditional diagnostic methods are still based on invasive techniques, e.g. using blood or tissue samples.
During the last decade, the ion mobility spectrometer combined with a multi-capillary column (MCC/IMS) has become an established, inexpensive, and non-invasive bioanalytics technique for detecting VOCs with various potential applications in medical research. To pave the way for this technology towards daily usage in medical practice, different challenges still have to be solved. One of the main challenges is to establish an automated framework optimizing the processing algorithms in the pipeline yielding to an optimal performance for the final goals: Disease prediction and biomarker detection.
Although equivalent computational methods and standard procedures exist for other biomedical applications (e.g. sequence and microarray analysis) we still are lacking such a standard protocol in breath research. In four recently published papers presented here[HBJ12, HKD$^+$13, HSP$^+$12, SHBB13] we aimed at solving this challenge.

1 Introduction

Developed in the early 1970, the ion mobility spectrometers (IMS) has mainly been applied in the military domain, e.g. for detecting explosives. The resolution and precision was dramatically increased by combining the IMS with the multi capillary column for pre-separation. This technology leap opened up the horizon to many other applications. Furthermore, the IMS is inured to the moisture in exhaled air, comparably cheap, robust, and easy to use in every day practice, which makes it especially interesting for biotechnological and medical

applications [Bau09]. Due to these developments the IMS can now also be used for patient breath analysis and monitoring [WLM+11], identification of bacterial strains and fungi [BW06, PJV+11] and cancer sub-typing [WLM+11], just to name a few.

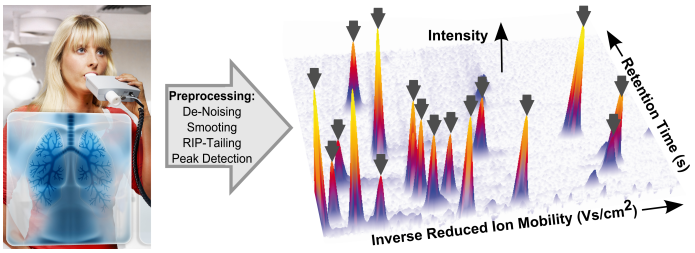


Figure 1: Result of the MCC/IMS measurement after preprocessing.

2 Computational Methods for MCC/IMS Data

Many challenges on the way to a comprehensive standard procedure in breath research have been solved in the last decade: the definition of a data format, the pre-processing (RIP detailing, smoothing, de-noising), the visualization and the evaluation using simple statistical techniques, see Figure 1. Nevertheless, to enable reliable disease prediction and biomarker detection we require an automated framework, following a standard protocol of carefully selected and adapted processing steps, see Figure 2.

Pre-processing and peak detection: With the increasing amount of performed measurements, a reliable and robust automated peak detection without manual intervention is an inevitable processing step. Although sophisticated peak detection approaches have been studied extensively in the last five to ten years, non of the proposed methods prevailed. This is most likely due to the fact that the assessment of the quality of these techniques has been done, if at all, solely by visual comparison with the manually selected peak lists[HSP+12]. Therefore, in a recent study we carefully evaluated four state-of-the-art approaches for automated MCC/IMS-based peak detection: local maxima search, watershed transformation, merged peak cluster localization, and peak model estimation (PME). We manually generated a gold standard with the aid of a domain expert (manual) and compared the performance of the automated methods with respect to two distinct criteria: (1) we systematically studied the classification performance of established machine learning methods (linear support vector machine and random forest) trained on the four peak detectors'

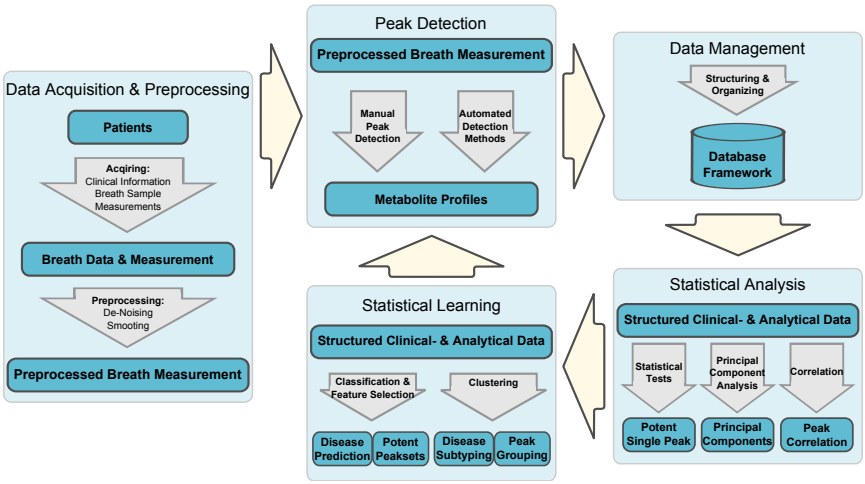


Figure 2: Automated framework for data analysis in breath research.

results and (2) we investigated the variance and robustness of the results, regarding overfitting and training set perturbations. In summary, all methods, though small differences exist, perform equally well in terms of classification error and are similarly robust against perturbations, while PME is most robust against overfitting. However, the trade off between a slightly higher accuracy (manual) and a huge increase in processing speed (automatic) has to be considered carefully [HKD+13].

Database and analysis platform: Another main part of an automated framework for MCC/IMS research is a flexible and comprehensive centralized data repository. To identify discriminating compounds as biomarkers, it is inevitable to have a clear understanding of the detailed composition of human breath and its potential confounders. Therefore, the challenge in MCC/IMS database development is not storing metabolic data in a fixed scheme but rather flexibly storing a huge set of heterogeneous clinical variables, varying for each study. To tackle this problem, we designed a comprehensive database application and analysis platform (IMSDB), combining metabolic maps with heterogeneous biomedical data in a well-structured manner. The model consists of a hybrid of the entity-attribute-value (EAV) model and the EAV-CR, incorporating the concepts of classes and relationships. Additionally, the IMSDB offers an intuitive user interface providing easy and quick access to the platform's functionalities: automated data integration and integrity validation, versioning and roll-back strategy, data retrieval as well as semi-automatic data mining

and basic machine learning capabilities. The standalone software system is based on the programming language Java and the WEKA learning package [HFH+09]. In conclusion, we are providing an intuitive software system for biologists, chemists, physicists and physicians conducting MCC/IMS-based research and diagnostics without demanding any computer skills [SHBB13].
Statistical learning: Finally, with respect to modern biomarker research and clinical diagnostics, one of the most important tasks is the automated classification of patient-specific data sets into different groups, e.g. healthy or not. In a recent pilot study, we investigated the potential of sophisticated statistical learning techniques for (1) VOC-based feature selection and (2) supervised classification on metabolic MCC/IMS data. Therefore, we utilized breath measurements of patients, either suffering from chronic obstructive pulmonary disease (COPD) or both, COPD and bronchial carcinoma (COPD+BC), as well as a healthy control group (CG). When distinguishing healthy from COPD all tested methods showed a reasonable performance between 84% (decision tree) and 94% (random forest) accuracy. However, the results indicated that further examination of the impact of bronchial carcinoma on COPD/no-COPD classification performance is necessary. The set of most important VOC features found by random forest coincided to a large extend with previous studies. Our findings demonstrate a generally high but improvable potential of statistical learning methods when applied to well-structured, medical MCC/IMS data. For more details see [HBJ12].

3 Summary

The MCC/IMS has become an established inexpensive, non-invasive bioanalytics technology for detecting VOCs with various potential applications in medical research. To pave the way for this analytical technique towards daily usage in medical practice, an automated framework following a standard protocol of carefully selected and adapted processing steps is needed. Therefore, our research focuses on the establishment of such a framework, with regard to the overall goal of disease prediction and biomarker detection. In this context, we evaluated different automated peak detection approaches and reported the peak model estimation as most robust [HKD+13]. The IMSDB combines two powerful computational tools, an extendible database as well as a machine learning toolkit, fully accessible through an intuitive graphical user interface, which will accelerate and expand the opportunities of clinical diagnostic research in the near future[SHBB13]. Finally, the application of sophisticated statistical learning methods for disease prediction and biomarker detection enables us to optimize the single steps of the proposed computational standard procedure for

breath research. The combination of the MCC/IMS methodology and the proposed automated standard procedure has the potential to successfully address a broad range of biomedical questions.

References

[Bau09] J. I. Baumbach. Ion mobility spectrometry coupled with multi-capillary columns for metabolic profiling of human breath. *Journal of Breath Research*, 3(3):1–16, 2009.

[BW06] J. I. Baumbach and M. Westhoff. Ion mobility spectrometry to detect lung cancer and airway infections. *Spectroscopy Europe*, 18(6):22–27, 2006.

[HBJ12] A.C. Hauschild, J. I. Baumbach, and Baumbach J. Integrated statistical learning of metabolic ion mobility spectrometry profiles for pulmonary disease identification. *Genet. Mol. Res.*, 11(3):2733–2744, 2012.

[HFH+09] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[HKD+13] A.C. Hauschild, D. Kopczynski, M. D'Addario, J. I. Baumbach, S. Rahmann, and J. Baumbach. Peak Detection Method Evaluation for Ion Mobility Spectrometry by Using Machine Learning Approaches. *Metabolites*, 3(2):277–293, 2013.

[HSP+12] A.C. Hauschild, T. Schneider, J. Pauling, K. Rupp, M. Jang, J. I. Baumbach, and J. Baumbach. Computational Methods for Metabolomic Data Analysis of Ion Mobility Spectrometry Data - Reviewing the State of the Art. *Metabolites*, 2(4):733–755, 2012.

[PJV+11] T. Perl, M. Juenger, W. Vautz, J. Nolte, M. Kuhns, B. Zepelin, and M. Quintel. Detection of characteristic metabolites of Aspergillus fumigatus and Candida species using ion mobility spectrometry - metabolic profiling by volatile organic compounds. *Mycoses*, 54(6):828–837, 2011.

[SHBB13] T. Schneider, A.C. Hauschild, J. I. Baumbach, and J. Baumbach. An Integrative Clinical Database and Diagnostics Platform for Biomarker Identification and Analysis in Ion Mobility Spectra of Human Exhaled Air. *Journal of Integrative Bioinformatics*, 10:733–755, 2013.

[WLM+11] M. Westhoff, P. Litterst, S. Maddula, B. Bödeker, and J. I. Baumbach. Statistical and bioinformatical methods to differentiate chronic obstructive pulmonary disease (COPD) including lung cancer from healthy control by breath analysis using ion mobility spectrometry. *International Journal for Ion Mobility Spectrometry*, pages 1–11, 2011.

Continuous rapid expansion of the mutually exclusive spliced exome in *Drosophila* species

Klas Hatje and Martin Kollmar

Abteilung NMR basierte Strukturbiologie, Max-Planck-Institut für Biophysikalische Chemie, Am Fassberg 11, D-37077 Göttingen, Germany

mako@nmr.mpibpc.mpg.de

Abstract: Mutually exclusive splicing is an important mechanism in a wide range of eukaryotic branches to expand proteome diversity but the extent of its distribution within a single species and its evolutionary conservation is unknown. Here, we present a genome-wide analysis of mutually exclusive spliced exons (MXEs) in *Drosophila melanogaster* at unprecedented depth. Most of the new MXE candidates are supported by evolutionary conservation, transcriptome data analysis and identification of competing RNA secondary structural elements. The enrichment of the genes with MXEs in transmembrane transporters and ion channel activity is consistent with findings in human, although the MXEs appeared independently and in non-homologous genes, supporting the idea of a universal benefit of adapting ion channel and receptor properties by tandem exon duplications. The comparison of the mutually exclusive spliced exomes within the Drosophila clade shows high numbers of MXE gain and loss events implicating a role of these processes in speciation.

1 Introduction

Alternative processing of primary RNA transcripts is an important driver of increased proteome diversity and regulated gene expression in eukaryotes. Alternative splicing has been reported for alveolates and stramenopiles, green algae and plants, the cryptophyte *Guillardia theta* and the chlorarachniophyte *Bigelowiella natans*, fungi and metazoa, and has therefore been an essential characteristic of the last common ancestor of the eukaryotes. The prevalence of the splice types and the overall number of events strongly differ between branches and species. Mutually exclusive splicing is a particularly interesting

type of generating alternative transcripts: The *Drosophila Down Syndrome Cell Adhesion Molecule* (*Dscam*) gene contains 95 mutually exclusive spliced exons (MXEs) representing the most extensively alternatively spliced gene known. Mutations in MXEs and regions regulating their splicing cause human diseases like the Timothy syndrome, cardiomyopathy or cancer. Mutually exclusive splicing has been shown to be regulated by competing RNA secondary structures. We reported on the continuous gain and loss of MXEs across twelve *Drosophila* species [KH13].

2 Results and Discussion

The algorithm that has been developed for the search for MXEs is based on criteria derived from biological knowledge. A) MXEs must be translated in the same reading frame and the splice sites must be compatible. B) MXEs must have about the same length, because they code for the same structural region in the resulting protein, and length differences are only possible in loop regions. C) The protein sequences coded by the MXEs are supposed to be similar, because they code for the same region in the protein and developed most probably by exon duplication during evolution. As input the software requires the exon-intron structure of the gene. Subsequently the surrounding introns of each original exon are searched for candidates for MXEs. The new algorithm is fully integrated into WebScipio [FO08], the web interface to the Scipio software. Figure 1 shows clusters of MXEs as found in the *DSCAM* gene of *Drosophila melanogaster*.

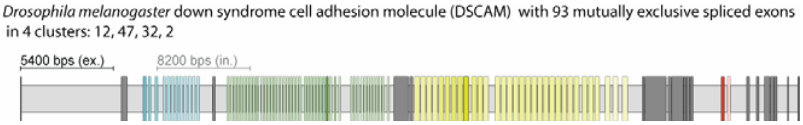


Figure 1. The *DSCAM* gene containing clusters of MXEs (coloured bars). Constitutive exons and introns are denoted by dark grey and light grey bars, respectively.

To characterize the mutually exclusive spliced exome of *Drosophila melanogaster*, we identified 1,297 exons that are mutually exclusive in annotated isoforms of the same gene. Of these 291 had similar length and sequence, including 218 internal MXEs. We predicted 539 exons of similar length and sequence that could be spliced in a mutually exclusive way (two times the annotated exons; Fig. 2). 419 of the MXE candidates were internal including 218 of the already annotated MXEs. Evidence for the predicted MXE candidates was obtained through additional data (Fig. 2): A) Mapping of EST and RNA-Seq data. B) Conservation of the MXE candidates in other

arthropods. C) Ab initio prediction of exonic regions in the respective introns using AUGUSTUS. D) Identification of competing RNA secondary structures. Of the internal MXEs 57% were supported by multiple data types, 21% were supported by EST data. Of the 44 newly predicted internal MXEs eight were supported by EST and/or RNA-Seq data. 94.5% of the annotated and reconstructed internal MXEs and 76.6% of the total predicted internal MXEs are evolutionarily conserved in at least one of the eighteen further analyzed species.

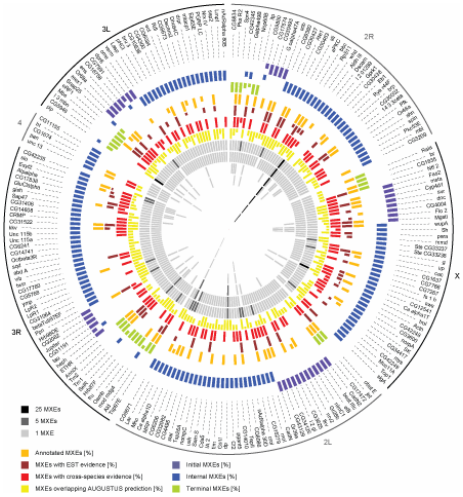


Figure 2. The mutually exclusive exome of *Drosophila melanogaster*. All genes containing predicted MXEs are listed.

In order to determine the extent of conservation within the Drosophila mutually exclusive spliced exomes we compared the data from *D.melanogaster* (dmel) with the reconstructed corresponding exomes of 11 further Drosophila species (Fig. 3A). In total, 2640 clusters were identified most of which are shared among several species, resulting in 770 unique clusters. Surprisingly, many of the clusters are unique to one of these groups like 164 clusters within the Drosophila subgenus group or 95 clusters within the obscura group. Only 68 clusters are conserved in all twelve species. To determine exon gain and loss during the evolution of the *Drosophila* species we counted these events based on maximum parsimony requiring the least exon loss events (Fig. 3B). The last common ancestor of the *Drosophila* species contained at least 186 clusters of MXEs. 456 clusters are unique to any of the *Drosophilas* and 111 clusters have been gained in certain branches.

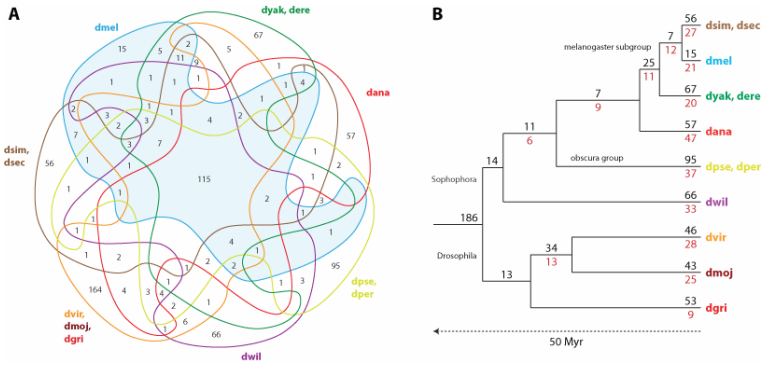


Figure 3. A) The Venn diagrams48 show the number of clusters of MXEs shared between species and subsets of species groups..B) The gain and loss of clusters of MXEs plotted onto the evolutionary tree of the Drosophila species.

3 Conclusion

Our analysis of the mutually exclusive exome of *D.melanogaster* considerably increased the number of mutually exclusive splicing events. Specifically, we have identified two times more internal MXE candidates than already annotated of which almost 80% are supported by evolutionary conservation or experimental transcript data.

References

[FO08] Florian Odronitz, Holger Pillmann, Oliver Keller, Stephan Waack, and Martin Kollmar. WebScipio: An online tool for the determination of gene structures using protein sequences. *BMC Genomics*, 9(422), 2008.

[HK13] Klas Hatje and Martin Kollmar. Predicting Tandemly Arrayed Gene Duplicates with WebScipio. *Nature Communications*, in revision.

[PH11] Holger Pillmann, Klas Hatje, Florian Odronitz, Björn Hammesfahr, and Martin Kollmar. Predicting mutually exclusive spliced exons based on exon length, splice site and reading frame conservation, and exon sequence homology. *BMC Bioinformatics*, 12(270), 2011.

Application of a Novel Triclustering Method (δ-TRIMAX) to Mine 3D Gene Expression Data of Breast Cancer Cells

Anirban Bhar[1], Martin Haubrock[1], Anirban Mukhopadhyay[2] and Edgar Wingender[1*]

1. Institute of Bioinformatics, University Medical Center Goettingen, Georg August University, Goettingen, Germany

2. Department of Computer Science and Engineering, University of Kalyani, Kalyani, India

1* edgar.wingender@bioinf.med.uni-goettingen.de

Abstract: We have proposed a novel triclustering algorithm δ-TRIMAX to mine 3D gene expression data sets by introducing a mean squared residue (MSR) score as a measure of coherence of the resultant triclusters. Applying our proposed algorithm on a time series gene expression dataset from an estrogen induced breast cancer cell, we identified key drivers for each resultant tricluster and found a number of hub genes that are known to be associated with breast cancer or estrogen responsive elements. Additionally, our coregulation analysis reveals synergistic regulatory effects of transcription factors.

1 Introduction

With the advent of microarray and other high-throughput technologies, it is feasible to measure expression profiles of thousands of genes across a set of samples and a set of time points. Exploratory approaches facilitate to analyze such high-throughput datasets and thus help to understand the phenotype of a cell. Coexpression analysis is instrumental in identifying genes that exhibit similar expression profiles in molecular networks. Highly interconnected genes in such lists of coexpressed genes are often called hub genes, the analysis of which may reveal underlying disease mechanisms. Clustering algorithms are useful to extract groups of genes or samples having similar expression profiles over all samples or genes, respectively. However, genes are not necessarily similarly expressed over all samples. To find local patterns in two-dimensional gene expression datasets, biclustering algorithms are used. However, to detect groups of genes that are coexpressed over a subset of samples during a subset

of time points, triclustering algorithms are required. Attempts to apply biclustering approaches to higher dimensional data would result in a disrupture of the time-dependent structure [SSW+07] and in an inappropriate amalgamation of the different dimensions, requiring extra efforts for postprocessing of the results. In a recent work we have proposed one triclustering algorithm δ-TRIMAX that aims to find triclusters from such 3D gene expression datasets [BHM+13]. We have delineated the coherence of a tricluster by introducing a novel measurement, called mean squared residue (MSR) score; each resultant tricluster must have an MSR score below a threshold δ [BHM+13]. In this work we have applied δ-TRIMAX on a time series gene expression dataset from an estrogen-induced breast cancer cell line to apprehend the underlying disease mechanisms, regulatory effects of transcription factors etc. Additionally, we have compared the capability of δ-TRIMAX with that of an existing triclustering algorithm using an artificial dataset and a real life dataset.

2 Method

Suppose D (G × C × T) represents a 3D gene expression dataset containing G, C and T number of genes, samples and time points, respectively. M(I, J, K) is a tricluster where I ⊆ G, J ⊆ C and K ⊆ T. We define Mean Squared Residue (MSR) to estimate the quality of a tricluster, i.e. the level of coherence among the elements of a tricluster as follows [BHM+13]: $\mathbf{MSR} = \frac{1}{|I||J||K|} \sum_{i \in I, j \in J, k \in K} (m_{ijk} - m_{iJK} - m_{IjK} - m_{IJk} + 2m_{IJK})^2$, where each element of the dataset is m_{ijk} and m_{iJK}, m_{IjK}, m_{IJk} correspond to the mean expression value of ith gene, jth sample, kth time point, respectively. m_{IJK} represents the mean over all genes, samples and time points. For further details of the method and for a description of the whole workflow, we refer to the original publication [BHM+13].

3 Results and Conclusions

For validation of our algorithm, we have applied it first on a synthetic dataset with implanted triclusters and different levels of noise. Comparing δ-TRIMAX with another algorithm, we found that δ-TRIMAX was more reliable in re-identifying the artificial triclusters. We have then applied our

algorithm on a time series gene expression dataset from estrogen-induced breast cancer cells to understand the underlying mechanisms of transcriptional regulation during different stages of estrogen response [CMS+06]. We have compared the performance of our algorithm with that of an existing algorithm in terms of coverage, statistical difference from background (SDB) and triclustering quality index (TQI) using the real life dataset [BHM+13]. We could demonstrate that δ-TRIMAX outperforms the existing algorithm according to each of these criteria. To assess the biological significance of genes belonging to each resultant tricluster we performed Gene Ontology biological process (GOBP) and KEGG pathway enrichment analysis. We have observed GOBP enrichment for genes belonging to each tricluster. We used the singular value decomposition (SVD) method to represent each tricluster by its eigen gene. Then we detected hub genes for the co-expression network of each resultant tricluster by calculating Pearson correlation coefficients between eigen gene and expression values of each gene of a tricluster over the samples and time points that are present in that tricluster. The genes (more specifically, the probeset IDs) were sorted in descending order of correlation coefficients. From the 10 topmost probeset IDs, hub genes of each tricluster were identified. This way, we have identified *NPC1L1, TMEM161B-AS1, POU5F1P3, POU5F1P4, POU5F1B, CCL2* as those hub-genes that are coexpressed over all-time points and samples. The chemokine *CCL2* has already been reported to play a role in breast cancer development [TCW+12]. Isoforms or pseudogenes of the transcription factor POU5F1 / OCT4, in particular POU5F1P4, have been found to play specific roles in other types of cancer [WGZ+13], while our results suggest for three of them that they are involved in breast cancer as well. The intestinal cholesterol absorption protein *NPC1L1* has already been inferred to be a target of liver X receptors (LXR) which play an instrumental role in breast carcinogenesis [VDI+04,DTT+06]. Additionally, a previous study infers that estrogen plays an important role in the upregulation of *NPC1L1* [VCR07]. It is already known that the intestinal cholesterol absorption can be used as a drug target for reducing the plasma cholesterol level below the threshold where it promotes the development of tumors and aggravates their aggressiveness [LDM+11, TD03]. Thus we can hypothesize *NPC1L1* as a potential drug target to prevent the growth of breast tumors. We have identified such key drivers for other triclusters as well and found that many of those hub genes are already reported to be associated with breast cancer or estrogen responsive elements. Moreover we performed TFBS enrichment analysis to identify statistically enriched transcriptional regulatory elements in the promoter regions of coexpressed genes using the TRANS-

FAC library (version 2009.4). From this analysis potential coregulation of the coexpressed genes could be inferred. The TFBS found to be enriched also suggested synergistic regulatory effects of transcription factors such as CREB, ATF3, Sp1 etc., which are already known to play crucial roles in breast cancer. We thus feel that our triclustering approach is very suitable to provide biologically meaningful hypotheses, in the example shown about the development of breast cancer.

References

[BHM+13] A Bhar, M Haubrock, A Mukhopadhyay, U Maulik, S Bandyopadhyay, and E Wingender. Coexpression and coregulation analysis of time-series gene expression data in estrogen-induced breast cancer cell. *Algorithms for molecular biology*, 8(1), March 23 2013.

[CMS+06] J S Carroll, C A Meyer, J Song, W Li, T R Geistlinger, J Eeckhoute, A S Brodsky, E K Keeton, K C Fertuck, G F Hall, Q Wang, S Bekiranov, V Sementchenko, E A Fox, P A Silver, T R Gingeras, X S Liu, and M Brown. Genome-wide analysis of estrogen receptor binding sites. *Nature Genetics*, 38(11):1289–1297, November 2006.

[DTT+06] C Duval, V Touche, A Tailleux, J C Fruchart, C Fievet, V Clavey, B Staels, and S Lestavel. Niemann-Pick C1 like 1 gene expression is down-regulated by LXR activators in the intestine. *Biochemical and Biophysical Research Communications*, 340(4):1259–1263, February 24 2006.

[LDM+11] G Llaverias, C Danilo, I Mercier, K Daumer, F Capozza, T M Williams, F Sotgia, M P Lisanti, and P G Frank. Role of cholesterol in the development and progression of breast cancer. *The American Journal of Pathology*, 178(1):402–412, January 2011.

[SSW+07] J Supper, M Strauch, D Wanke, K Harter, and A Zell. EDISA: extracting biclusters from multiple time-series of gene expression profiles. *BMC Bioinformatics*, 8(334), September 12 2007.

[TCW+12] A Tsuyada, A Chow, J Wu, G Somlo, P Chu, S Loera, T Luu, A X Li, X Wu, W Ye, S Chen, W Zhou, Y Yu, Y Z Wang, X Ren, H Li, P Scherle, Y Kuroki, and S E Wang. CCL2 mediates cross-talk between cancer cells and stromal fibroblasts that regulates breast cancer stem cells. *Cancer Research*, 72(11):2768–79, June 1 2012.

[TD03] S D Turley and J M Dietschy. The intestinal absorption of biliary and dietary cholesterol as a drug target for lowering the plasma cholesterol level. *Preventive Cardiology*, 6(1):29–33, Winter 2003.

[VCR07] M A Valasek, S L Clarke, and J J Repa. Fenofibrate reduces intestinal cholesterol absorption via PPARα-dependent modulation of NPC1L1 expression in mouse. *Journal of Lipid Research*, 48(12):2725–2735, December 2007.

[VDI+04] D M Vigushin, Y Dong, L Inman, N Peyvandi, J P Alao, C Sun, S Ali, E J Niesor, C L Bentzen, and R C Coombes. The nuclear oxysterol receptor LXRalpha is expressed in the normal human breast and in breast cancer. *Medical Oncology*, 21(2):123–131, 2004.

[WGZ+13] L Wang, Z Y Guo, R Zhang, B Xin, R Chen, J Zhao, T Wang, W H Wen, L T Jia, L B Yao, and A G Yang. Pseudogene OCT4-pg4 functions as a natural micro RNA sponge to regulate OCT4 expression by competing for miR-145 in hepatocellular carcinoma. *Carcinogenesis*, 00(00), May 23 2013.