Expose Master thesis

Integrating miRNAs into gene regulatory networks for identification of lymphoma-relevant genes

Yvonne Mayer
<yvonne.mayer@fu-berlin.de>

Humboldt-Universität zu Berlin
Faculty of Mathematics and Natural Sciences *II*
Department of Computer Science
Knowledge Management in Bioinformatics

Berlin, 30. Mai 2012

1 Motivation and Background

Over the past decades it has been thought that only the protein coding part of the genome is relevant for the body functions and their regulation. In 1993 Victor Ambros et al. [1] discovered a gene, *lin-4*, that affected development in *C. elegans*, although its product is not a protein but a short RNA molceule. This was the first discovered MicroRNA (miRNA). But only in the 2000s researchers started to consider these MicroRNAs as a distinct class of biological regulators beside transcription factors. MicroRNAs are very short, the average length is 22 nucleotides. As negative, post-transcriptional regulators of gene expression they bind to partially complementary sites of mRNAs 3' UTR region and cause inhibition of translation or, in most cases, degradation of the mRNA, resulting in decreased mRNA levels [2]. Today $\sim$ 1200 miRNAs in humans are known. Over 30% of mRNAs are supposed to be regulated by miRNAs. The mRNAs regulated by miRNAs mostly have roles in multiple biological processes like differentiation, cell proliferation and apoptosis. Several studies indicate that miRNAs play a critical role in the development and progression of various diseases, leading to abnormal expression profiles of miRNA and mRNA levels [3]. Moreover several miRNAs are supposed to act as tumor suppressors and oncognes in different cancer types, including lymphoma. These miRNAs are often referred to as "oncomirs" [4].

In the context of the DFG funded project TRR54 *"Growth and Survival, Plasticity and Cellular Interactivity in Lymphoid Malignancies"* a dataset comprising miRNA high-throughput sequencing data and mRNA microarray data from 108 patients suffering from seven different lymphoma types is available. This dataset also includes control patients.

In my master thesis I want to concentrate on one lymphoma subtype and combine available miRNA and mRNA data in gene regulatory networks. I want to identify those genes that have different centrality in healthy and diseased persons. It is conceivable that genes with different centrality can have an essential role in the development of disease patterns (here lymphoma) and could be potential therapeutic targets [5, 6, 7]. In addition I want to analyse network motifs (see section 1.1) and compare their occurences in healthy and diseased persons. The analysis of networks motifs can be used to further an understanding of the networks behaviour and they offer potential targets for trying to interfere with the networks behaviour [8]. Furthermore I will investigate whether it is beneficial to integrate miRNAs in gene regulatory networks for the identification of such lymphoma-relevant genes.

1.1 Network Motifs

A network motif is a subgraph and is defined as a pattern of interaction between nodes, which recur in different parts of a network significantly more often than would be expected in a random network. Most gene regulatory networks of well-studied microorganisms appear to consist of a small set of network motifs, which are occuring again and again, thus constituting almost the entire network [9]. Network motifs can be seen as simple building blocks, regulating each others transcription rate, from which the network is composed [10]. It has been shown that these network motifs can be responsible for specific information-processing functions [9].

The most simple network motifs are a transcription factor (TF) X regulating gene Y, negative autoregulation (NAR) and positive autoregulation (PAR). NAR/PAR occurs when a TF repres-

ses/activates its own transription. NAR has two important functions, first it speeds up the response time of gene circuits and second it can reduce cell-cell variation in protein levels. The effects of PARs are opposite. Another class of network motifs is the feedforward loop (FFL). The motif consits of three nodes, gene X which regulates Y and Z. Z is additionally regulated by Y. The regulatory interactions can bee inhibitory or activating, thus leading to eight possible FFLs, all having different functions. A summary of all typical network motifs and their associated functions was published 2007 by Uri Alon [9]. *Escherichia coli* was the first organisms, in which network motifs were systematically discovered. The gene regulatory network of *Escherichia coli* is largeley composed of three main motif families [11]. Since then the same motifs have been found in several organisms, from bacteria to human.

The common procedure to find significant network motifs in a given network is described e.g. in [10]. It should be noted, that motif search also has provoked criticism: For determining which motifs occur at higher frequency than in random networks, a specified random graph model must be established, and this models are highly controversial [12, 13]. In my master thesis I will therefore primarily count network motifs and compare their occurences between healthy and diseased cells.

2 Goal

The goal of this thesis is to investigate the influence of miRNAs in gene regulatory networks and to elucidate whether their integration improves the identification of genes playing a role in lymphoma.

For this purpose a network will be built from high-throughput sequencing and microarray data which combines miRNAs, their targets and the targets of the miRNA targets as well. Therefore the interactions in the network will be extracted from public databases and the expression data will be used to characterize the interactions. The edges in the network should be directed, weighted (according to the correlation of expression values) and signed (positive or negative regulation). Such a network will be built for healthy and cancerous cells and both will be characterized with different methods (see section 3). Results will be compared to identify genes that behave differentially in healthy and cancerous cells, leading to a list of genes potentially important for the development of lymphoma. Secondarily, for evaluating the influence of miRNAs, the same two networks without miRNA will be constructed (i.e. networks only consist of mRNA and their regulating transcription factors [1]) and analysed in the same way.

For evaluating the influence of miRNA the obtained lists are compared with a manually curated list of genes relevant to lymphoma (gold standard list).

[1] In case that the network decays one could instead build a gene co-expression network from microarray data.

3 Approach

Experimental High-Throughput Datasets

The datasets available in the TRR54 project comprise miRNA high-throughput sequencing data (platform: ABI SOLiD) and mRNA microarray data (platform: exon arrays) from 103 patients suffering from seven different lymphoma types. These datasets also include control patients (Tonsil).

Interaction Data

Targets of miRNAs will be taken from the databases *TarBase* [14] and *miRTarBase* [15], since these data are experimentally validated. Additionally, targets, which are listed in at least two of the databases *TargetScan* [16], *PicTar* [17], *miRanda* [18] or *microCosm* [19] will be extracted. These databases contain computationally predicted targets. For definitions of proteins as TFs and their targets a list combining regulatory interactions from *TRANSFAC* [20],*ORegAnno* [21] and *TRRD* [22] and containing manually revised interactions obtained by text mining methods will be used. This list is available in the WBI group. TF $\rightarrow$ miRNA regulations will be taken from the database *TransmiR* [23]. *TransmiR* contains 649 manually curated regulatory relationships in humans (survey of $\sim$ 5000 papers) including signs of interactions.

Construction of Networks

The interaction data will be used to construct the networks. As described in section 2 alltogether four gene regulatory networks will be built. At first two networks integrating miRNA data will be constructed (one for healthy and one for cancerous cells). These networks have three kinds of nodes (miRNA, gene and TF), leading to four different types of directed edges in the network: miRNA $\rightarrow$ gene, TF (gene) $\rightarrow$ gene, TF (gene) $\rightarrow$ miRNA and TF (gene) $\rightarrow$ TF (gene). The networks without miRNA are obtained by simply removing miRNA nodes and their interactions.

For each edge a weight w is assigned, which corresponds to the Pearson correlation of expression values beetween the two nodes. If the absolute value of w is below a certain threshold the edge will be removed from network. For interactions of type TF (gene) $\rightarrow$ gene and TF (gene) $\rightarrow$ TF (gene) signs of regulatory relationships are determined from correlation between the two genes, i.e. the interaction from A to B is activating (inhibitory), if the Pearson correlation is positive (negative). Signs of TF (gene) $\rightarrow$ miRNA interactions are taken from *TransmiR*. Regulatory interactions by miRNAs are regarded as negative.

Network Analysis

The networks will be characterized with different methods including the basic topology (number of nodes and edges, clustering coefficient, degree districutions, path lengths), identification of network hubs with a centrality measure and occurence of network motifs.

To measure the centrality of nodes most likely degree centrality and betweeness centrality will be used. For identification of (enriched) network motifs I will use an existing tool or algorithm, most

likely this will be the sampling tool *FANMOD* [24], which allows to search for (enriched) network motifs in networks with different node and edge types. For the enrichment analysis for a set of genes (e.g. network hubs, genes participating in certain FFLs) the *DAVID*- [25] or *BiNGO*-tool [26] could be used.

Evaluation

For evaluating the influence of miRNA the obtained lists containing genes ranked differentially in healthy and diseased cells need to be compared with a manually curated list of genes relevant to lymphoma (gold standard list). This list will be obtained from lymphoma experts working in the TRR54 project at the Charité - Universitätsmedizin Berlin.

4 Related Work

Several studies have attempted to integrate regulation by TFs and miRNA into a network. Mostly these networks are limited in terms of their datasets, e.g. they are purely based on computational predictions, interactions between TFs and genes or miRNAs are not incorporated or the signs of regulatory interactions are not considered.

Schramm et al. (2010) [27] investigated regulation patterns in signaling networks of eleven different tumor types, based on correlation of gene expression between interacting proteins. Therefore microarray datasets of malignant cells and the corresponding non-malignant tissue samples were used. They included PPI and distinguished between TFs, receptors and normal proteins. Although Schramm et al. did not include miRNAs in their networks, their work is interesting for my master thesis, since they compared networks which were inferred from malignant cells with that inferred from non-malignant cells. A variety of network features was calculated and significance tests of the pairwise differences between tumor and normal (paired, non-parametric Wilcoxon-rank test) were performed. In summary they discovered that non-malignant cells form a more centralized network: the network is smaller with a higher frequency of hubs and a higher clustering coefficient. In normal cells the same hubs are more often used for different signaling tasks. Schramm et al. defined signaling motifs describing this fact: Two different signals are transmitted from two receptors to a transcription factor (TF). In normal cells the motif where the pathways from the receptors to the TFs share common links is counted to a significantly higher number than in tumor cells. Networks inferred from malignant cells show shorter path lengths in a larger network with a higher robustness to removal of hubs.

Cheng et al. (2011) [28] recently constructed a network similar to my approach for analyzing gene regulation in *C. elegans*. Additionally they included protein-protein interactions (PPI). The TF (gene) $\rightarrow$ miRNA and TF (gene) $\rightarrow$ TF (gene) interactions were extracted from ChIP-Seq binding profiles. Predicted targets of miRNAs were identified by the *PicTar*- [17] and *TargetScan*-algorithm [16]. Expression levels of mRNA and miRNA were obtained from RNA-seq experiments and are used to weight the edges by correlation. As a first step Cheng et al. analysed the basic topology of the network (number of nodes and edges, degree distributions for the different kinds of nodes, correlation between in- and out-degrees for miRNAs etc.). For a better visualization the hierarchy of the network was analysed. RNAi screening of the TFs in the different layers showed that downstream TFs are more essential for survival of *C. elegans*. Besides the TFs in different layers have different topological properties (e.g. TFs in the bottom layer are more likely to be regulated by miRNAs).

Lastly significant network motifs in the signed and unsigned regulatory network of *C. elegans* were determined. For this purpose they used the sampling tool *FANMOD* [24]. The most interesting enriched motifs including miRNA that were discovered are a TF↔miRNA composite feedback loop (a TF that regulates a miRNA is regulated itself by that same miRNA), the 3-node motif with a pair of interacing proteins being regulated by a common miRNA and the 3-node motif where a TF and its downstream target (gene, TF or miRNA) are simultaneously repressed by a common TF. Enriched motifs were responsible for robustness during development and play a role in periodic process (e.g. molting in different larvae stages).

The authors suggest to improve the network by including interaction data from public databases (both experimentally verified and computationally predicted interactions) and by filtering out interactions of low confidence.

For analysing the role of miRNAs in human ovarian cancer, Schmeier et al. (2011) [8] built a network, comprising the same type of interactions as Cheng et al.. They did not use experimental data and hence did not compare a network inferred from healthy with that inferred from cancerous tissue. Instead the whole network was constructed with knowledge from various public databases. As a starting point they selected genes and miRNAs, which are related to ovarian cancer (e.g. from Dragon Database for Exploration of Ovarian Cancer Genes). Interactions for this set of genes and miRNAs were also extracted from different databases. TF↔miRNA interactions were inferred from ab initio transcription factor binding site predictions.

In order to identify nodes with high influence on the overall network Schmeier et al. developed their own measure, instead of an established centrality measure. It is based on the outgoing edges of a node and considers nodes up to the second degree. The score S_n of a node n is defined as:

$$S_n = \sum e_{n1} + w * \sum e_{n2},$$

where w is the weighting factor for second-degree edges e_{n2}, since they should have less influence on the score. Weights were sampled randomly (average of 10.000 samplings). In this procedure they ensured that the edges between a miRNA and its targets have the highest weight, since they are experimentally verified. The undirected edges between proteins got the lowest weight.

Next they identified three-element network motifs containing one miRNA and two proteins, one being a TF. An enrichment analysis revealed that genes involved in such kinds of feedback loops are enriched in the highly relevant pathways cell cycle regulation and apoptosis.

In order to understand the gene expression regulation in T-cell acute lymphoblastic leukemia (T-ALL), Ye et al. [29] started from experimentally verified T-ALL related miRNAs, genes and TFs and combined them into 120 FFLs, finally they constructed a network from these FFLs. Interations between miRNAs, TFs and genes were extracted from databases containing predicted targets (e.g. *TargetScan* [16]). In this network they identified network hubs, which were simply defined as nodes with more than eight ingoing or outgoing edges. Four hub genes and miRNAs were selected and their subnetworks extracted and investigated. The analysis of this subnetworks revelaed that miR-19, CYLD and NF-KB form a regulatory FFL, which provides new clues for sustained activation of NF-KB in T-ALL. Ye et al. confirmed these findings with different experiments (e.g. luciferase reporter assays). All in all the constructed network from Ye et al. differs from my approach, as I will have weighted edges from correlation between expression values.

Literaturverzeichnis

[1] Rosalind C. Lee, Rhonda L. Feinbaum, and Victor Ambros. The c. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, December 1993.

[2] Huili Guo, Nicholas T. Ingolia, Jonathan S. Weissman, and David P. Bartel. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–840, August 2010.

[3] Qinghua Jiang, Yadong Wang, Yangyang Hao, Liran Juan, Mingxiang Teng, Xinjun Zhang, Meimei Li, Guohua Wang, and Yunlong Liu. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Research*, 37(Database issue):D98–D104, January 2009. PMID: 18927107 PMCID: PMC2686559.

[4] Aurora Esquela-Kerscher and Frank J. Slack. Oncomirs - microRNAs with a role in cancer. *Nature Reviews Cancer*, 6(4):259–269, April 2006.

[5] Diana Ekman, Sara Light, Asa K. Björklund, and Arne Elofsson. What properties characterize the hub proteins of the protein-protein interaction network of saccharomyces cerevisiae? *Genome Biology*, 7(6), June 2006.

[6] Gordana Apic, Tijana Ignjatovic, Scott Boyer, and Robert B Russell. Illuminating drug discovery with biological pathways. *FEBS Letters*, 579(8):1872–1877, March 2005. PMID: 15763566.

[7] Trey Ideker and Roded Sharan. Protein networks in disease. *Genome Research*, 18(4):644–652, April 2008.

[8] Sebastian Schmeier, Ulf Schaefer, Magbubah Essack, and Vladimir B Bajic. Network analysis of microRNAs and their regulation in human ovarian cancer. *BMC Systems Biology*, 5:183, November 2011. PMID: 22050994 PMCID: 3219655.

[9] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, June 2007.

[10] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, October 2002.

[11] Shai S. Shen-Orr, Ron Milo, Shmoolik Mangan, and Uri Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, 31(1):64–68, April 2002.

[12] Sebastian Wernicke. A faster algorithm for detecting network motifs. In Rita Casadio and Gene Myers, editors, *Algorithms in Bioinformatics*, volume 3692 of *Lecture Notes in Computer Science*, pages 165–177. Springer Berlin / Heidelberg, 2005.

[13] Reid Ginoza and Andrew Mugler. Network motifs come in sets: Correlations in the randomization process. *Physical Review E*, 82(1):011921, July 2010.

[14] Thanasis Vergoulis, Ioannis S. Vlachos, Panagiotis Alexiou, George Georgakilas, Manolis Maragkakis, Martin Reczko, Stefanos Gerangelos, Nectarios Koziris, Theodore Dalamagas, and Artemis G. Hatzigeorgiou. TarBase 6.0: capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Research*, 40(D1):D222–D229, January 2012. PMID: 22135297 PMCID: 3245116.

[15] Sheng-Da Hsu, Feng-Mao Lin, Wei-Yun Wu, Chao Liang, Wei-Chih Huang, Wen-Ling Chan, Wen-Ting Tsai, Goun-Zhou Chen, Chia-Jung Lee, Chih-Min Chiu, Chia-Hung Chien, Ming-Chia Wu, Chi-Ying Huang, Ann-Ping Tsou, and Hsien-Da Huang. miRTarBase: a database curates experimentally validated microRNA–target interactions. *Nucleic Acids Research*, 39(Database issue):D163–D169, January 2011. PMID: 21071411 PMCID: PMC3013699.

[16] 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, January

2005.

[17] Azra Krek, Dominic Gr|[uuml]|n, 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, April 2005.

[18] Bino John, Anton J Enright, Alexei Aravin, Thomas Tuschl, Chris Sander, and Debora S Marks. Human MicroRNA targets. *PLoS Biology*, 2(11), November 2004. PMID: 15502875 PMCID: PMC521178.

[19] Sam Griffiths-Jones, Harpreet Kaur Saini, Stijn van Dongen, and Anton J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Research*, 36(Database issue):D154–D158, January 2008. PMID: 17991681 PMCID: PMC2238936.

[20] V. Matys and V. Kel-Margoulies. TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Research*, 34, 2005.

[21] Obi L. Griffith, Stephen B. Montgomery, Bridget Bernier, Bryan Chu, Katayoon Kasaian, Stein Aerts, Shaun Mahony, Monica C. Sleumer, Mikhail Bilenky, Maximilian Haeussler, Malachi Griffith, Steven M. Gallo, Belinda Giardine, Bart Hooghe, Peter Van Loo, Enrique Blanco, Amy Ticoll, Stuart Lithwick, Elodie Portales-Casamar, Ian J. Donaldson, Gordon Robertson, Claes Wadelius, Pieter De Bleser, Dominique Vlieghe, Marc S. Halfon, Wyeth Wasserman, Ross Hardison, Casey M. Bergman, and Steven J.M. Jones. ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Research*, 36(Database issue):D107–D113, January 2008. PMID: 18006570 PMCID: PMC2239002.

[22] N A Kolchanov, O A Podkolodnaya, E A Ananko, E V Ignatieva, I L Stepanenko, O V Kel-Margoulis, A E Kel, T I Merkulova, T N Goryachkovskaya, T V Busygina, F A Kolpakov, N L Podkolodny, A N Naumochkin, I M Korostishevskaya, A G Romashchenko, and G C Overton. Transcription regulatory regions database (TRRD): its status in 2000. *Nucleic Acids Research*, 28(1):298–301, January 2000. PMID: 10592253.

[23] Juan Wang, Ming Lu, Chengxiang Qiu, and Qinghua Cui. TransmiR: a transcription factor–microRNA regulation database. *Nucleic Acids Research*, 38(Database issue):D119–D122, January 2010. PMID: 19786497 PMCID: 2808874.

[24] Sebastian Wernicke and Florian Rasche. FANMOD: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, May 2006.

[25] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1):44–57, December 2008.

[26] Steven Maere, Karel Heymans, and Martin Kuiper. BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21(16):3448–3449, August 2005.

[27] Gunnar Schramm, Nandakumar Kannabiran, and Rainer König. Regulation patterns in signaling networks of cancer. *BMC Systems Biology*, 4(1):162, 2010.

[28] Chao Cheng, Koon-Kiu Yan, Woochang Hwang, Jiang Qian, Nitin Bhardwaj, Joel Rozowsky, Zhi John Lu, Wei Niu, Pedro Alves, Masaomi Kato, Michael Snyder, and Mark Gerstein. Construction and analysis of an integrated regulatory network derived from High-Throughput sequencing data. *PLoS Computational Biology*, 7(11), November 2011. PMID: 22125477 PMCID: 3219617.

[29] Huashan Ye, Xiaowen Liu, Meng Lv, Yuliang Wu, Shuzhen Kuang, Jing Gong, Ping Yuan, Zhaodong Zhong, Qiubai Li, Haibo Jia, Jun Sun, Zhichao Chen, and An-Yuan Guo. MicroRNA and transcription factor Co-Regulatory network analysis reveals miR-19 inhibits CYLD in T-Cell acute lymphoblastic leukemia. *Nucleic Acids Research*, February 2012.