International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:5, No:8, 2011

Full-genomic network inference for non-model organisms: A case study for the fungal pathogen Candida albicans

Jörg Linde, Ekaterina Buyko, Robert Altwasser, Udo Hahn, Reinhard Guthke

Abstract—Reverse engineering of full-genomic interaction networks based on compendia of expression data has been successfully applied for a number of model organisms. This study adapts these approaches for an important non-model organism: The major human fungal pathogen *Candida albicans*. During the infection process, the pathogen can adapt to a wide range of environmental niches and reversibly changes its growth form. Given the importance of these processes, it is important to know how they are regulated. This study presents a reverse engineering strategy able to infer full-genomic interaction networks for *C. albicans* based on a linear regression, utilizing the sparseness criterion (LASSO). To overcome the limited amount of expression data and small number of known interactions, we utilize different prior-knowledge sources guiding the network inference to a knowledge driven solution. Since, no database of known interactions for *C. albicans* exists, we use a text-mining system which utilizes full-text research papers to identify known regulatory interactions. By comparing with these known regulatory interactions, we find an optimal value for global modelling parameters weighting the influence of the sparseness criterion and the prior-knowledge. Furthermore, we show that soft integration of prior-knowledge additionally improves the performance. Finally, we compare the performance of our approach to state of the art network inference approaches.

Keywords—Pathogen, Network Inference, Text-Mining, *Candida albicans*, LASSO, Mutual Information, Reverse Engineering, Linear Regression, Modelling

I. INTRODUCTION

THE esearch community has successfully predicted full-genomic interaction networks of model organisms, such as *Escherichia coli* [1]. From a methodological point of view, the various full-genomic network inference methods can be divided into approaches based on (partial) correlation [2], information theory [1], [3], [4], and linear regression [5]. The integration of prior-knowledge based on additional data sources to gene expression data significantly improves the reverse engineering approach [6]–[8].

Understanding the interaction networks of human pathogenic microorganisms is important for the identification of drug targets and design of medical treatment. So far, only small scale models for pathogenic bacteria [9] or fungi [8], [10], [11] have been suggested. While for model organisms large databases with known interactions exist [12] and a large amount of expression data is available [13], for *C. albicans* less expression data is available and only a few interactions are known(but not collected in a data base).

During the last decades, the morbility and mortality rates due to *C. albicans* infections have been increasing, making this organisms one of the most important human fungal pathogens [14]. The infection process is characterized by a change from a harmless commensal to an aggressive pathogen, phenotipic growth form switches, and adaptations to changing environmental parameters (pH, temperature, nutrient availability...) [15]. All these processes lead to dramatic changes in gene expression patterns [15], [16]. Unrevealing the underlying interaction network will improve our understanding on how the pathogen is able to start and maintain infection. Since there are only a few regulatory interactions known, it is important to use computer based models to predict gene interactions in *C. albicans* . During the last decades many *C. albicans* researchers have performed gene expression studies and a large amount data has become publicly available [13], facilitating the inference of full-genomic network models.

Presently, human professionals called “biocurators” create and maintain *gold-standard* databases of scientific knowledge from molecular biology. The curation task is known to be an extremely time-consuming and manual process. As Baumgartner *et al* [17] have shown, the exponential growth rate of publications already outpaces human capabilities to keep track with the speed of publication of documents relevant for database curation. Hence, the current state of the art of database curation requires a profound change of methodologies for accessing and structuring information from the biomedical literature. Hahn *et al* [18] promote applications of automatic text mining procedures that would render a reasonable support since considerable progress has been made during the past years in this field. Indeed, the automatic harvesting of information from biomedical literature has caught high attention in recent years, and is witnessed by various challenges such as BIOCREATIVE(Critical Assessment of Information Extraction systems in Biology) [19] and the BIONLP SHARED TASK ON EVENT EXTRACTION [20] series. The success of text mining tools for automatic database synthesis has already been demonstrated for the REGULONDB [12], world’s largest manually curated reference database for the transcriptional regulation network of *E. Coli* (e.g., [21], [22]).

This study reverse engineers full-genomic gene interaction

J. Linde, R. Altwasser, and R. Guthke are with the Research Group Systems Biology / Bioinformatics at the Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute (HKI), D-07745 Jena, Germany, Email: joerg.linde@hki-jena.de, www.sysbio.hki-jena.de/

Ekaterina Buyko and Udo Hahn are with the Jena University Language and Information Engineering (JULIE) Lab at the Friedrich-Schiller-University D-07743 Jena, Germany, www.julielab.de/

networks of *C. albicans* . In order to monitor the expression of the fungal genes over a large number of conditions, we collected a compendium of micro-array data. To support the inference approach, we compiled a number of additional data sets proposing potential interactions (prior-knowledge). Finally, we apply a reverse engineering strategy based on linear regression, which softly integrates this prior-knowledge. In order to evaluate the performance of our network inference approach, we use a text-mining algorithm together with full-text research papers to compile a list of known interactions in *C. albicans* ("gold-standard").

II. Data and Methods

Expression data

We collected expression data for the *C. albicans* strain SC5314 from the Gene Expression Omnibus Database [13]. In cases of internal replicates on one array (replicated spots), we calculated the median of those replicates. The collected compendium of expression data consists of 6269 genes whose expression was monitored under 491 conditions. The compendium contains 19.5% missing values. We imputed those missing values following a similar approach as Albrecht *et al* [23]. First, we removed 504 genes, and 9 experiments, which had more than 50% missing values. Then, we tested seven imputation methods: from the R package *impute* [24]) : K-nearest-neighbour; and all six methods from the R package *pcaMethods* [25]. The K-nearest-neighbour method turned out to perform best (data not shown) and was chosen to impute missing values on the original gene expression data set.

Prior-knowledge

This study compiles three different prior-knowledge sources for the prediction of gene interactions, softly integrates them during the network inference and evaluates the performance of the network inference based on these sources. For source 1-3, we translated each interaction from *Saccharomyces cerevisiae*, where both interaction partners have orthologs, into a putative interaction in *C. albicans* , using the orthologe definition of the CGDB [26].

Source 1 PPI: Protein-protein interactions (PPI) for *S. cerevisiae* downloaded from the MPACT section of the CYGD database [27].

Source 2 BIND: The Biomolecular Interaction Network Database (BIND) [28] includes interactions which are either individually submitted, based on PDB or on large scale interaction experiments.

Source 3 TRANS: This source utilizes a network of transcriptional relations of *S. cerevisiae* [29].

Source 4 FAC: This source uses known physical transcription factor - target gene interactions from the TRANSFAC database [30]. After downloading all fungal interactions, we blasted the protein sequences of the transcription factors and their targets against the *C. albicans* proteome. We defined best BLAST hits with a sequence similarity larger than 25% and an E-value smaller than 0.001 as orthologs.

Table I compares the prior-knowledge sources with each other and the gold-standard (see Gold-Standard). All sources differ in the number of genes and interactions. TRANS and FAC consist of few interactions which are mainly not involved in the gold-standard. Reasons for this might be that transcriptional interactions are hardly mentioned in the literature, the text-mining algorithm has problems in identifying these interactions, *C. albicans* has a very different transcriptional regulatory network than other fungi. On the other hand, there are more PPI interactions and this prior-knowledge source has the most interactions in common with the gold-standard. Since the BIND database includes results from a number of large scale experiments, it predicts by far the most interactions.

TABLE I: Comparison of the different prior-knowledge sources.

	PPI	BIND	TRANS	FAC
#(genes)	2290	6333	1502	226
#(goldgenes ∩ prior-knowledge genes)	281	266	208	92
#(interactions)	6674	2288	2689	249
#(interactions between goldgenes)	341	266	232	57
#(interactions between goldgenes ∩ prior-knowledge interactions)	47	40	15	14

III. Automatically Generated Gold-Standard

In order to automatically generate a gold-standard, we chose the information extraction approach of [31] to automatically extract gene regulation relations from biological documents. It is based on JReX, a high-performance machine-learning-based relation extraction system, which scored on second rank in the BioNLP Shared Task on Event Extraction [20]. Furthermore, JReX has already been successfully applied for the reconstruction of the RegulonDB [12] [22]. Generally speaking, the JReX system classifies pairs of genes in sentences as *interaction pairs* using various syntactic and semantic information (cf. [32] for a deeper account). In this work, we applied the JReX system for the extraction of gene expression regulation relations on a full-text article collection (5996 freely available *C. albicans* research papers). JReX extracted from the collected document set 971 interactions between 500 genes which are included in the gene expression compendium (called goldgenes hereafter). This list was used as gold-standard in order to compare the performance between different network inference approaches, different prior-knowledge sets and modelling parameter settings.

Network Inference

This study uses a system of linear equations to model the interactions between genes of *C. albicans* based on gene expression data $x_j(k)$ under the experimental condition k. For all conditions $k = [1 \ldots M]$, the expression of gene i is modelled as the weighted sum of the expressions of the other genes:

$$\hat{x_i}(k) = \sum_{\substack{j=1,\\ j\neq i}}^{N} \beta_{i,j} x_j(k) \tag{1}$$

where N is the number of genes. $\beta_{i,j}$ is the coefficient (weight), describing the influence of gene j on the expression of gene i. The estimated coefficients $\beta_{i,j}$ describe the interactions between the genes: Positive coefficients describe an activation, negative describe an inhibition, zero coefficients describe no interaction. The absolute value of the coefficient describes the strength of the interaction. These interactions may be direct physical protein-gene interactions, or have a more conceptual meaning (for example if there is a pathway or signal cascade between two interactors).

The basic task when inferring interaction networks is to use the expression data in order to define the coefficients of the model, so that the modelled expression data $\hat{x_i}(k)$ fits well to the measured expression data $x_i(k)$. However, the equation system is strongly underestimated. The number of genes is larger than the number of measured conditions. Therefore, we apply the biological constraint of sparseness. Utilizing sparseness, means minimizing the the number of non-zero model parameters.

In an attempt to combine model selection and keep the number of edges (non-zero parameters) small, Tibshirani et al. [33] developed the *Least absolute shrinkage and selection operator* (LASSO) algorithm. This algorithm infers the interaction matrix $B = \{\beta_{i,j}\}$ which is filled with a maximum number of zeros, while the linear model is still able to adapt to the measured expression values.

This leads to the least squares criterion with a limit:

$$\hat{B} = \arg\min_{B}\Big(\sum_{k=1}^{M}\sum_{i=1}^{N}\big(x_i(k) - \sum_{\substack{j=1,\\ j\neq i}}^{N}\beta_{i,j}\hat{x_j}(k)\big)^2 + \lambda\sum_{i=1}^{N}\sum_{\substack{j=1,\\ j\neq i}}^{N} \mid \beta_{i,j} \mid\Big) \qquad (2)$$

where λ is a parameter modelling the influence of the sparseness criterion. The more λ increases the more influence does the L_1 penalty get and the coefficients shrink to zero.

In order to further increase the perfromance quality, Zou presented the adaptive LASSO [34], facilitating the possibility of using prior-knowledge during the inference process.

In this study, we applied one global value ω for all interactions predicted by one prior-knowledge source.

$$\hat{B} = \arg\min_{B}\Big(\sum_{k=1}^{M}\sum_{i=1}^{N}\big(x_i(k) - \sum_{\substack{j=1,\\ j\neq i}}^{N}\beta_{i,j}\hat{x_j}(k)\big)^2 + \lambda\sum_{i=1}^{N}\sum_{\substack{j=1,\\ j\neq i}}^{N} \omega\delta_{i,j} \mid \beta_{i,j} \mid\Big) \qquad (3)$$

If an edge from gene i to gene j is predicted by the prior-knowledge then $\delta_{i,j} = 1$, otherwise $\delta_{i,j} = 0$.

The *Least Angle Regression* by Efron et al. [35] effectively implements the LASSO. To compute the network inference, we used the R package *lars* [36] and adapted the functions “lars” and “cv.lars” according to the changes of Zou [34] to use adaptive LASSO.

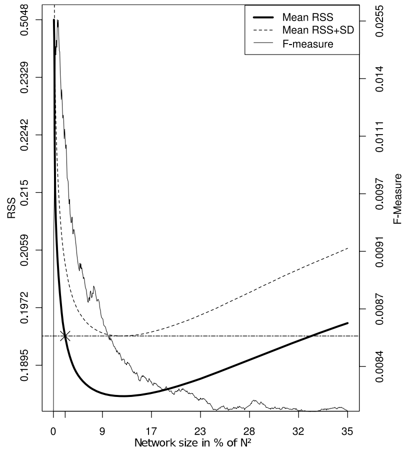

Fig. 1: Optimising the network size for BIND prior-knowledge.

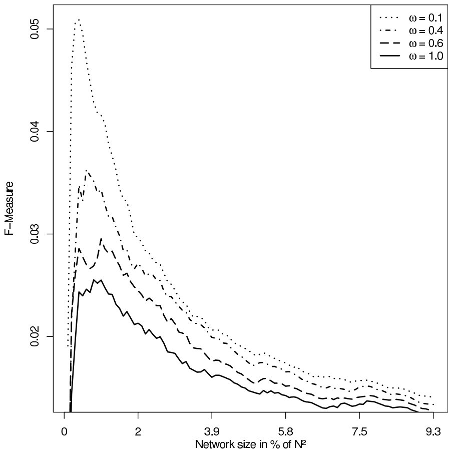

Fig. 2: Optimising the influence of the prior-knowledge for BIND prior-knowlede.

The computation of the coefficients for a gene is time consuming, but at the same time independent from the computation of other genes, since the linear equation system is uncoupled. This offers the poissibility to speed up the computation time of the algorithm by using parallel computing. In order to perform parallel computing, we distributed the calculation for all $\beta_{i,.}$ to different CPUs working in parallel.

Results and Discussion

Model Parameters: The proposed modelling approach contains two global modelling parameters: ω and λ. Since each inference takes long computation time, we worked on a test gene expression data consisting only of the goldgenes (see table I).

The parameter λ determines the influence of the shrinkage factor in 3 and thus correspondents with the predicted network size (number of edges). We sampled λ values corresponding to network sizes of $1 - 10$ % of a fully connected network. We performed 10 fold cross-validation and calculated the residual sum of squares (RSS, see first term in equation 2

TABLE II: Performance comparison. The table shows different quality meassures for different network inference methods in combination with prior-knowledge sources.

	LASSO	PPI	BIND	TRANS	FAC	CLR	ARACNE	MR
#(Interactions)	748042	749069	749735	749379	748219	15686064	39986	15329450
FPR	0.00669	0.0067	0.0067	0.0067	0.00669	0.23405	0.00035	0.22514
Precision	0.00005	0.00007	0.00007	0.00007	0.00004	0.00003	0.00045	0.00003
Recall	0.03605	0.05252	0.05252	0.05046	0.03605	0.46289	0.01856	0.46082
F-Measure	0.00009	0.00014	0.00014	0.00013	0.00009	0.00006	0.00087	0.00006
AUC(ROC)	0.513	0.5127	0.502	0.51908	0.5004	0.50816	0.47814	0.52258

and 3) for all sampled λ values. Figure 1 visualizes the mean RSS, its standard deviation (SD), as well as the F-Measure (see performance evaluation) for different network sizes. After reaching a minimum, the mean RSS is growing with a larger network size, which indicates over-fitting. On the other hand, the F-Measure is going down after reaching a maximum indicating a worse performance quality. Following the experience of Hecker *et al* [6], we chose λ in that way to select the most parsimonious model within 1 SD of the mean RSS minimum. As figure 1) visualizes this choice (marked with a cross) corresponds to good performance values. We achieved very similar results using different prior-knowledges sources.

For the optimal choice of ω, we tested different values in $[0.1, 0.9]$. As figure 2 indicates we achieved best results when choosing a value of 0.1. This result is independent of the choice of the network size and the prior-knowledge source.

A. Performance evaluation

We calculated eight full-genomic networks. One is based on LASSO without prior-knowledge, four on adaptive LASSO integrating different prior-knowledge sources (see Table I) and three networks which apply inference based on mutual information: context likelihood of relatedness (CLR) [1], Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) [3], and Mutual information relevance networks (MR) [4]. In order to evaluate the quality of the different predicted networks we compared the results to the mined gold-standard and calculated a number of performance measurements: false positive rate (FPR), precision, recall, F-Measure and area under the receiver-operating characteristics curve (AUC(ROC), calculated following Stolovitzky *et al* [37]). Table II summarizes quality measures for the full-genomic networks. There are large differences in the number of interactions. While CLR and MR predict very large networks, LASSO is more sparse and ARACNE predicts a very small network. While large networks have a high recall value, their precision value is rather small. Contrary, small networks have a higher precision but smaller recall. F-measure and AUC(ROC) combine precision and recall. Using the AUC(ROC) as criterion the large network predicted by MR performs best, followed by networks based on LASSO. Our network inference approach uses a very small cutoff for the network size, so that the majority of predicted interactions for the AUC(ROC) are based on random samples. Contrary, MR predicts a large number of interactions, so that only a few interactions are based on random sampling when calculating the AUC(ROC). The F-Measure on the other hand, favours small networks, giving ARACNE the best value. Networks inferred using LASSO perform worse than ARACNE, but better than MR and CLR. Furthermore, the F-measure indicates an improvement of prediction quality when integrating prior-knowledge during the network inference. Best results are achieved when using BIND and PPI as prior-knowledge. In general, it seems that the performance strongly depends on the predicted number of interactions. This might be explained by a relatively small gold-standard. In future, larger gold-standards might help to improve performance evaluation.

IV. Conclusion

In this study, we predict the first full-genomic interactions network for a human fungal pathogen. The proposed modelling strategy is based on linear regression and softly integrates different prior-knowledge sources. Since, no database of known interactions for *C. albicans* exists, we use a text-mining system which utilizes full-text research papers to identify known regulatory interactions, making this the first study which combines network inference with text-mining. Using this list to compare the performance quality, we can identify optimal values for the modelling parameters determining the influence of the sparseness criterion and the prior-knowledge during the network inference. Finally, we show that the integration of prior-knowledge improves the inference approach. The influence of the prior-knowledge on the performance needs to be studied in more detail. Especially, the combination of different prior-knowledge sources will be in the focus of future studies. The combination of text-mining with network inference may be a future working frame for the prediction of interaction networks for non-model organisms. This may especially help to identify central (hub) genes in networks of pathogenic microorganisms, which are important drug targets. Together with the ever growing number of available expression data for *C. albicans* and the growing amount of known interactions, researchers will be able to improve the network inference results, thus making *C. albicans* to a model organisms for fungal infections.

Acknowledgment

JL was supported by the excellence graduate school Jena School for Microbial Communication (JSMC).

References

[1] J. J. Faith *et al.*, “Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles.” *PLoS Biol*, vol. 5, no. 1, p. e8, Jan 2007. [Online]. Available: http://dx.doi.org/10.1371/journal.pbio.0050008

[2] R. Opgen-Rhein and K. Strimmer, "From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data." *BMC Syst Biol*, vol. 1, p. 37, 2007. [Online]. Available: http://dx.doi.org/10.1186/1752-0509-1-37

[3] A. A. Margolin *et al.*, "Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context." *BMC Bioinformatics*, vol. 7 Suppl 1, p. S7, 2006. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-7-S1-S7

[4] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements." *Pac Symp Biocomput*, pp. 418–429, 2000.

[5] M. Gustafsson, M. Hörnquist, and A. Lombardi, "Constructing and analyzing a large-scale gene-to-gene regulatory network–lasso-constrained inference and biological validation." *IEEE/ACM Trans Comput Biol Bioinform*, vol. 2, no. 3, pp. 254–261, 2005. [Online]. Available: http://dx.doi.org/10.1109/TCBB.2005.35

[6] M. Hecker *et al.*, "Integrative modeling of transcriptional regulation in response to antirheumatic therapy." *BMC Bioinformatics*, vol. 10, p. 262, 2009. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-10-262

[7] M. Gustafsson, M. Hörnquist, J. Björkegren, and J. Tegnr, "Soft integration of data for reverse engineering," in *International Conference on Systems Biology,2008*, 2008, pp. 127–127.

[8] J. Linde, D. Wilson, B. Hube, and R. Guthke, "Regulatory network modelling of iron acquisition by a fungal pathogen in contact with epithelial cells." *BMC Syst Biol*, vol. 4, no. 1, p. 148, 2010. [Online]. Available: http://dx.doi.org/10.1186/1752-0509-4-148

[9] H. Yoon *et al.*, "Coordinated regulation of virulence during systemic infection of salmonella enterica serovar typhimurium." *PLoS Pathog*, vol. 5, no. 2, p. e1000306, Feb 2009. [Online]. Available: http://dx.doi.org/10.1371/journal.ppat.1000306

[10] R. Guthke *et al.*, "Discovery of gene regulatory networks in aspergillus fumigatus ." *Lect Notes Bioinf*, vol. 4366, pp. 22–41, 2007.

[11] Y.-C. Wang *et al.*, "Global screening of potential candida albicans biofilm-related transcription factors via network comparison." *BMC Bioinformatics*, vol. 11, p. 53, 2010. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-11-53

[12] A. M. Huerta, H. Salgado, D. Thieffry, and J. Collado-Vides, "RegulonDB: a database on transcriptional regulation in Escherichia coli," *Nucleic Acids Res.*, vol. 26, no. 1, pp. 55–59, 1998.

[13] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: Ncbi gene expression and hybridization array data repository." *Nucleic Acids Res*, vol. 30, no. 1, pp. 207–210, Jan 2002.

[14] L. S. Wilson *et al.*, "The direct cost and incidence of systemic fungal infections." *Value Health*, vol. 5, no. 1, pp. 26–34, 2002.

[15] B. Hube, "From commensal to pathogen: stage- and tissue-specific gene expression of candida albicans." *Curr Opin Microbiol*, vol. 7, no. 4, pp. 336–341, Aug 2004. [Online]. Available: http://dx.doi.org/10.1016/j.mib.2004.06.003

[16] K. Zakikhany *et al.*, "In vivo transcript profiling of candida albicans identifies a gene essential for interepithelial dissemination." *Cell Microbiol*, vol. 9, no. 12, pp. 2938–2954, Dec 2007. [Online]. Available: http://dx.doi.org/10.1111/j.1462-5822.2007.01009.x

[17] W. A. Baumgartner(Jr.) *et al.*, "Manual curation is not sufficient for annotation of genomic databases." in *ISMB/ECCB (Supplement of Bioinformatics)*, 2007, pp. 41–48.

[18] U. Hahn, J. Wermter, R. Blasczyk, and P. A. Horn, "Text mining: Powering the database revolution (correspondence)," *Nature*, vol. 448, no. 7150, p. 130, 2007.

[19] L. Hirschman, A. S. Yeh, C. Blaschke, and A. Valencia, "Overview of biocreative: Critical assessment of information extraction for biology," *BMC Bioinformatics*, vol. 6, no. Supplement 1: S1, 2005.

[20] J.-D. Kim *et al.*, "Overview of BioNLP'09 Shared Task on Event Extraction," in *Proceedings BioNLP 2009. Companion Volume: Shared Task on Event Extraction*. Boulder, Colorado, USA, June 4-5, 2009, 2009, pp. 1–9.

[21] C. Rodríguez-Penagos, H. Salgado, I. Martínez-Flores, and J. Collado-Vides, "Automatic reconstruction of a bacterial regulatory network using natural language processing," *BMC Bioinformatics*, vol. 8, no. 293, 2007. [Online]. Available: http://www.biomedcentral.com/1471-2105/8/293

[22] U. Hahn *et al.*, "How feasible and robust is the automatic extraction of gene regulation events? a cross-method evaluation under lab and real-life conditions," in *Proceedings of the NAACL workshop on BioNLP 2009*. Association for Computational Linguistics, 2009, pp. 37–45.

[23] D. Albrecht, O. Kniemeyer, A. A. Brakhage, and R. Guthke, "Missing values in gel-based proteomics." *Proteomics*, vol. 10, no. 6, pp. 1202–1211, Mar 2010. [Online]. Available: http://dx.doi.org/10.1002/pmic.200800576

[24] T. Hastie *et al.*, "Imputing missing data for gene expression arrays," 1999.

[25] W. Stacklies *et al.*, "pcamethods–a bioconductor package providing pca methods for incomplete data." *Bioinformatics*, vol. 23, no. 9, pp. 1164–1167, May 2007. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btm069

[26] M. B. Arnaud *et al.*, "Candida genome database," http://www.candidagenome.org/.

[27] U. Güldener *et al.*, "MPact: the MIPS protein interaction resource on yeast," *Nucleic Acids Research*, vol. 34, no. Database issue, pp. D436–441, Jan. 2006, PMID: 16381906. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/16381906

[28] G. D. Bader, D. Betel, and C. W. V. Hogue, "Bind: the biomolecular interaction network database." *Nucleic Acids Res*, vol. 31, no. 1, pp. 248–250, Jan 2003.

[29] S. Balaji *et al.*, "Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast." *J Mol Biol*, vol. 360, no. 1, pp. 213–227, Jun 2006. [Online]. Available: http://dx.doi.org/10.1016/j.jmb.2006.04.029

[30] E. Wingender, P. Dietze, H. Karas, and R. Knüppel, "TRANSFAC: a database on transcription factors and their DNA binding sites," *Nucleic Acids Research*, vol. 24, no. 1, pp. 238 –241, Jan. 1996. [Online]. Available: http://nar.oxfordjournals.org/content/24/1/238.abstract

[31] E. Buyko, E. Faessler, J. Wermter, and U. Hahn, "Syntactic simplification and semantic enrichment - Trimming dependency graphs for event extraction," *Computational Intelligence, in print*, 2011.

[32] E. Buyko, E. Faessler, J. Wermter, and U. Hahn, "Event extraction from trimmed dependency graphs," in *Proceedings BioNLP 2009. Companion Volume: Shared Task on Event Extraction*. Boulder, Colorado, USA, June 4-5, 2009, 2009, pp. 19–27.

[33] R. Tibshirani, "Regression shrinkage and selection via the lasso," *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 58, pp. 267—288, 1994. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.35.7574

[34] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, pp. 1418–1429, December 2006. [Online]. Available: http://ideas.repec.org/a/bes/jnlasa/v101y2006p1418-1429.html

[35] I. J. B Efron, T Hastie and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[36] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," 2009. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.149.3333

[37] G. Stolovitzky, D. Monroe, and A. Califano, "Dialogue on reverse-engineering assessment and methods: the dream of high-throughput pathway inference." *Ann N Y Acad Sci*, vol. 1115, pp. 1–22, Dec 2007. [Online]. Available: http://dx.doi.org/10.1196/annals.1407.021