A Software Architecture for *de Novo* Induction of Regulatory Networks from Expression Data

Frank RÜGHEIMER[12], Ashish ANAND[12] and Benno SCHWIKOWSKI[12]

[1] Institut Pasteur, Laboratoire de Biologie Systémique, Dept Génomes et Génétique, F-75015 Paris, France
frueghei@pasteur.fr, anand.ashish@gmail.com, benno@pasteur.fr
[2] CNRS, URA2171, F-75015 Paris, France

Keywords systems biology, regulatory networks, structure learning, bioinformatics tools

1 Introduction

Understanding the regulatory networks in cells to explain or even predict phenotypical effects has been one of the key goals of systems biology. Although extensive information is available for specific, well understood systems, that knowledge still covers only a fraction of the expected global regulatory interactions. Unfortunately, the huge number of potential regulatory structures limits purely computational approaches to the network learning task, as even significant numbers of large-scale transcriptomics experiments do not supply sufficient data to discriminate between all alternative solutions.

Here we present a modular architecture for the induction of biologically meaningful regulatory networks from quantitative data and their iterative refinement by new experiments (Section 2). By integrating the experiment selection with the computional framework we aim to maximize the benefits of conducted experiments with respect to their impact on searching the space of regulatory hypotheses. The approach is supported by tools that identify hypotheses consistent with observed data and propose experiments to further assess them.

An application of the proposed approach is presented in Section 3. In that application we use a set genes with potential regulatory function as identified by existing annotations. Similar annotations, such as those using the Gene Ontology standard [1], are now available for an increasing number of model organisms.

2 Global Search Strategy

The proposed strategy (Fig. 1) iterates between three phases: the induction of connectivity scores for individual edges from data (I), a search for likely regulatory hypotheses and the selection of suitable experiments (II), and the assessment of these hypotheses in direct experiments (III). During the first phase pairwise interaction measures are used to compute a connectivity score that assesses the plausibility of pertubations being transmitted along edges between pairs of regulators and targets. The second phase integrates these local evaluations into a global hypotheses of regulatory paths form the origin of a pertubation to the effectors. From superpositions of high scoring pathway hypotheses it is then possible to identify critical experiments that allow to distinguish between large subsets of hypotheses and test important putative links. Both phase (I) and phase (II) allow for the integration of external data via the selected connectivity measures and scoring functions. Results of phase (III) are fed back to the next interation, e.g. by penalizing links that failed validation. The same method can used to draw on prior knowledge from public databases if so desired.

3 Application Example and Software

We applied our approach to transcriptome data collected from *bacillus subtilis* grown in liquid culture. In these experiments a pertubation was caused by the addition of malate to a glucose-based medium. Expression change was subsequently detected in pathways previously thought to be unconnected to carbon metabolism.

First we identified genes involved in carbon metabolism, genetic regulation and affected pathways using a recently released functional annotations available via http://www.subtiwiki.uni-goettingen.de

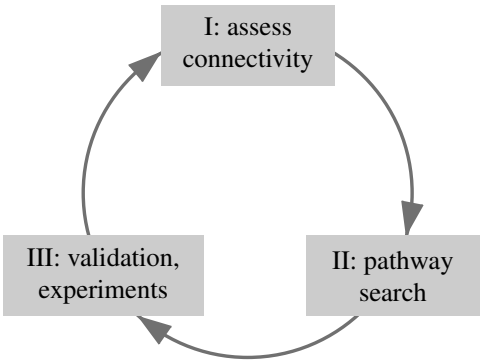

Figure 1. Three phase architecture for regulatory network induction

gene	score	gene	score	gene	score
glcR	0.006311	hxlR	0.006262	rapH	0.006081
comGA	0.006311	manP	0.006229	ybdK	0.005989
cggR	0.006311	yydG	0.006147	phrH	0.005803
kinD	0.006294	yydF	0.006147	yydI	0.005739
ccpA	0.006294	phoR	0.006147	comK	0.005711

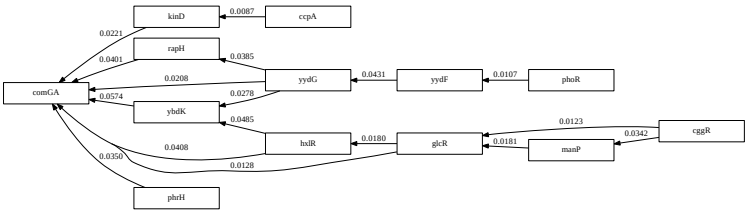

Figure 2. 15 highest scoring nodes and superposition of top 10 pathways for nutrient shift problem (extracted from 409 genes)

For the assessment of link connectivity (phase I) we use the GENIE3 algorithm [2] to induce edge weights from transcriptome data. For the second phase we implemented a path search strategy in the scoreKO command line tool. ScoreKO reads a list of weighted edges in a table format and can be configured to either directly report node assessments or to produce pathway reconstructions. The latter mode reports superpositions of all regulatory pathways up to a specified quality rank allowing to visualize critical players in the selection of regulatory hypotheses. Supplementary scripts that convert the program output for visualization and further processing in Cytoscape [3] are included in the source code archive.

The edge weights induced by GENIE3 were aggregated into pathway scores using the Hamacher product. Our implementation draws on the monotonicity property of that operator for efficient search. This monotonicity property is a natural requirement for any conjunctive aggregation function. Depending on the interpretation of edge weights the operator can be replaced, e.g. by other t-norms. As of version 1.2 our tool also supports the minimum and product as pathway scoring operators.

Results of the second phase (Figure 2) including proposed experiments are assessed in an ongoing collaboration with the Medical Microbiology group at the University of Groningen.

The command line-based pathway search and network induction program we developed has been made available via the the website `http://www.ruegheimer.org/scoreKO`. A complementary tool named findGenes (available from our software website `http://proteomics.fr/Sysbio/Software`) calculates interaction measures, which serve as input for scoreKO, from expression data. We are planning to provide plug-in versions of both tools for the upcoming version 3.0 of the Cytoscape software.

Acknowledgements

This work was supported by a grant of the European Union (FP6, BaSysBio, grant LSHG-CT-2006-037469)

References

[1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[2] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel and P. Geurts, Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE*, 5(9), 2010.

[3] M. S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, N. Landys, C. Workman, R. Christmas, I. Avila-Campilo, M. Creech, B. Gross, K. Hanspers, R. Isserlin, R. Kelley, S. Killcoyne, S. Lotia, S. Maere, J. Morris, K. Ono, V. Pavlovic, A. R. Pico, A. Vailaya, P.-L. Wang, A. Adler, B. R. Conklin, L. Hood, M. Kuiper, C. Sander, I. Schmulevich, B. Schwikowski, G. J. Warner, T. Ideker and G. D. Bader, Integration of biological networks and gene expression data using cytoscape. *Nat Protoc*, 2(10):2366–82, 2007, URL `http://www.ncbi.nlm.nih.gov/pubmed/17947979`.