BIOINFORMATICS

Vol. 20 no. 10 2004, pages 1517–1521
doi:10.1093/bioinformatics/bth112

New methods for joint analysis of biological networks and expression data

Florian Sohler[1,*], Daniel Hanisch[2] and Ralf Zimmer[1]

[1]Institut für Informatik, Ludwig-Maximilians-Universität München, Amalienstraße 17, 80333 München, Germany and [2]Institute for Algorithms and Scientific Computing (SCAI), Fraunhofer Gesellschaft, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

Received on October 12, 2003; accepted on February 3, 2004

ABSTRACT

Summary: Biological networks, such as protein interaction, regulatory or metabolic networks, derived from public databases, biological experiments or text mining can be useful for the analysis of high-throughput experimental data. We present two algorithms embedded in the ToPNet application that show promising performance in analyzing expression data in the context of such networks. First, the Significant Area Search algorithm detects subnetworks consisting of significantly regulated genes. These subnetworks often provide hints on which biological processes are affected in the measured conditions. Second, Pathway Queries allow detection of networks including molecules that are not necessarily significantly regulated, such as transcription factors or signaling proteins. Moreover, using these queries, the user can formulate biological hypotheses and check their validity with respect to experimental data. All resulting networks and pathways can be explored further using the interactive analysis tools provided by ToPNet program.

Contact: florian.sohler@ifi.lmu.edu

INTRODUCTION

As gene expression measurements are still one of the most promising approaches to high-throughput elucidation of biological processes and pathomechanisms of diseases, the analysis of these data receives considerable attention. In this paper, we will focus on interactive analysis of expression data in the context of biological networks and functional annotations. Incorporation of this additional biological knowledge into analysis methods enables researchers to assess quickly the functional context and the relevant interaction partners of significantly regulated genes. This context might be neglected when relying on results from expression measurements only. Most methods incorporate biological annotations after processing the gene expression data alone (e.g. using a clustering procedure). For example, Robinson *et al*. (2002) implemented a Web-based tool for statistical evaluation of cluster results according to functional categorizations, such as gene ontology (GO; The Gene Ontology Consortium, 2001). Methods for integration of biological network and expression data have also been suggested. For example, Zien *et al*. (2000) and Hanisch *et al*. (2002) demonstrated methods for detection of co-regulated metabolic subnetworks. Ideker *et al*. (2002) presented a method for identifying regulatory mechanisms integrating protein–protein interaction network information and expression data. Recently, Yeang and Jaakola (2003) suggested a method of extracting physical pathways supported by expression measurements.

Here, we demonstrate two new algorithms implemented in the ToPNet program, namely Significant Area Search and Pathway Queries, for contextual analysis of expression data. Both methods aim at detecting subnetworks relevant with respect to experimental data. Whereas the first algorithm identifies significantly regulated subnetworks, the second one additionally incorporates user-specified constraints in order to generate biologically more plausible hypotheses.

THE ToPNet FRAMEWORK

ToPNet (http://www.biosolveit.de/topnet/) is a tool for handling several biological networks from multiple sources. Each network has several associated properties (e.g. color, size and hyperlinks of nodes and edges) that can be linked to annotation data, such as a matrix of gene expression data. The glue between networks and annotations is provided by a third data type, the mappings.

ToPNet is able to import networks from various sources (e.g. network databases), which can then be explored interactively using the algorithms provided, edited manually and stored for later inspection.

Data maps handle annotation data in ToPNet by providing standardized information about their content. For example, expression data are often available in a tabular format where rows represent genes and columns correspond to specific experimental conditions. The table itself might contain e.g. probability values quantifying differential expression, or fold changes of expression. A corresponding data map then provides information for a gradient color coding or tooltip annotation of corresponding vertices. As a second example,

[*]To whom correspondence should be addressed.

terms from the GO (The Gene Ontology Consortium, 2001) can be treated as a data map in ToPNet, thereby associating a set of GO terms and corresponding hyperlinks with each vertex.

To connect annotation data to network properties, a mapping is essential. As several major gene and protein databases exist and a general nomenclature for protein and gene names is still missing, ToPNet is able to load mappings for different sets of identifiers interactively and visualize the results. If necessary, ToPNet automatically generates transitive mappings by computing shortest paths in the graph of all mappings.

ALGORITHMS AND DATA ANALYSIS

To analyze expression data in the context of biological networks, we developed several algorithms that require user interaction to various degrees. To demonstrate the capabilities of these algorithms, we selected the yeast compendium dataset (Hughes *et al*., 2000). In this work, 300 mutations and chemical treatments in *Saccharomyces cerevisiae* were analyzed using expression profiling techniques. To assess the degree of differential expression in a certain condition, an error model of expression in the wildtype has been developed and calibrated by Hughes *et al*. (2000). Using this model, probability values quantifying differential expression are available for each gene in each experimental condition.

In this case study, we focus on two gene knockout experiments (HPT1, FUS3/KSS1) and construct corresponding data maps associating each of the approximately 6000 measured open reading frames (ORFs) with the p-value of differential expression provided. It has been found previously that HPT1 mutations affect the expression of purine biosynthesis genes in yeast (Guetsova *et al*., 1997) and that Fus3 and Kss1 are mitogen activated protein (MAP) kinases involved in the pheromone response pathway. Both are activated by Ste7 and, in turn, activate the transcription factor Ste12 (Tedford *et al*., 1997; Bardwell *et al*., 1996).

For analysis of the expression dataset with ToPNet, we use three different networks: The DIP protein interaction network contains experimentally determined interactions (Xenarios *et al*., 2000). The genome-wide location analysis of Lee *et al*. (2002) provides data for a transcriptional regulatory network. In this network, an edge between a transcription factor and a gene exists if the transcription factor binds to the upstream region of that gene with a p-value less than 0.01. Furthermore, we compute a literature network using the name recognition procedure of Hanisch *et al*. (2003). We search for gene names in all PubMed abstracts (NIH, 2000, http://www.ncbi.nlm.nih.gov/entrez/) and construct an edge between two genes if both co-occur in one abstract. The resulting network contains 6262 genes and approximately 42 000 edges.

Hulls, paths and queries

For interactive exploration of the data, gene sets can be selected according to user-defined criteria. These criteria are specified via boolean functions defined on data maps. For example, given that probability values and GO annotations are available, the following expression would select all apoptosis-related genes with a significant p-value: GO biological process like apoptosis & pValueMap ≤ 0.05.

Selected gene sets can be visualized as a network or further extended by graph operations. These operations include computing hulls around genes, i.e. exploring the neighborhood of genes or computing all shortest paths among selected molecules. In conjunction, the selection and manipulation options provide the basis for efficient interactive exploration of the gene expression data.

From the HPT1 knockout data, we select the 11 most significantly regulated genes in the literature network (Fig. 1a), eight of which belong to the purine biosynthesis pathway and form a connected component. The other three genes (BDH1, SAM3 and YBL098w) are unconnected and have functions that appear unrelated. Further investigation of these genes might be worthwhile, but there is also a chance that they are false positives. In contrast, the evidence that the purine biosynthesis is affected by the knockout of HPT1 is striking.

Significant Area Search

The Significant Area Search algorithm aims at detecting connected parts of the network that are significant according to specified p-values. These might correspond to co-regulated pathways in metabolic networks or functionally related proteins in literature networks. The algorithm selects a set of seed genes according to a specified threshold and starts a greedy expansion by including the most significant neighboring molecule in each step. The significance of the selected gene set is quantified by combination of individual p-values using Fisher's inverse χ^2 method (Fisher, 1932), which quantifies the probability that all individual values result from their respective null distributions. The individual p-values are adjusted for greedy selection based on local graph topology. This avoids the construction of subnetworks that are connected only via unspecific high-degree nodes. The detected significant areas are collected and pruned for highly overlapping redundant graphs. The resulting graphs are reported to the user in order of decreasing significance for further interactive exploration.

Applying this algorithm to the HPT1 knockout experiment on the literature network yields only one significant area containing 11 genes from the purine biosynthesis pathway (Fig. 1b). Thus, we can identify three more regulated genes from the purine biosynthesis pathway, namely MTD1, ADE5,7 and ADE12, that are not found by simply selecting the most significantly regulated genes. These genes further support the hypothesis that the purine biosynthesis pathway is affected. On the other hand, the chances that the unconnected three genes with significant p-values are true positives is further diminished since there are no more

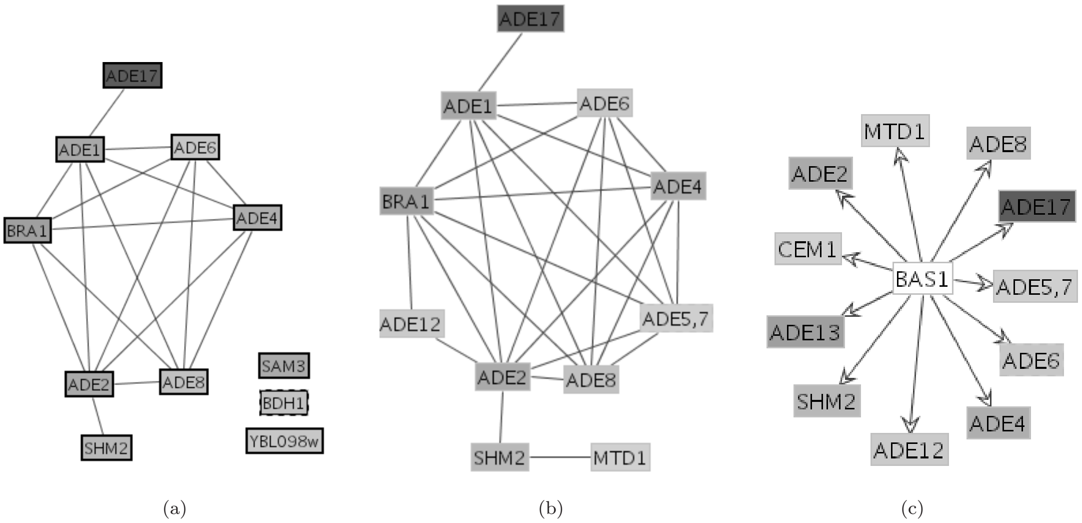


Fig. 1. Examples of significantly regulated subgraphs of the yeast literature network and the regulatory network according to the HPT1 experiment. Genes are shaded according to significance of differential expression. (**a**), The 11 most significantly regulated genes of the network are shown. (**b**) A network found by Significant Area Search using the same data is depicted. Using Significant Area Search on the regulatory network derived from Lee *et al*. (2002), we find the subgraph shown in (**c**).

regulated genes in their neighborhood within the literature network; otherwise they would also form a significant area. In the regulatory network, a significant area containing almost the same set of genes connected by the transcription factor Bas1 is found (Fig. 1c). Bas1 is a known regulator of the purine biosynthesis pathway (Daignan-Fornier and Fink, 1992).

Based on the FUS3/KSS1 knockout, Significant Area Search provides interesting subgraphs of the text mining and regulatory network as well (Fig. 1b and c). The text mining graph contains significantly downregulated genes essential in filamentous growth (e.g. TEC1 and FIG1) and mating pheromone response (e.g. FUS1, SST2 and STE12). This is consistent with findings that the knockout genes Fus3 and Kss1 are required for regulation of filamentation in conjunction with TEC1 (Zeitlinger *et al*., 2003). Moreover, both genes are known to regulate the pheromone response pathway (Tedford *et al*., 1997). Using the regulatory network, the transcription factor STE12 and to a lesser degree MCM1 can be identified as being involved in regulation of differentally expressed genes. This has been described in the literature as well (Kirkman-Correia *et al*., 1993).

Pathway Query Language and Pathway Search

The main goal of ToPNet is to assist in the identification of pathways or subnetworks that are interesting with respect to experimental data. Users from different areas of application will have specific restrictions as to what they consider interesting. For example, in pharmaceutical research focus may be on pathways containing ‘druggable’ targets like kinases or phosphatases. A kinase could be considered interesting only if it phosphorylates a transcription factor that regulates genes that show a significant change in their expression pattern in a certain experiment.

To allow for such complex queries, we have developed an XML-based query language. In this language, pathway templates can be formulated as graph-like structures where vertices describe properties of the genes or proteins (e.g. must be a kinase or a transcription factor) and edges pose restrictions on the connections (e.g. the path between the kinase and the transcription factor must not be longer than 2 or it must involve a phosphorylation).

The template is first transformed into an instance graph by finding all instances of the template nodes in the underlying network and connecting them if all conditions on the paths between them are met. With that preprocessing, a maximal clique in the instance graph that contains instances of all template nodes is a valid realization of the pathway. To find all pathway instances, we enumerate all maximal cliques in the instance graph. This is a computationally hard problem (known to be NP-hard), but exploiting the special structure of the instance graph, most practical queries can be computed within a few seconds.

Figure 2 (left-hand-side) shows a simple pathway template, that specifies a network containing a transcription factor connected to one or more genes that are significantly regulated (p-value < 0.005). In the yeast literature network, we find three instances of that template with at least three regulated genes. These instances are merged into a single graph and shown in Figure 2 (right-hand side). They contain 11 regulated genes from the purine synthesis pathway and the transcription factors BAS1, PHO2/BAS2 and GCN4, which are known to regulate this pathway (Daignan-Fornier and Fink, 1992; Rolfes and Hinnebusch, 1993).

```
<?xml version=”1.0”?>
<pw:TheNet xmlns:pw=
      ”http://bio.informatik.uni−muenchen.de/Pathways”
  name=”Transcription_Factor_+_Regulatee”>
  <pw:Subnet name=”Transcription_Factor”>
    <pw:Node name=”Transcription_Factor”>
      GO: molecular_function like Transcription Factor
    </pw:Node>
  </pw:Subnet>
  <pw:Subnet name=”Regulatee” multiple=”true”>
    <pw:Node name=”Regulatee”>
      pvals &lt; 5e−3
    </pw:Node>
  </pw:Subnet>
  <pw:Connection minEdges=”1” maxEdges=”1”>
    <pw:ConnectFrom>Transcription Factor</pw:ConnectFrom>
    <pw:ConnectTo>Regulatee</pw:ConnectTo>
  </pw:Connection>
</pw:TheNet>
```

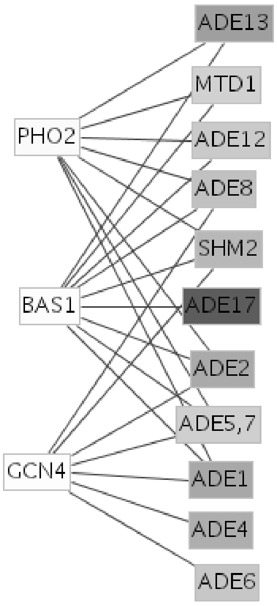
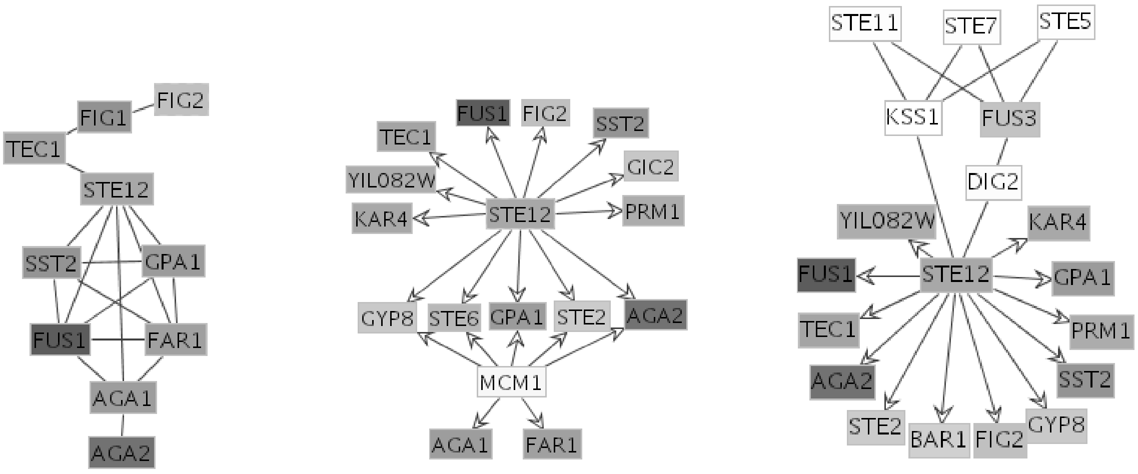


Fig. 2. An example of a pathway query. The template on the left describes a subgraph that consists of a transcription factor and all connected genes that are significantly regulated. On the right, the three most significant results obtained on the text mining network merged into a single graph are shown.



(a) (b) (c)

Fig. 3. Effects of the knockout of FUS3 and KSS1. In (**a**) and (**b**) the two subgraphs show the results of a Significant Area Search conducted on the literature network and the regulatory network, respectively. (**c**) The result of a specific Pathway Query on a combination of the DIP and regulatory networks. Details of the Pathway Query are described in the text.

To demonstrate the flexibility of pathway templates, we apply a more complex query to the FUS3/KSS1 knockout data. We look for kinases that are directly connected to both Fus3 and Kss1. Fus3 and Kss1 must be connected via at most one additional protein to a transcription factor that regulates genes that are differentially expressed in the knockout experiment. Thus, the underlying network must contain protein interactions as well as gene regulations by transcription factor binding. To obtain such a network, we merge the DIP- and the regulatory network and apply the query to the resulting graph. The only matching template instance (Fig. 3c) strongly resembles the downstream part of the pheromone response signaling pathway as it can be found, e.g. in CYGD (Mewes *et al.*, 2002).

SUMMARY AND FUTURE WORK

The presented algorithms embedded in the ToPNet application show promising performance in analyzing expression data in the context of biological networks. Here, we demonstrated their merits in analyzing experiments of the yeast compendium dataset in the context of different networks. Significant Area Search detects subnetworks consisting of significantly regulated genes. These networks often provide hints as to which biological processes are affected in the measured conditions. The Pathway Queries allow detection of networks including molecules that are not necessarily significantly regulated, such as transcription factors or kinases. Moreover, the user can formulate biological hypotheses and check their validity with respect to experimental data. All resulting networks and pathways can be explored further using the interactive analysis tools provided by ToPNet.

In the future, we plan to construct a library of biologically meaningful pathway templates. It might even be possible to learn such templates from databases of known pathways using machine learning techniques. Furthermore, we would like to incorporate ideas from genetic network reconstruction algorithms to enable the discovery of potential pathways that are not completely present in the network sources.

ACKNOWLEDGEMENTS

We thank the Aventis bioinformatics group for encouraging feedback. Part of this work was supported by Aventis Pharma, Frankfurt (Project BEX) and the BMBF (Project BOA).

REFERENCES

Bardwell,L., Cook,J., Chang,E., Cairns,B. and Thorner,J. (1996) Signaling in the yeast pheromone response pathway: specific and high-affinity interaction of the mitogen-activated protein (MAP) kinases Kss1 and Fus3 with the upstream MAP kinase kinase Ste7. *Mol. Cell Biol.*, **16**, 3637–3650.

Daignan-Fornier,B. and Fink,G. (1992) Coregulation of purine and histidine biosynthesis by the transcriptional activators BAS1 and BAS2. *Proc. Natl Acad. Sci., USA*, **89**, 6746–6750.

Fisher,R. (1932) *Statistical Methods for Research Workers*, 4th edn. Oliver and Boyd, London.

Guetsova,M.L., Lecoq,K. and Daignan-Fornier,B. (1997) The isolation and characterization of *Saccharomyces cerevisiae* mutants that constitutively express purine biosynthetic genes. *Genetics*, **147**, 383–397.

Hanisch,D., Fluck,J., Mevissen,H.T. and Zimmer,R. (2003) Playing biology's name game: identifying protein names in scientific text. *Pac. Symp. Biocomput.*, 403–414.

Hanisch,D., Zien,A., Zimmer,R. and Lengauer,T. (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics*, **18** (Suppl. 1), S145–S154.

Hughes,T.R., Marton,M.J., Jones,A.R., Roberts,C.J., Stoughton,R., Armour,C.D., Bennett,H.A., Coffey,E., Dai,H., He,Y.D. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Ideker,T., Ozier,O., Schwikowski,B. and Siegel,A.F. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18** (Suppl. 1), S233–S240.

Kirkman-Correia,C., Stroke,I. and Fields,S. (1993) Functional domains of the yeast Ste12 protein, a pheromone-responsive transcriptional activator. *Mol. Cell. Biol.*, **13**, 3765–3772.

Lee,T., Rinaldi,N., Robert,F., Odom,D., Bar-Joseph,Z., Gerber,G., Hannett,N., Harbison,C., Thompson,C., Simon,I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Mewes,H.W., Frishman,D., Guldener,U., Mannhaupt,G., Mayer,K., Mokrejs,M., Morgenstern,B., Munsterkotter,M., Rudd,S. and Weil,B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.

NIH (2000) Pubmed–national library of medicine.

Robinson,M.D., Grigull,J., Mohammad,N. and Hughes,T.R. (2002) Funspec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, **3**, 35.

Rolfes,R.J. and Hinnebusch,A.G. (1993) Translation of the yeast transcriptional activator GCN4 is stimulated by purine limitation: implications for activation of the protein kinase GCN2. *Mol. Cell Biol.*, **13**, 5099–5111.

Tedford,K., Kim,S., Sa,D., Stevens,K. and Tyers,M. (1997) Regulation of the mating pheromone and invasive growth responses in yeast by two MAP kinase substrates. *Curr. Biol.*, **7**, 228–238.

The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425.

Xenarios,I., Rice,D.W., Salwinski,L., Baron,M.K., Marcotte,E.M. and Eisenberg,D. (2000) DIP: the database of interacting proteins. *Nucleic Acids Res.*, **28**, 289–291.

Yeang,C.-H. and Jaakola,T. (2003) Physical network models and multi-source data integration. *The Seventh Annual International Conference on Research in Computational Molecular Biology* (RECOMB 2003). Berlin, April 10–13, ACM Press, NY, USA.

Zeitlinger,J., Simon,I., Harbison,C., Hannett,N., Fink,T.V.G. and Young,R. (2003) Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell*, **113**, 395–404.

Zien,A., Küffner,R., Zimmer,R. and Lengauer,T. (2000) Analysis of gene expression data with pathway scores. In Altman,R. *et al.* (eds), *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*. AAAI, La Jolla, CA, pp. 407–417.