A GO-Based Method for Assessing the Biological Plausibility of Regulatory Hypotheses

Jonas Gamalielsson, Patric Nilsson, and Björn Olsson

Systems Biology Group, University of Skövde,
Box 407, 541 28 Skövde, Sweden
jonas.gamalielsson@his.se, patric.nilsson@his.se,
bjorn.olsson@his.se

Abstract. Many algorithms have been proposed for deriving regulatory networks from microarray gene expression data. The performance of such algorithms is often measured by how well the resulting network can recreate the gene expression data that it was derived from. However, this kind of performance does not necessarily mean that the regulatory hypotheses in the network are biologically plausible. We therefore propose a method for assessing the biological plausibility of regulatory hypotheses using prior knowledge in the form of regulatory pathway databases and Gene Ontology-based annotation of gene products. A set of templates is derived by generalising from known interactions to typical properties of interacting gene product pairs. By searching for matches in this set of templates, the plausibility of regulatory hypotheses can be assessed. We evaluate to what degree the collection of templates can separate true from false positive interactions, and we illustrate the practical use of the method by applying it to an example network reconstruction problem.

1 Introduction

It is highly desirable to be able to derive causal gene regulatory networks using gene expression data. This is known as reverse engineering of genetic networks[1]. Time series expression data is often used, and the task of the reverse engineering algorithm is to find a set of activation rules that fits these data. Methods for reverse engineering of genetic networks that have been proposed include techniques such as boolean [2, 3], neural [4, 5] and Bayesian networks [6, 7, 8]. In most cases, however, many different reverse engineered networks are consistent with the observed data, but we can expect only a few of these networks to be biologically plausible. A drawback of reverse engineering methods which are based solely on fit to the data is that they do not provide any way of distinguishing between biologically plausible and implausible networks. Therefore, we here propose a method for assessing the biological plausibility of regulatory hypotheses using prior biological knowledge in the form of Gene Ontology (GO) annotation of gene products and examples of regulatory interactions from pathway databases. Using GO, we derive templates encoding general knowledge by generalising from examples of known interactions to typical properties of interacting gene product pairs. By matching regulatory hypotheses to such templates, their plausibility

V.N. Alexandrov et al. (Eds.): ICCS 2006, Part II, LNCS 3992, pp. 879–886, 2006.

can be assessed. The assumption is that if the properties of the gene products in a regulatory hypothesis are similar to those of gene products which are already known to interact, this increases the plausibility of the hypothesis.

Gene Ontology [9] is a structured vocabulary of molecular biology. It contains three different sub-ontologies covering the molecular functions, biological processes and cellular components of gene products, and is structured as a directed acyclic graph showing how terms are related to each other, using inheritance (IS-A) and aggregation (PART-OF). Using these relations, abstraction hierarchies of terms with different specificity are created. A gene product can be associated with several terms in each sub-ontology. One important use of an ontology is the calculation of semantic similarity between two terms. Different information theoretic measures have been proposed and applied for this purpose [10, 11, 12] and ideas from these measures are used in our method.

The concept of templates is appealing for identifying hypothetical regulatory relations that are similar to known regulatory relations. Templates have previously been used in the context of association rule discovery and expression data in [13], where a template-language was designed that allows users to group, filter and inspect a large number of rules produced by association rule discovery algorithms. Of particular interest for our work are rule templates $F1 \rightarrow F2$ which detect rules where the antecedent part contains any of the genes in a predefined functional group $F1$ and the consequent part contains a gene from another predefined group $F2$.

The basic idea behind the method proposed here is to use what we know about regulation in documented pathways, generalise this knowledge, and apply it for assessment of the plausibility of regulatory hypotheses. The GO molecular function classification of the gene products participating in regulatory relations in pathways is used to derive templates. The templates encode the types of gene products known to be involved in a particular type of regulatory relation.

2 Methods and Algorithms

The three main steps of the method are: (1) GO term probability calculation, (2) binary relation extraction, and (3) template derivation.

Step 1: GO term probability calculation. The GO term probability calculation is done as specified in [14] using annotations of the *S. cerevisiae* Genome Database (SGD) available at the GO website (http://www.geneontology.org, September 2005). The term probability calculation is defined as follows, where D_A is a GO annotation database for gene products (in this case SGD):

- For each gene product $GP_i \in D_A$:
  - Increment a counter CNT_j for each GO term GT_j in the annotation of GP_i and increment the counter of each ancestor term of GT_j.
- For each GO term GT_k:
  - Calculate the term probability $p(GT_k) = \frac{CNT_k}{N_A}$, where N_A is the total number of annotations in D_A.

The term probabilities indicate how specific, or common, different GO terms are, and they are used as a measure of specificity of relation templates. The specificity increases with decreasing term probability.

Step 2: Binary relation extraction. Each relation between protein complexes is decomposed into a set of binary relations between pairs of gene products. These binary relations are subsequently used to derive templates. In the following, D_P is a regulatory pathway database, C_i and C_j refer to complexes containing sets of gene products, and rel is a specific relation type, e.g. "expression":

- For each complex relation $C_i[rel]C_j \in D_P$:
  - Create $|C_i| \cdot |C_j|$ binary relations $GP_k[rel]GP_l$ by combining each gene product in C_i with each gene product in C_j, and add all new relations to a set MOD of model relations.

Step 3: Template derivation. Templates representing generalised knowledge of regulatory relations are derived. For each template a GO-score GS is calculated, based on how common the two GO terms appearing in the template are in the annotation database. This is formalised as the average information content of the template terms, using standard information theoretic principles. Templates are derived by the following procedure, where GID represents a GO identifier:

- For each binary relation $GP_k[rel]GP_l \in MOD$:
  - Create templates $GID_i[rel]GID_j$ where $GID_i \in S_k$={GP_k terms with ancestors} and $GID_j \in S_l$={GP_l terms with ancestors}, yielding $|S_k| \cdot |S_l|$ templates. Add these templates to a set T.
- For each template $GID_i[rel]GID_j \in T$:
  $GS = -log_2((p(GID_i) + p(GID_j))/2)$.

Using this information-theoretic definition of GS we get $GS \in [0, -log_2(1/N_A)]$, where N_A is the total number of annotations in D_A. In order to facilitate interpretation of GS values and to make it easier to set a threshold for which hypotheses to consider plausible we also calculate an "expect" value E for each hypothesis H_i according to:

$$E(H_i) = |H| \cdot \frac{|\{t : t \in T \wedge GS(t) \geq GS(H_i)\}|}{|T|} \tag{1}$$

where H is the hypothesis set and $E(H_i)$ the expected number of template hits for H_i with $GS(t) \geq GS(H_i)$ if templates are randomly drawn from T.

We refer to templates based only on annotation level terms (terms directly associated with the gene products in D_A) as "basic templates", while templates containing GO terms of higher abstraction levels are referred to as "variant templates". When initially testing the approach using the *S. cerevisiae* cell-cycle pathway relations from KEGG as input, the method derived 68 basic templates, all with GO-scores in the interval [4.84,11.29], and 1236 variant templates with GO-scores in the interval [0,10.97]. 54% of these variant templates had a GS lower than the lower bound for the basic templates.

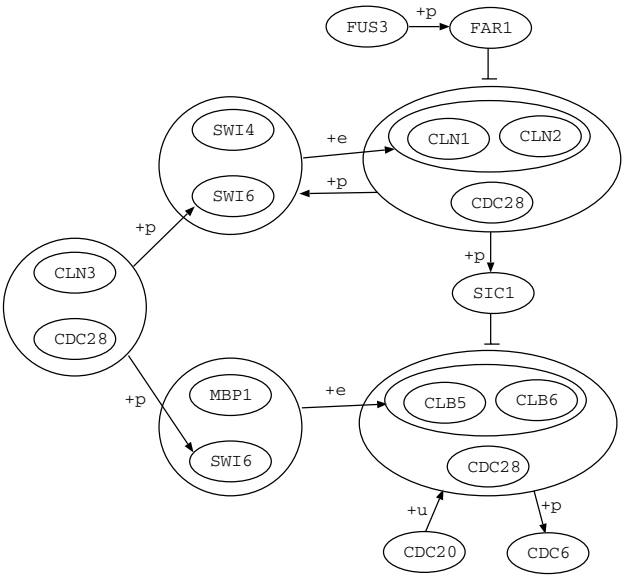

Fig. 1. Part of the cell-cycle regulatory network of *S. cerevisiae* (adapted from [7])

As an example to illustrate the process of template derivation, we here consider the complex-to-complex relation {SWI4,SWI6} [+e] {CLN1,CLN2} in the *S. cerevisiae* cell cycle pathway (figure 1). This relation between gene product complexes is decomposed into the four binary relations "SWI4 [+e] CLN1", "SWI4 [+e] CLN2", "SWI6 [+e] CLN1" and "SWI6 [+e] CLN2". Each of these are used for template derivation. When "SWI4 [+e] CLN1" is used, each of the terms in the GO subgraph of SWI4 in figure 2 are combined with those for CLN1 to create templates. The basic template in this case would be "GO:0003700 [+e] GO:0016538", with a GO-score of $-log_2((0.010+0.003)/2)) = 7.27$. Variant templates are for example "GO:0003700 [+e] GO:0019887" and "GO:0030528 [+e] GO:0019207", with GO-scores $-log_2((0.010+0.005)/2)) = 7.06$ and $-log_2((0.052+0.005)/2))=5.13$, respectively. With six terms for SWI4 and five terms for CLN1, there is a total of 30 templates for this relation (one basic and 29 variant templates).

The derived set of templates is used in the assessment of hypothetical regulatory relations derived using e.g. reverse engineering algorithms. A hypothetical relation (or hypothesis for short) is defined as a directed relation between two gene products. There can be different relation types, e.g. "expression" and "phosphorylation". Each hypothesis $GPh_k[rel]GPh_l$ from a set H of regulatory hypotheses is assessed by the templates using the following procedure:

- For each template $GID_i[rel]GID_j \in T$:
  - Report template match if $GID_i \in \{GPh_k$ terms with ancestors$\}$ and $GID_j \in \{GPh_l$ terms with ancestors$\}$.

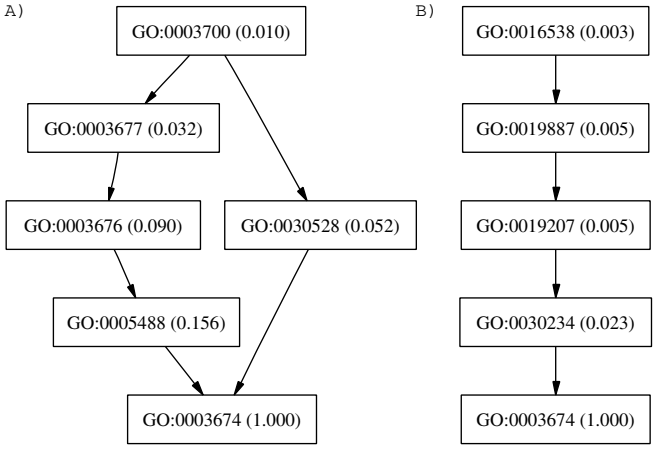

Fig. 2. GO subgraphs for SWI4 (A) and CLN1 (B). GO terms are represented by nodes with probabilities in brackets. Directed edges represent IS-A relationships. A) SWI4: GO:0003700 = transcription factor activity, 0003677 = DNA binding, 0003676 = nucleic acid binding, 0005488 = binding, 0003674 = molecular function, 0030528 = transcription regulator activity. B) CLN1: 0016538 = cyclin-dependent kinase regulator activity, 0019887 = protein kinase regulator activity, 0019207 = kinase regulator activity, 0030234 = enzyme regulator activity.

This procedure results in a list of templates conforming to the hypothesis, ranked in GO-score order. Subsequently, all hypotheses in H can be ranked in GO-score order according to the highest-scoring template of each hypothesis.

3 Results

This section presents experiments done to assess the performance and to show the practical use of the proposed method for plausibility assessment. Out of the 12 types of regulatory relations supported by KEGG, we selected "activation", "inhibition", "expression", "phosphorylation" and "dephosphorylation". The only relation representing transcriptional regulation is "expression", while the other four represent post-translational regulatory mechanisms.

In the evaluation, we study sets of hypothetical relations derived by applying a dynamic Bayesian network technique to gene expression data measuring the expression of genes during the *S. cerevisiae* cell cycle. Hypotheses derived in [15] are shown together with their best-scoring templates in table 1. Each hypothesis in the table is classified into one of four different categories of conformance with respect to the target network in figure 1: correct, transitive, misdirected and incorrect [7]. The hypothetical relations are limited to cover only the part of the cell cycle shown in figure 1. Relations from the whole *S. cerevisiae* cell cycle pathway were used to derive the template set.

Table 1. Regulatory hypotheses and best matching templates. *Hypothesis*: a directed hypothetical relation between two gene products, derived in ref. [15]. *Template*: GO term identifier for the left hand side of the template, followed by type of relation in square brackets (E = "expression", P = "phosphorylation", I = "inhibition"), and GO term identifier for the right hand side of the template. *Sc*: GO-score of the template. *E*: the expected number of hits with a GO-score greater than or equal to the matching template's GO-score (see eq. 1). *Cl*: hypothesis class, coded as in ref. [7] (C = correct edge with respect to the known model, T = transitive edge which skips exactly one gene product, M = misdirected edge, I = incorrect edge). The left half of the table shows hypotheses with $E < 1$ and the right half those with $E \geq 1$. Note that all five correct hypotheses have $E < 1$. In addition, correct hypotheses had the lowest average E, followed by transitive hypotheses, while mis-directed and incorrect hypotheses had substantially higher average E (C: 0.62, T: 0.96, I: 3.07, M: 3.48).

Hypothesis	**Template**	**Sc**	**E**	**Cl**		**Hypothesis**	**Template**	**Sc**	**E**	**Cl**
FUS3→SIC1	4707[P]19210	10.7	0.07	I		CLN1→CLB6	19887[I]16538	7.97	1.29	T
CDC28→CDC6	4693[P]3689	9.97	0.18	C		CLB6→CLN2	19887[I]16538	7.97	1.29	I
CLN1→SIC1	16538[P]19210	9.04	0.47	C		CLN1→CLN3	19887[I]16538	7.97	1.29	I
SIC1→CLN2	19210[I]16538	9.04	0.47	M		CLB6→CLN1	19887[I]16538	7.97	1.29	I
CLN2→FAR1	16538[P]19210	9.04	0.47	M		CLN2→SWI4	16538[P]3700	7.24	2.36	M
CLB5→CDC6	16538[P]3689	8.97	0.52	C		FUS3→SWI4	4674[P]3700	6.57	3.59	I
CLN3→CDC6	16538[P]3689	8.97	0.52	I		CDC28→FUS3	4674[E]4674	6.51	3.68	I
FAR1→FAR1	4861[I]19887	8.59	0.59	I		CDC20→FAR1	30234[P]19210	6.38	4.03	I
FAR1→SIC1	19887[P]19210	8.53	0.63	T		CDC20→CLN2	30234[I]16538	6.23	4.60	I
FUS3→CLN3	4707[P]19887	8.53	0.63	I		CDC6→CLB6	3677[E]16538	5.82	7.07	M
CLB5→CDC28	19887[I]4693	8.33	0.96	C		CDC6→CLB5	3677[E]16538	5.82	7.07	M
CDC28→CLB5	4693[P]19207	8.24	0.99	C		SWI6→SWI6	30528[E]5515	4.26	12.9	I

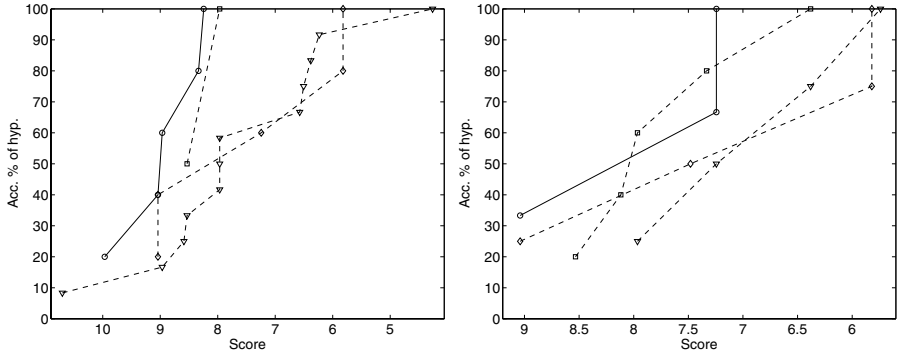

Fig. 3. Percentage of accumulated hypotheses as a function of GO-score for different hypothesis classes for the relations derived in ref. [15] (left) and ref.[7] (right). Circles = correct, squares = transitive, diamonds = misdirected, triangles = incorrect.

As can be seen in table 1, all correct hypotheses have $E < 1$, and the average E of the class of correct hypotheses is higher than for any of the other classes. This suggests that the approach is useful for reducing the number of false positives from a set of derived regulatory hypotheses by setting an E-value threshold

for their plausibility. The results are visualised in figure 3, showing the accumulation of hypotheses from different classes as a function of GO-score. It can be noted that the curve representing the class of "correct" hypotheses reaches 100% before any other class, while the class "incorrect" is the last to reach this level. This suggests that templates and their GO-scores are useful in the evaluation of hypothesis sets derived from data mining. For another hypothesis set, derived in [7], the diagram at the right in figure 3 shows that the results are similar.

4 Discussion

We proposed a method based on the generalisation of knowledge of regulatory pathways for assessing the biological plausibility of hypotheses derived during regulatory network reconstruction. Our results demonstrate that the method is able to filter out a large proportion of implausible hypotheses, thus improving the specificity of the regulatory network reconstruction process.

Gene products in some hypotheses are identical or very similar to gene products in relations used for template derivation, with respect to GO functional annotation. This means that the hypotheses are highly feasible, according to our current general knowledge, while at the same time current specific knowledge would suggest that they are incorrect. It is possible that a future more fine-grained GO annotation will allow us to better discriminate between such hypotheses. It is also possible that future experiments show that some of these hypothetical relations actually exist. A way to get higher accuracy for our method could be to incorporate other sources of biological knowledge, such as transcription factor binding site information or protein interactions.

Our method decomposes relations between gene product complexes into atomic binary relations between individual gene products. This leads to certain problems, as can be observed e.g. for the CLN3-CDC28 complex in figure 1. In order to form this complex, CLN3 will bind to the CDC28 kinase and the complex as a unit will phosphorylate SWI6. Hence, it is not entirely true that CLN3 and CDC28 individually phosphorylate SWI6 as our method suggests. We plan to extend our method so that complex relations can be handled without being decomposed into binary relations.

Our method automatically identifies the biologically most plausible hypotheses by a measure based on GO term specificity, but it would also be beneficial to have the top scoring hypotheses assessed by a domain expert in order to reduce the number of false positives.

References

1. D'haeseleer, P., Liang, S., Somogyi, R.: Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics **16** (2000) 707–726
2. Liang, S., Fuhrman, S., Somogyi, R.: REVEAL - a general reverse engineering algorithm for inference of genetic network architectures. Pac Symp Biocomput **3** (1998) 18–29

3. Akutsu, T., Miyano, S., Kuhara, S.: Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Pac Symp Biocomput **4** (1999) 17–28
4. D'haeseleer, P., Wen, X., Fuhrman, S., Somogyi, R.: Linear modeling of mRNA expression levels during CNS development and injury. Pac Symp Biocomput **4** (1999) 41–52
5. Weaver, D. C., Workman, C. T., Stormo, G. D.: Modeling regulatory networks with weight matrices. Pac Symp Biocomput **4** (1999) 112–123
6. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian networks to analyze expression data. RECOMB '00: Proceedings of the 4th Annual International Conference on Computational Molecular Biology (2000) 127–135
7. Kim, S. Y., Imoto, S., Miyano, S.: Inferring gene networks from time series microarray data using dynamic Bayesian networks. Briefings in Bioinformatics **4** (2003) 228–235
8. Husmeier, D.: Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics **19** (2003) 2271–2282
9. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G.: Gene Ontology: tool for the unification of biology. Nature Genetics **25** (2000) 25–29
10. Lin, D.: An information-theoretic definition of similarity. Proceedings of the 15th International Conference on Machine Learning (1998) 296–304
11. Jiang, J. J., Conrath, D. W.: Semantic similarity based on corpus statistics and lexical taxonomy. Proceedings of the 10th International Conference on Research on Computational Linguistics, ROCLING X (1997) 19–33
12. Resnik, P.: Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial Intelligence Research **11** (1999) 95–130
13. Tuzhilin, A., Adomavicius, G.: Handling very large numbers of association rules in the analysis of microarray data. Proceedings of the 8th ACM SIGKDD International Conference on Data Mining and Knowledge Discovery (2002) 396–404
14. Lord, P. W., Stevens, R. D., Brass, A. and Goble, C. A.: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics **19** (2003) 1275–1283
15. Sigursteinsdottir, G.: Learning gene interactions from gene expression data using dynamic Bayesian networks. Master's thesis, HS-IKI-MD-04-202, Univ. of Skövde (2004)