



## Superiority of network motifs over optimal networks and an application to the revelation of gene network evolution

S. Ott<sup>1,\*</sup>, A. Hansen<sup>2</sup>, S.-Y. Kim<sup>1</sup> and S. Miyano<sup>1</sup>

<sup>1</sup>Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, 108-8639, Japan, <sup>2</sup>Wolfson Institute for Biomedical Research, University College London, The Cruciform Building, Gower Street, London, WC1E 6AU, UK

Received on March 26, 2004; revised on August 2, 2004; accepted on August 17, 2004  
Advance Access publication September 17, 2004

### ABSTRACT

**Motivation:** Estimating the network of regulative interactions between genes from gene expression measurements is a major challenge. Recently, we have shown that for gene networks of up to around 35 genes, optimal network models can be computed. However, even optimal gene network models will in general contain false edges, since the expression data will not unambiguously point to a single network.

**Results:** In order to overcome this problem, we present a computational method to enumerate the most likely  $m$  networks and to extract a widely common subgraph (denoted as gene network motif) from these. We apply the method to bacterial gene expression data and extensively compare estimation results to knowledge. Our results reveal that gene network motifs are in significantly better agreement to biological knowledge than optimal network models. We also confirm this observation in a series of estimations using synthetic microarray data and compare estimations by our method with previous estimations for yeast. Furthermore, we use our method to estimate similarities and differences of the gene networks that regulate tryptophan metabolism in two related species and thereby demonstrate the analysis of gene network evolution.

**Availability:** Commercial license negotiable with Gene Networks Inc. (cherkis@gene-networks.com)

**Contact:** sascha-ott@gmx.net

### INTRODUCTION

The estimation of gene networks from expression level measurements has been one focus of bioinformatics research in recent years (Chen *et al.*, 1999; Friedman *et al.*, 2000; Hartemink *et al.*, 2001; Ideker *et al.*, 2002; Rung *et al.*, 2002; van Someren *et al.*, 2002). Knowledge of gene networks will

be important for understanding cellular processes, designing new strategies to combat disease and so on.

A widely used approach to model gene networks are Bayesian networks (Buntine, 1991; Cooper and Herskovits, 1992; Friedman and Goldszmidt, 1998; Heckerman, 1999; Friedman *et al.*, 2000; Hartemink *et al.*, 2001; Pe'er *et al.*, 2001; Imoto *et al.*, 2002; Ong *et al.*, 2002; Tamada *et al.*, 2003; Nariai *et al.*, 2004; Ott *et al.*, 2004), which model expression levels of genes as random variables and gene networks as joint probability distributions of expression patterns. These distributions are decomposed using directed acyclic graphs (DAGs), which we will call networks. Networks are scored using score functions based on the likelihood of networks, given the data.

A recent result shows that one can optimally estimate gene network models for gene networks of up to about 35 genes (Ott *et al.*, 2004).<sup>1</sup> This result holds for any score function  $s : G \times 2^G \rightarrow \mathbb{R}$  assigning a score to a pair of a gene  $g$  and a possible selection of parents for  $g$ . However, while optimal gene network models are the most likely models, there may still be very different models that have approximately the same likelihood, considering the large number of DAGs [in the case of 10 genes  $\approx 4.17 \times 10^{18}$  (Robinson, 1973)]. Therefore, even optimal gene network models will in general not match the target gene network.

Our endeavour to tackle this problem is 3-fold. First, we provide an algorithm for the enumeration of optimal and sub-optimal networks in the order of their likelihood, and extract frequent subgraphs of the most likely  $m$  networks, denoted as gene network motifs in this work.<sup>2</sup> Second, we rigorously and extensively compare estimated network models to available

<sup>1</sup>This boundary of feasibility holds in the biologically relevant case of a limited number of regulators for every gene.

<sup>2</sup>Therefore, our definition of a gene network motif seems to differ from (Milo *et al.*, 2002), but might turn out to be closely related.

\*To whom correspondence should be addressed at Wolfson Institute for Biomedical Research, University College London, The Cruciform Building, Gower Street, London WC1E 6AU, UK.

knowledge about gene networks. Third, we apply the gene network motif extraction to microarray data of *Bacillus subtilis* and *Escherichia coli* in order to demonstrate the analysis of gene network evolution.

Our evaluations show that gene network motifs are in significantly better agreement with knowledge about transcriptional regulation than optimal network models. We also derived the conclusion that the gene networks governing the regulation of tryptophan metabolism in the above species have probably changed significantly during their evolution while other parts of the network have been conserved.

## METHODOLOGY

Throughout this work, we assume a set of genes  $G$  and a score function  $s : G \times 2^G \rightarrow \mathbb{R}$ . The score of a network  $N$  is defined as  $score(N) = \sum_{g \in G} s(g, P^N(g))$ , where  $P^N(g)$  denotes the set of  $g$ 's parents in  $N$ . In order to find the model that explains the given data best, we need to find the DAG  $N$  that minimizes  $score(N)$ .

Since this problem is NP-hard (Chickering, 1996), and the search space is of super-exponential size (Robinson, 1973), heuristic approaches have frequently been applied (Friedman et al., 2000; Hartemink et al., 2001; Pe'er et al., 2001; Imoto et al., 2002; Ong et al., 2002; van Someren et al., 2002; Tamada et al., 2003; Nariai et al., 2004), though the accuracy of heuristics is uncertain. However, we have recently shown that optimal networks can be found using  $(|G|/2 + 1) \cdot 2^{|G|}$  dynamic programming steps that leads to an exact algorithm applicable in all kinds of research settings (Ott et al., 2004).

For larger gene networks, empiric or heuristic methods can be used to select a biologically meaningful subspace of the search space. If this is done as described in Ott and Miyano (2003), the approach of Ott et al. (2004) can be used to perform an optimal search within the selected subspace.

We have extended the algorithm of Ott et al. (2004) in order to solve the enumeration task. Since the likelihood of gene network motifs is the sum of the likelihood of the networks containing it, they will turn out to be more reliable than single network estimations.

### Enumerating optimal gene networks

Without loss of generality, we assume networks with equal score to be sorted in some way, therefore allowing the notion 'the  $k$ -th best network'. For  $m \in \mathbb{N}$ , we use  $\mathcal{N}_{\leq m}$  to denote  $\{1, \dots, m\}$ . An ordering of a set  $A \subseteq G$  can be described as a permutation  $\pi : \{1, \dots, |A|\} \rightarrow A$ . We use  $\Pi^A$  to denote the set of all permutations of  $A$ . For  $\pi \in \Pi^A$ , we say that a network  $N \subseteq A \times A$  is  $\pi$ -linear if for all  $(g, h) \in N$   $\pi^{-1}(g) < \pi^{-1}(h)$  holds, that is, all edges in  $N$  comply with the direction given by  $\pi$ .

Our strategy is to divide the space of DAGs on a set  $A \subseteq G$  into subspaces of  $\pi$ -linear networks, for all  $\pi \in \Pi^A$ , and

to decompose the problem of finding optimal and suboptimal networks into the following two problems:

- (1) Find the subspace of the search space that contains the (sub)optimal network searched for.
- (2) Find the (sub)optimal network within the selected subspace.

We first define some functions and then show how these functions can be computed for gene networks of considerable size.

**DEFINITION 1.** Let  $m \in \mathbb{N}$ . We define  $F^m : G \times 2^G \times \mathcal{N}_{\leq m} \rightarrow 2^G$  inductively.<sup>3</sup> First, for all  $g \in G$  and  $A \subseteq G$ , we define

$$F^m(g, A, 1) = \arg \min_{B \subseteq A} s(g, B).$$

Then, denoting the set of all previous solutions  $\{F^m(g, A, p) \mid p < k\}$  as  $J(k)$ ,

$$F^m(g, A, k) = \arg \min_{\substack{B \subseteq A \\ B \notin J(k)}} s(g, B)$$

for all  $1 < k \leq m$ .

**DEFINITION 2.** Let  $m \in \mathbb{N}$ . We define  $S^m : G \times 2^G \times \mathcal{N}_{\leq m} \rightarrow \mathbb{R}$  as

$$S^m(g, A, k) = s(g, F^m(g, A, k))$$

for all  $g \in G$ ,  $A \subseteq G$ , and  $k \in \mathcal{N}_{\leq m}$ .

By the definitions,  $F^m(g, A, k)$  is the  $k$ -th best choice of parents for a gene  $g$  when the parents have to be selected from  $A$ , and  $S^m(g, A, k)$  is the score for this choice. When  $m$  is clear from the context, we use  $F$  and  $S$  instead of  $F^m$  and  $S^m$ , respectively. Note that  $F^m$  and  $S^m$  are partially defined functions, since  $m$  may be larger than the number of subsets of  $A$ .

In order to find optimal and suboptimal networks for a given subspace specified by a permutation  $\pi$ , we need the following function  $Q^A$ . For a given gene  $g$ , let us denote the set of all genes that precede  $g$  in  $\pi$  as  $V(\pi, g) = \{h \mid \pi^{-1}(h) < \pi^{-1}(g)\}$ .

**DEFINITION 3.** Let  $A \subseteq G$ . We define  $Q^A : \Pi^A \times \mathcal{N}^{|A|} \rightarrow 2^{A \times A}$  as

$$Q^A(\pi, v) = \{(h, g) \mid h \in F(g, V(\pi, g), v_{\pi^{-1}(g)})\}$$

for all  $\pi \in \Pi^A$  and  $v \in \mathcal{N}^{|A|}$ .

In Definition 3, we have used a vector  $v \in \mathcal{N}^{|A|}$  to determine the rank of the selection of parents for the particular genes. Below it will be shown that  $Q^A(\pi, v)$  is the set of edges of

<sup>3</sup>We define  $\mathbb{N}$  as  $\{1, 2, \dots\}$  in this work.

**Table 1.** Intuitive meanings of the functions used to define the algorithm

Function	Functionality	Meaning
$F^m$	$G \times 2^G \times \mathbb{N}_{\leq m} \rightarrow 2^G$	$F^m(g, A, k)$ is the $k$ -th best choice of parents for $g$ from $A$
$S^m$	$G \times 2^G \times \mathbb{N}_{\leq m} \rightarrow \mathbb{R}$	$S^m(g, A, k)$ is the score of the $k$ -th best choice of parents for $g$ from $A$
$Q^A$	$\Pi^A \times \mathbb{N}^{ A } \rightarrow 2^{A \times A}$	$(A, Q^A(\pi, v))$ is a $\pi$ -linear network
$M^m$	$2^G \times \mathbb{N}_{\leq m} \rightarrow \bigcup_{A \subseteq G} \Pi^A$	The $k$ -th best network on $A$ is $M^m(A, k)$ -linear
$D^m$	$2^G \times \mathbb{N}_{\leq m} \rightarrow \bigcup_{i=0}^{ G } \mathbb{N}^i$	The $k$ -th best network on $A$ is $(A, Q^A(M^m(A, k), D^m(A, k)))$

an optimal or suboptimal  $\pi$ -linear network on  $A$ , its rank depending on  $v$ . Next, we define two functions  $M^m$  and  $D^m$  that specify subspaces, in which (sub)optimal networks can be found, and the choice of a network from the subspace, respectively.

**DEFINITION 4.** Let  $m \in \mathbb{N}$ . We inductively define functions  $M^m : 2^G \times \mathbb{N}_{\leq m} \rightarrow \bigcup_{A \subseteq G} \Pi^A$  and  $D^m : 2^G \times \mathbb{N}_{\leq m} \rightarrow \bigcup_{i=0}^{|G|} \mathbb{N}^i$  over their second parameter. Let  $A \subseteq G$ . First, we define

$$D^m(A, 1) = (1, \dots, 1) \in \mathbb{N}^{|A|}$$

and

$$M^m(A, 1) = \arg \min_{\pi \in \Pi^A} \text{score}((A, Q^A(\pi, D^m(A, 1)))).$$

Let  $k \in \mathbb{N}_{\leq m}$  with  $k > 1$  and let  $N$  be a network on  $A$  with optimal score among networks not in  $\{(A, Q^A(M^m(A, p), D^m(A, p))) \mid p < k\}$ . Let  $\pi^* \in \Pi^A$  be a permutation such that  $N$  is  $\pi^*$ -linear. We define

$$M^m(A, k) = \pi^*.$$

Let  $v^* \in \mathbb{N}^{|A|}$  such that for every  $g \in A$ , the set of  $g$ 's parents,  $Pa^N(g)$ , equals

$$F^m(g, V(\pi^*, g), v_{\pi^*^{-1}(g)}^*).^4$$

We define:

$$D^m(A, k) = v^*$$

As  $F^m$  and  $S^m$ ,  $M^m$  and  $D^m$  are partial functions.

The functions and their intuitive meanings are summarized in Table 1, the definition of the algorithm is given in Table 2.

The algorithm computes the functions  $F^m$  and  $S^m$  in Step 1 and in Step 2 for all  $g \in G$ ,  $A \subseteq G$ , and  $j \leq m$ . This can be

<sup>4</sup>Since network  $N$  is among the best  $m$  networks on  $A$ , the choice of parents for each gene  $g$  must also be among the best  $m$  choices.

**Table 2.** The algorithm

Step 1:	Set $F^m(g, \emptyset, 1) = \emptyset$ , $S^m(g, \emptyset, 1) = s(g, \emptyset)$ for all $g \in G$ .
Step 2:	For all $g \in G$ and all $A \subseteq G - \{g\}$ , $A \neq \emptyset$ , do the following two steps for all $j \leq m$ :
Step 2a:	Select $B^* \subseteq A$ that minimizes $s(g, B^*)$ from $\{B \subseteq A \mid B = A \vee B = F^m(g, A - \{h\}, p), h \in A, p \leq m\} - \{F^m(g, A, p) \mid p < j\}$ .
Step 2b:	Set $F^m(g, A, j) = B^*$ , $S^m(g, A, j) = s(g, B^*)$ .
Step 3:	Set $M^m(\emptyset, 1) = \emptyset$ and $D^m(\emptyset, 1) = \emptyset$ .
Step 4:	For all $A \subseteq G$ , $A \neq \emptyset$ , do the following three steps for all $j \leq m$ :
Step 4a:	Choose a triple $(g, p, q) \in A \times \mathbb{N}_{\leq m} \times \mathbb{N}_{\leq m}$ such that $\text{score}((A - \{g\}, Q^{A - \{g\}}(M^m(A - \{g\}, p), D^m(A - \{g\}, p)))) + S^m(g, A - \{g\}, q)$ is minimized and $(g, p, q)$ induces a network different from $(A, Q^A(M^m(A, r), D^m(A, r)))$ for $r < j$ .
Step 4b:	Set $M^m(A, j)(i) = M^m(A - \{g\}, p)(i)$ for all $i <  A $ , and $M^m(A, j)( A ) = g$ .
Step 4c:	Let $v$ denote $D^m(A - \{g\}, p)$ . Set $w \in \mathbb{N}^{ A }$ as $w_i = v_i$ for all $i <  A $ and $w_{ A } = q$ . Set $D^m(A, j) = w$ .
Step 5:	Return $Q^G(M^m(G, i), D^m(G, i))$ for all $i \leq m$ .

done by applying dynamic programming, since only function values of  $F^m$  for a set  $A$  of lower cardinality or lower  $j$  are needed in order to select  $B^*$  in Step 2a.

In Steps 3 and 4, functions  $M^m$  and  $D^m$  can be computed similarly using dynamic programming, since for the selection of a triple in Step 4a only function values of  $M^m$  and  $D^m$  for a set  $A$  of lower cardinality or lower  $j$  are needed. The triple  $(g, p, q)$  specifies a network on  $A$ ,  $g$  being a candidate for the last element in the permutation searched for,  $p$  specifying the remaining permutation that can be chosen from up to  $m$  previously computed permutations of  $A - \{g\}$ . Then, to form a network in the subspace defined by the resulting permutation, the  $q$ -th best selection of parents for  $g$  is used, while for the other genes parents are selected as indicated by  $D^m(A - \{g\}, p)$ .

Functions  $F^m$ ,  $S^m$ ,  $M^m$  and  $D^m$  are partially defined in the case of high  $m$  which we did not explicitly mention in the description of the algorithm. In our implementation of the algorithm, we store permutations  $\pi$  and make use of the restrictions for edges in  $\pi$ -linear networks, yielding a memory- and time-efficient coding of networks. Moreover, the two applications of dynamic programming in Steps 1/2 and Steps 3/4, respectively, can be performed in an alternating way, thus reducing the memory requirements substantially. The number of network comparisons in Step 4a can be minimized in practice, since networks with different scores are different. Moreover, the algorithm can well be parallelized.

### Correctness

Let us denote the  $k$ -th best network on a set  $A \subseteq G$  by  $N_{A,k}^*$ . We first reformulate two lemmata from Ott *et al.* (2004).

LEMMA 1. Let  $A \subseteq G$  and  $\pi \in \Pi^A$ . Let  $N^*$  be a  $\pi$ -linear network on  $A$  with minimal score. Then,  $\text{score}((A, Q^A(\pi, (1, \dots, 1)))) = \text{score}(N^*)$  holds.

LEMMA 2. Let  $A \subseteq G$  and  $m \in \mathbb{N}$ . Let  $g^* = \arg \min_{g \in A} (S^m(g, A - \{g\}, 1) + N_{A - \{g\}, 1}^*)$ . Define  $\pi \in \Pi^A$  by  $\pi(i) = M(A - \{g^*\}, 1)(i)$  for  $i \in \{1, \dots, |A| - 1\}$ , and  $\pi(|A|) = g^*$ . Then,  $\pi = M^m(A, 1)$ .

The following theorem provides the correctness. We regard as one dynamic programming step the computation that is executed for one  $g \in G$  and one  $A \subseteq G$  in Step 2, respectively for one  $A \subseteq G$  in Step 4. We use  $n$  to denote  $|G|$ .

THEOREM 1. Let  $m \in \mathbb{N}$ . The best  $m$  networks can be found using  $(n/2 + 1) \cdot 2^n$  dynamic programming steps, where the complexity of a dynamic programming step depends on  $m$ .

PROOF. By the definitions, the output of the algorithm,  $Q^G(M^m(G, i), D^m(G, i))$ ,  $i \leq m$ , are the best  $m$  networks on  $G$ . We only need to prove that the recursive formulas given in the algorithm are correct. The equations given in Step 1 are correct by the definitions of  $F^m$  and  $S^m$ . When we select a subset of a set  $A \subseteq G$  in Step 2, we have basically two choices: the whole set  $A$  or a true subset. In the former case, we can compute the score of the choice directly, in the latter, we can use previously computed values of  $F^m$  and  $S^m$ , which gives the correctness of Step 2.

After the execution of Step 2, we have computed all values of  $F^m$  and  $S^m$ . Using these values, function  $Q$  can be computed directly. Therefore, we only need to compute functions  $M^m$  and  $D^m$  in order to be able to produce the output in Step 5. The equations in Step 3 are again correct by the definitions. We observe that with Lemma 1 in combination with Definition 4, the following equation follows by induction:

$$N_{A,k}^* = Q^A(M^m(A, k), D^m(A, k)) \quad (1)$$

From this equation and Lemma 2 we see that the recursion in Step 4 is correct for  $k = 1$  (variable  $j$  in the algorithm). For  $k > 1$ , we compute the suboptimal permutation  $M^m(A, k)$  and the suboptimal choice of parents  $D^m(A, k)$  in the same way, restricting to a network not previously chosen.

The dynamic programming in Steps 1 and 2 requires  $2^{n-1}$  steps for each gene, since a gene may not be one of its parents. In the dynamic programming in Steps 3 and 4 a total number of  $2^n$  steps is needed, which completes the proof.

### Taking advantage of a combinatorial explosion

The time required for one dynamic programming step depends on the number of solutions  $m$ , but can be regarded as a feasible constant for  $m \leq 200$ . Here, we show how  $m$  can be chosen as high as 20 000 in practice, with only slightly increasing the computation time compared to  $m = 200$ .

The algorithm as stated in Table 2 computes a fixed number of solutions, regardless of the size of the set  $A$ . However, it is often possible to derive the best  $m$  solutions for a set  $A$  in layer  $|A| = j$  ( $2 \leq j \leq |G|$ ) from the knowledge of the best  $m'$  solutions for sets in layer  $j - 1$  with  $m' < m$ . The number of derivable solutions may vary from no increase in the worst case to a quadratic increase in the best case (Ott, 2004). A quadratic increase in one step means a super-exponential increase of derived solutions with increasing layers. In the practical case, it is unlikely to encounter one of the extreme cases as we validated in computational experiments using microarray data.

We found that the number of derivable solutions usually increases exponentially. This allows us to compute a lower number of solutions  $m'$  for the layers exhibiting a high number of subsets, and to exponentially increase the number of solutions for higher layers of rapidly declining size. This strategy takes advantage of the intrinsic combinatorial explosiveness of the search space and works well for  $m' = 100$  and  $m = 10\,000$ , or  $m' = 200$  and  $m = 20\,000$  in order to compute the best  $m$  networks (Ott, 2004). The observed increase in the number of solutions was by about a factor 1.5 when climbing up one layer.

### Extracting gene network motifs

A straightforward approach to exploit the information given by an enumeration of the most likely  $m$  networks would be to count the occurrences of each edge in the networks, select only the edges which have a count above some threshold, and compose a partial network from these edges.<sup>5</sup> However, a partial network composed from edges of high counts does not have to be frequent in the networks at hand and may, therefore, be unlikely. This leads to the following problem, which we refer to as the gene network motif extraction problem:

Given graphs  $N_1 = (G, E_1), \dots, N_m = (G, E_m)$ , and  $k \in \mathbb{N}$ , find a set  $M \subseteq G \times G$  with  $|M| = k$  maximizing the number of graphs  $N_i$  that include  $M$ , i.e.  $M \subseteq E_i$ .

The motif extraction problem is equivalent to the well-known problem of finding maximal frequent item sets. The problem of finding balanced complete bipartite subgraphs (Garey and Johnson, 1979) can be reduced to both problems, so they are NP-hard. However, the problem can be solved for practical instances as encountered in this work by exhaustively searching over subsets  $M$  of the set of edges with at least  $c$  occurrences ( $c \in \mathbb{N}$ ), using the fact that edges in a motif  $M$  with at least  $c$  occurrences must also have at least  $c$  occurrences.

<sup>5</sup>This approach was used in order to compose partial networks based on networks enumerated by the bootstrap method (Pe'er et al., 2001). It was shown that, assuming independency of edge counts, regions of significantly many edges with high edge counts can be found in gene network estimations. However, this assumption does not account for the fact that the edge counts of all edges connected to the same gene  $g$  depend on the expression measurements of  $g$ .

Using this search strategy, we find the optimal solution  $M^*$  as long as the empirically chosen constant  $c$  is not set too high.

## EVALUATING GENE NETWORK ESTIMATIONS

The principled evaluation of gene network estimations is an important issue which has been rarely addressed in the literature. Since transcription and translation co-localize in prokaryotes, rendering mRNA expression data sufficient for estimating gene networks, we chose *B.subtilis* and *E.coli* as targets for evaluating the proposed methods, using the available microarray data and biological knowledge of both species.

### Data and score function

For *E.coli*, we selected the data sets GDS95–GDS100 from the Gene Expression Omnibus.<sup>6</sup> Changes in gene expression levels were elicited by perturbations of tryptophan metabolism, UV exposure and novobiocin treatment (Khodursky *et al.*, 2000a,b; Courcelle *et al.*, 2001). We also received data of transcriptional regulation from RegulonDB (Salgado *et al.*, 2001) for comparison with estimation results.

For *B.subtilis*, we used a data set of 70 microarrays from time-course experiments under various treatments, a data set of 99 microarrays from gene disruptant experiments (unpublished data) and a data set of known transcriptional regulation from DBTBS (Ishii *et al.*, 2001; Makita *et al.*, 2004).

The score functions ( $s : G \times 2^G \rightarrow \mathbb{R}$ ) used in our implementation include the MDL score (Friedman and Goldszmidt, 1998), the BDe score (Cooper and Herskovits, 1992; Friedman and Goldszmidt, 1998) and the BNRC score (Imoto *et al.*, 2002). We selected the BNRC score for our computations because it can be applied without discretization of the data, avoiding additional parameters and loss of information, and gene interactions are modeled using B-splines, allowing for non-linear relationships. Furthermore, in computational experiments on the *B.subtilis* data set, optimal networks with respect to the BNRC score turned out to be in the best agreement with textbook knowledge among these score functions (Ott, 2004; Sonenshein *et al.*, 2001).

### Selection of target networks

We selected all relations from the knowledge data set for *E.coli*, for which experimental evidence is provided (evidence level three or higher) yielding a set of 899 known relations. From the *B.subtilis* data set, we selected 840 regulatory relations with evidence in the literature. We applied a random procedure to select target networks from these data. Since we need to select genes in a way that there are some known regulatory relations, we select the first few genes randomly, and then iteratively select genes that are connected to the previously selected genes. In each iteration, we randomly select

a connected component of the intermediate network and a new gene with a known relation to at least one gene in the component. Since trivial choices of target networks should be avoided, we choose a gene not connected to the previously selected genes when five connected genes have been selected in a row.

The selection procedure yields a partially known gene network  $N$  of non-trivial structure represented as a matrix. Each pair of genes with (without) a known relation is represented with 1 (0) in the corresponding entry of the matrix, but 0.5 is set instead of 0 for pairs  $(g, h)$  fulfilling at least one of the following conditions:

- (1)  $h$  regulates  $g$ .
- (2) A gene  $i$  in the target network regulates  $g$  and  $h$ .
- (3) A gene  $i$  in the target network is regulated by  $g$  ( $h$ ) and regulates  $h$  ( $g$ ).
- (4) Condition 2 or Condition 3 hold for a gene  $i$  outside the target network.

Using these conditions, nearly correct estimations are distinguished from wrong estimations. If edge  $(g, h)$  is estimated and  $(h, g)$  is a known regulatory relation, then the fact that these two genes interact was correctly estimated (Condition 1). In the same way, indirect relations established by Conditions 2–4 are also not entirely wrong, if estimated.

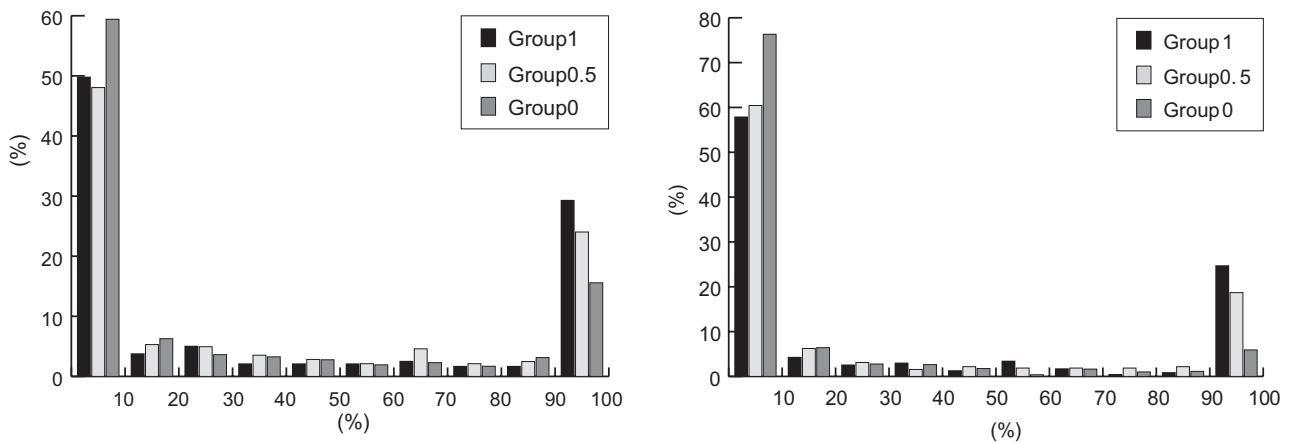
### Evaluating edge counts

The above algorithm allows to enumerate most likely networks. We, therefore, asked whether the number of occurrences (edge count) of a given edge in the most likely  $m$  networks carries biologically relevant information. We applied the procedure described above to select 30 target networks of 10 genes for *B.subtilis* and used the exact algorithm to enumerate the best 500 network models for each set of genes  $G_i$ . For each  $G_i$  ( $i = 1, \dots, 30$ ) and for each possible directed edge  $(g, h)$  or undirected edge  $\{g, h\}$  in a network on  $G_i$ , we counted the number of occurrences in the 500 estimated networks for  $G_i$ . We then examined how each group of edges (0, 0.5 and 1) distributes over the range of edge counts.

The results (Figure 1) show that most edges have an edge count below 50 or above 450 (first/last interval). In these two intervals, group 0 separates clearly from group 0.5 and group 1, whereas group 0.5 and group 1 do not show a clear separation, pointing to the difference between indirectly correct estimations (group 0.5) and estimations that can not be justified from current knowledge. Group 0 shows fewer high edge counts, but dominates the interval of lowest edge counts, indicating that edges with high edge count are more likely to be true than edges with low edge count.

Disruptant data yield a stronger separation of group 0 than time-course data, which might be due to the (different) amount of data rather than to the experimental method. Despite the lack of information about the target networks (since they had

<sup>6</sup><http://www.ncbi.nlm.nih.gov/geo/>.



**Fig. 1.** Result for *B.subtilis* data; x-axis: discretized ratio of edge counts to the number of estimated networks (500), y-axis: proportion of edges of the group 0, 0.5 and 1 per interval. Left: time-course data, Right: disruptant data.

**Table 3.** Evaluation results for the motif extraction approach using 80 and 95 as thresholds

Method	Number of selected edges		Number of correct edges		p-value	
	80	95	80	95	80	95
Edge from directed motif	1000	997	552	581	$2.32 \times 10^{-22}$	$2.32 \times 10^{-31}$
Edge from undirected motif	1000	1000	538	564	$9.05 \times 10^{-19}$	$1.04 \times 10^{-25}$
Edge from optimal network	1000		486		$2.25 \times 10^{-8}$	

been randomly selected, irrespective of the microarray data) in the microarray data and true relations possibly being unknown or estimated as transitive edges, the observed separation of group 0 is an encouraging result.

A computation on the *E.coli* data set did not yield a clear separation of group 0, which might be due to the lower number of microarrays (53) and the low number of affected genes in these specific experiments (see above).

We conclude that the above evaluation scheme can be applied not only for evaluating score functions, but also for evaluating the significance of the data.

### Superiority of gene network motifs

In order to evaluate the motif extraction approach, we selected the disruptant data set for *B.subtilis* and 1000 target gene networks  $N_i$  of 10 genes in a random way as described above, enumerated the most likely 100 network models for each target network, and extracted motifs with two or more edges (highest-scoring motif among largest motifs), at threshold  $c = 80$ , resp. 95. We randomly selected one edge of each optimal network, and one edge of each motif. We then checked the correctness of both edges, using the DBTBS data, and computed the probability  $p_i$  of randomly guessing a single edge from  $N_i$  as the ratio of the number of edges of  $N_i$  to the number of possible edges. According to the results of the previous subsection, we judged 1 entries and 0.5 entries as correct. We computed an upper bound for the probability of

guessing at least  $k$  single edges correctly among  $n$  networks,  $P(n, k)$ , by using the inequality  $P(n, k) \leq \sum_{i=k}^n \binom{n}{i} p^i \times (1-p)^{n-i}$ , where  $p$  denotes  $\max_{i=1}^{1000} p_i$ .<sup>7</sup> We observe that the results (Table 3) for the motif extraction approach as well as for the optimal models are in significant agreement with the knowledge. But the former approach clearly outperforms the results for optimal gene networks. We conclude that gene network motifs are even more reliable than the network with the best score. We note that our results also hold for the estimated networks as a whole, since we selected arbitrary edges for the evaluation.

This result was confirmed in further evaluations for gene networks of 15 genes and in the case of accepting only one entry of the knowledge matrix as correct (Ott, 2004). Examining the optimal number of enumerated networks  $m$ , we found that higher numbers yield better results, if  $m$  is chosen from  $\{1, \dots, 150\}$ .

### Evaluations using synthetic data

In order to evaluate whether the observed higher reliability of network motifs holds in general, we produced synthetic microarray data using the following procedure. First, we selected a DAG of 11 vertices (genes) as an artificial gene

<sup>7</sup>In order to avoid too rough estimations of  $p$ -values, random selection of target networks is repeated when a target network of extreme  $p_i$  was selected. The actual value of  $p$  was about 0.4 in our evaluations.

network. Four genes in our network have no parents, four genes have one parent, two genes have two parents, and one gene has three parents. We then selected a linear or non-linear function for each gene to describe the dependency of its expression values and the expression values of its parents. The expression levels of genes without parents were modelled as normally distributed with standard deviation set to 1. For genes with parents we added a normally distributed system error to the input from its parents. The standard deviation of this system error was set as half the standard deviation of its input. Finally, we added a measurement error of varying standard deviation to the data.

We assessed the accuracy of optimal networks and network motifs by comparing these graphs to the artificial network and computing the sensitivity and the specificity.<sup>8</sup> Figures 2 and 3 summarize the results for varying measurement error. Each data point is an average value of 10 repetitions of data generation and network estimation.

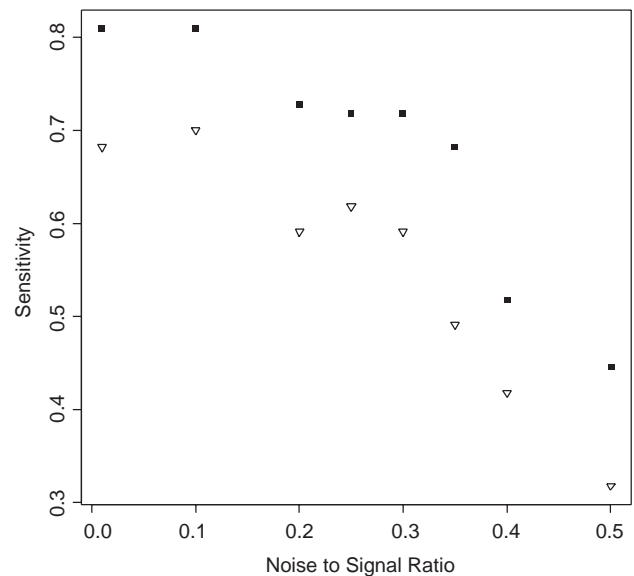
We observe that optimal networks show a higher sensitivity, while the specificity of network motifs outperforms optimal networks. This result is consistent with our above finding on real data and shows that network motifs provide the biologist with more reliable estimations of gene regulation than optimal networks do. The increase in reliability comes at the expense of sensitivity, but reliability might be of higher priority, considering the high complexity of gene network models and the noisy data.

In a second series of estimations we used synthetic data to evaluate the influence of the number of microarrays on the sensitivity and specificity of estimations. Figures 4 and 5 summarize these results, each data point is an average value after 10 repetitions. As expected, the accuracy of estimations increases with an increasing number of arrays, while the observed differences between optimal networks and network motifs do not seem to depend on the number of arrays.

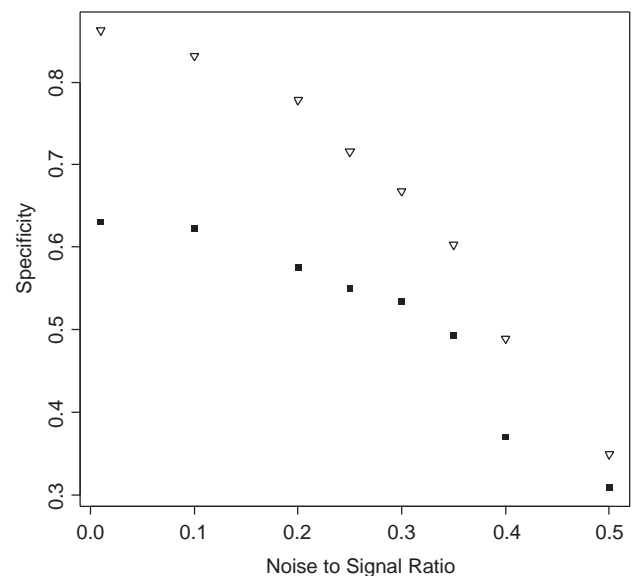
### Application to yeast data

In Friedman *et al.* (2000), the bootstrap method has been applied to assess the confidence of features (e.g. edges) of networks estimated by a heuristic algorithm. In the bootstrap approach, the data set is modified using a random procedure  $m$  times, and a network estimation is done for each modified data set. The confidence of estimated features is then computed as the number of occurrences in  $m$  estimated graphs. In our approach, we do not modify the given data, but instead analyze the subspace of the most likely network structures. Still, both methods are similar in the sense that they aim to distinguish reliable parts of estimations from less reliable ones.

In Friedman *et al.* (2000), gene networks have been estimated for yeast, using the data of Spellman *et al.* (1998). As a



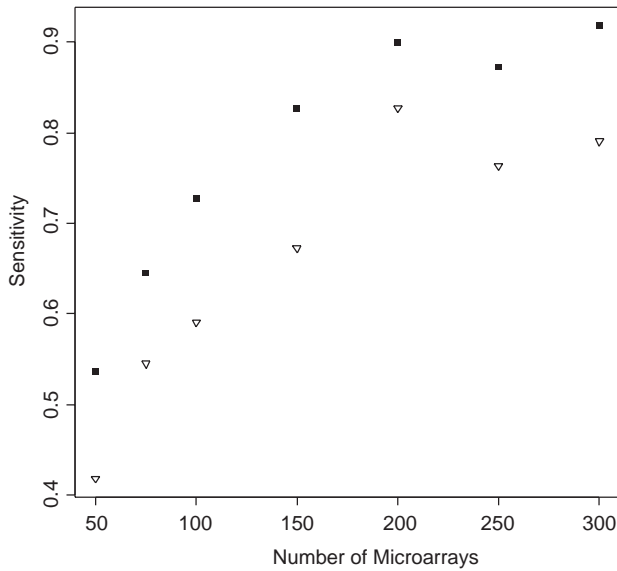
**Fig. 2.** Average sensitivity as a function of noise-to-signal ratio. Rectangles represent average values for optimal networks, triangles represent average values for network motifs.



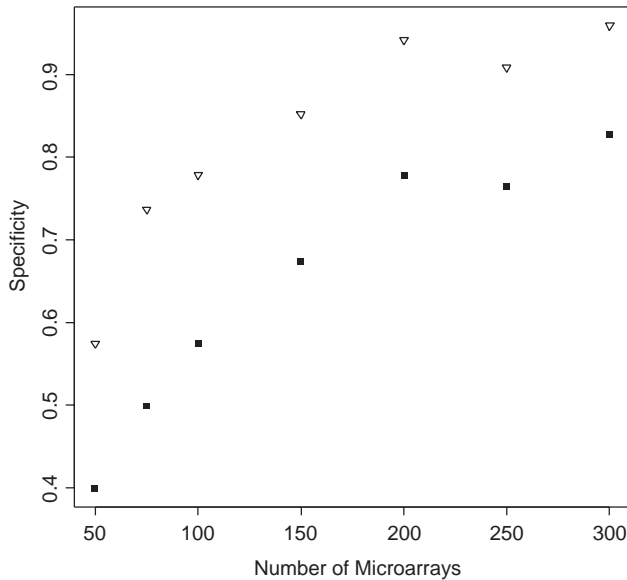
**Fig. 3.** Average specificity as a function of noise-to-signal ratio. Rectangles represent average values for optimal networks, triangles represent average values for network motifs.

comparison of both approaches, we have extracted network motifs for two small networks using the same microarray data. Figures 6 and 7 show the extracted network motifs and regulatory relations estimated in Friedman *et al.* (2000). We observe that out of 13 edges estimated by the bootstrap approach, we find 6 edges in the network motif, 3 edges that are reversed, and 4 edges that do not appear in the network motif. However, in these four cases we find directed paths of length 2 that

<sup>8</sup>Sensitivity is defined as  $TP/(TP+FN)$ , specificity as  $TP/(TP+FP)$ , where TP, FP and FN denote the number of true positives, false positives, and false negatives, respectively.



**Fig. 4.** Average sensitivity as a function of the number of microarrays. Rectangles represent average values for optimal networks, triangles represent average values for network motifs.

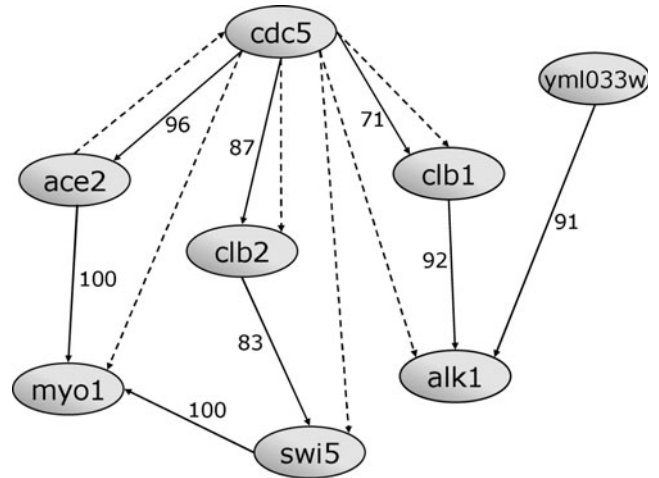


**Fig. 5.** Average specificity as a function of the number of microarrays. Rectangles represent average values for optimal networks, triangles represent average values for network motifs.

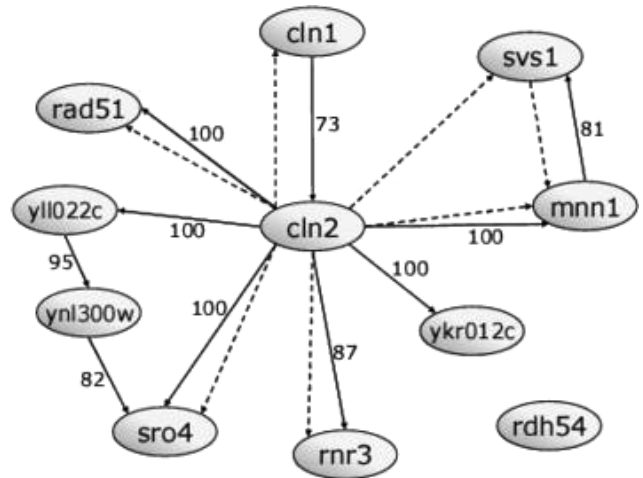
represent the same relationships indirectly. Therefore, both estimation results are quite consistent.

### APPLICATION TO GENE NETWORK EVOLUTION

There are indications that evolution may be driven more strongly by changes in expression levels than changes in



**Fig. 6.** Estimation result for yeast data, subnetwork around *cdc5*. Solid arrows show edges included in the network motif estimated by our method, dashed arrows represent estimations given in Friedman *et al.* (2000), edge counts are included for solid arrows.



**Fig. 7.** Estimation result for yeast data, subnetwork around *cln2*. Solid arrows show edges included in the network motif estimated by our method, dashed arrows represent estimations given in Friedman *et al.* (2000), edge counts are included for solid arrows.

protein structures (Enard *et al.*, 2002; Oleksiak *et al.*, 2002), thus indicating a kind of gene network evolution. The methods described can help to unravel similarities and evolutionary changes in gene regulatory networks of related species. As we did in the above evaluations, we choose Gram-negative rod-shaped *E.coli* and Gram-positive rod-shaped *B.subtilis* as promising targets for our estimations. But instead of using the whole data set described above, we now restrict the data according to the matter of investigation, selecting specific experiments and specific genes only, thus increasing specificity.

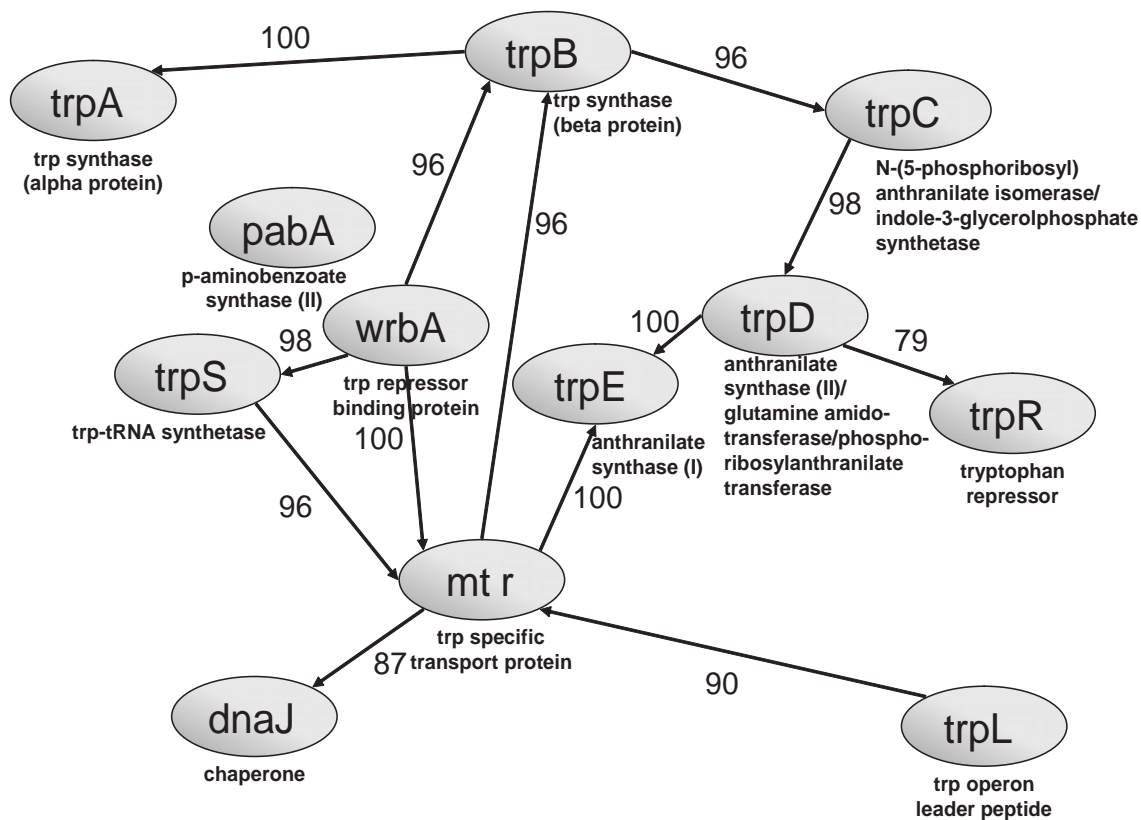


Fig. 8. Extracted motif for *E.coli*.

Given the above data, the most promising target network is the tryptophan network because for *E.coli* there is few (18 microarrays), but highly specific data from experiments designed to unravel details of the regulation of the genes involved in tryptophan metabolism, and for *B.subtilis* there is a comparatively high number of 59 microarrays obtained from time-course experiments under various nutritional conditions, likely to affect the tryptophan network. We selected genes known to be involved in the well-studied tryptophan network and extracted directed motifs based on 100 enumerated optimal network models, using 50 as the threshold. A graph was derived from the output file showing the largest motif with highest score extracted from the enumerated networks, genes are presented by nodes, and each edge is labelled with its weight.

The largest motif obtained from the data set for *E.coli* is found 54 times and contains 13 edges with weights ranging from 79 to 100 (Figure 8), the one for *B.subtilis* data was found 61 times and contains 16 edges, weights ranging from 77 to 100 (Figure 9).

*trpA*, *trpB*, *trpC*, *trpD* and *trpE* are linked by highly weighted edges in both motifs, forming the core of this network, corresponding to the fact that they are positioned close to each other in the *trp*-operon in both species. Even the

order of position in the *trp*-operon can partially be recognized in the graphs. But in *B.subtilis*, seven genes code for the enzymes in the *trp* biosynthesis pathway, as opposed to five in *E.coli*. The two extra genes, *trpF* and *pabA* are also contained in the derived motif. *trpF* is connected to *trpB* and *trpD*, corresponding to its approximate position on the *trp*-operon. *pabA*, also called *trpG* in *B.subtilis*, is found closely connected to *trpA* and *trpB* although it is not located in the *trp*-operon but in the folate operon (Sonenshein *et al.*, 2001). This close connection may indicate the close functional relation in the tryptophan biosynthesis pathway. The *E.coli* *pabA* gene, which has not been found to be deeply involved in this pathway, remains unconnected in the extracted motif.

On the other hand, *mtrB* is closely linked to *trpE* and *trpF* in the graph for *B.subtilis*. This can be explained by the fact that its gene product, tryptophan RNA-binding attenuation protein (TRAP) has been found to be among the key regulators of tryptophan biosynthesis in *B.subtilis* (Valbuzzi and Yanofsky, 2001). Interestingly, *mtr*, coding for a tryptophan specific transport protein in *E.coli* (Ong *et al.*, 2002) is also associated with core genes of tryptophan biosynthesis like *trpB* and *trpE* in the graph. The structure of the network motifs suggests that the function of *mtr* in *E.coli* might be similar to the one of *mtrB* in *B.subtilis*.

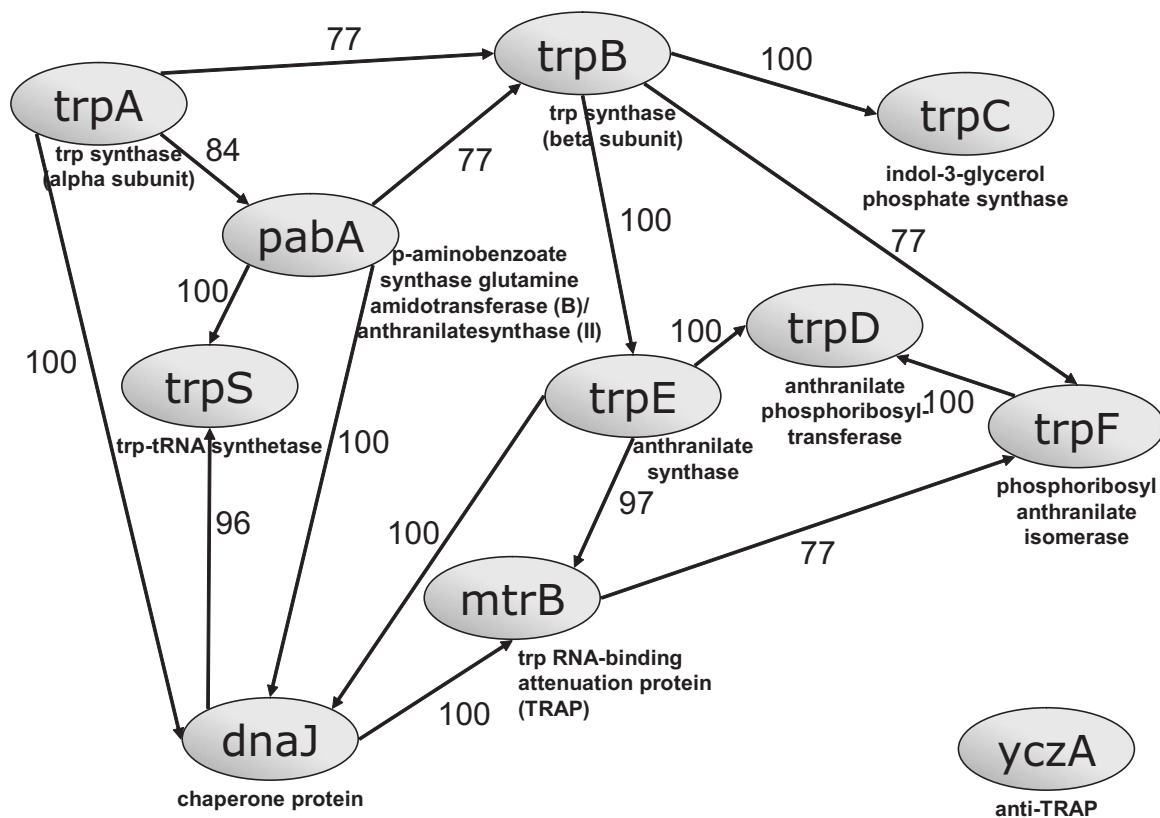


Fig. 9. Extracted motif for *B.subtilis*.

Recently, a TRAP-inhibitory protein (anti-TRAP, AT) has been found and has been identified as the gene product of *yczA* (Valbuzzi and Yanofsky, 2001). However, no association between *yczA* and the other genes examined can be found in the results; but *yczA*-mRNA levels may be low since AT does not act alone but depends on tryptophan levels in the cell and only binds to Trp-activated TRAP (Valbuzzi and Yanofsky, 2001). No homologue of *yczA* has been found so far in *E.coli*, but BLASTp searches revealed that an amino acid sequence of AT shows similarities to that of the cystein-rich domain of the chaperone DnaJ (Szabo *et al.*, 1996; Martinez-Yamout *et al.*, 2000). Therefore, *dnaJ* had been included into this calculation, the results suggest an association with the tryptophan network in *E.coli* and even more in *B.subtilis*.

The reason might be cross-hybridization with *yczA*-mRNA or interaction on the protein level as a chaperone or as a TRAP regulator. This would explain the strong link between *dnaJ* and *mtrB* in the figure for *B.subtilis* (Figure 9). *trpS*, coding for the tryptophanyl-tRNA synthetase has been included into this analysis because tRNA<sup>Trp</sup> had been shown to play an important, yet different, role in the regulation of tryptophan biosynthesis in both bacteria (Valbuzzi and Yanofsky, 2001). The structures of the respective network motifs differ indeed. The figure for *E.coli* (Figure 8) shows *trpS* being linked to *mtr*

and *wrbA*. *mtr* may be involved in transcription termination which can be observed in abundance of tRNA<sup>Trp</sup> (Valbuzzi and Yanofsky, 2001), stalling at the leader peptide, which is encoded by *trpL*, corresponding to the edge from *trpL* to *mtr* in the figure. *wrbA* encodes a tryptophan repressor binding protein, so this edge corresponds to Trp activating the tryptophan repressor. However, there is no direct link between *trpS* and *trpR*, the tryptophan repressor gene. This can be explained by the fact that the tryptophan repressor protein acts through conformational change when Trp binds, not through mRNA expression. Yet the graph shows an association of *trpR* with *trpD*, which may indicate a general production of tryptophan repressor protein together with tryptophan biosynthesis enzymes, though *trpR* is not located in the *trp*-operon. In the *B.subtilis* graph, *trpS* expression is linked to *pabA* and *dnaJ*. If *dnaJ* functions as *yczA*, this is consistent with the finding that tRNA<sup>Trp</sup> has been found to be associated with the formation of AT-inactivated TRAP (Valbuzzi and Yanofsky, 2001).

We conclude that comparing estimated gene networks can be valuable to confirm previous knowledge and to derive new hypotheses. The meaning of edges may vary, ranging from transcriptional regulation to membership in the same operon or more complex interactions, and has, therefore, to be assessed in the light of previous knowledge after gene network

estimation, and the understanding of such estimations has to be developed further. Our method can thus be helpful in both medical and biological applications (e.g. finding candidate target genes).

## DISCUSSION AND CONCLUSION

We have provided a theoretical basis for the enumeration of optimal and suboptimal gene network models for gene networks of considerable size, presented results of a comprehensive comparison of estimations to biological knowledge, showed that gene network motifs are superior to optimal networks and applied the motif extraction approach to the intriguing problem of gene network evolution. We note that our evaluation criterion for estimated networks is very demanding and our work is one among very few that perform a principled evaluation of the estimated networks.<sup>9</sup>

The algorithmic methodology described is generally applicable in all situations where a score  $s$  with functionality  $s: G \times 2^G \rightarrow \mathbb{R}$  is given, which is the case for all scores within the Bayesian network framework, but also for most other score functions. This is an important property for gene network estimation techniques, since work on new scores incorporating previous knowledge is on-going (Imoto *et al.*, 2003; Tamada *et al.*, 2003; Nariai *et al.*, 2004). In contrast to heuristic approaches, the exact algorithm guarantees finding the most likely networks.

Since the number of significantly distinct gene expression patterns observed in typical data sets is low, the number of essential (groups of) genes will be within the range of feasibility in many cases. For gene networks of size beyond the computational limits of our algorithm, techniques as in Ott and Miyano (2003) can be applied, such that our approach of finding gene network motifs is not limited to small gene networks.

Rigorously assessing the accuracy of gene network estimations using available knowledge as we conducted in the above evaluations, is a promising approach to develop standards for comparing the strength of gene network estimation methods. Since there is little work on the principled evaluation of gene network estimations, this should be further pursued.

## ACKNOWLEDGEMENTS

The authors would like to thank the members of the RegulonDB team for processing our query in natural language, and Yuko Makita for providing yet unpublished data from DBTBS. Furthermore, we are grateful for discussions and advice from Michiel de Hoon and Seiya Imoto. A.H. was supported by a HFSP LT fellowship and a Wellcome Trust Functional Genomics Programme grant (066790/E/02/Z).

<sup>9</sup>The probably most demanding evaluation so far is given in (Rung *et al.*, 2002).

## REFERENCES

- Buntine, W. (1991) Theory refinement on Bayesian networks. *UAI '91*, **7**, 52–60.
- Chen, T., He, H.L. and Church, G.M. (1999) Modeling gene expression with differential equations. *Pac. Symp. Biocomput.*, **4**, 29–40.
- Chickering, D.M. (1996) Learning Bayesian networks is NP complete. In Fisher, D. and Lenz, H.-J. (eds), *Learning from Data: Artificial Intelligence and Statistics V*. Springer-Verlag, NY, pp. 121–130.
- Cooper, G.F. and Herskovits, E. (1992) A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, **9**, 309–347.
- Courcelle, J., Khodursky, A., Peter, B., Brown, P.O. and Hanawalt, P.C. (2001) Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient *Escherichia coli*. *Genetics*, **158**, 41–64.
- Enard, W., Khaitovich, P., Klose, J., Zöllner, S., Heissig, F., Giavalisco, P., Nieselt-Struwe, K., Muchmore, E., Varki, A., Ravid, R. *et al.* (2002) Intra- and interspecific variation in primate gene expression patterns. *Science*, **296**, 340–343.
- Friedman, N. and Goldszmidt, M. (1998) Learning Bayesian networks with local structure. In Jordan, M.I. (ed.), *Learning and Inference in Graphical Models*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 421–459.
- Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000) Using Bayesian networks to analyze expression data. *J. Comp. Biol.*, **7**, 601–620.
- Garey, M.R. and Johnson, D.S. (1979) *Computers and Intractability*. W.H. Freeman and Company, San Francisco, CA.
- Hartemink, A.J., Gifford, D.K., Jaakkola, T.S. and Young, R.A. (2001) Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pac. Symp. Biocomput.*, **6**, 422–433.
- Heckerman, D. (1999) A tutorial on learning with Bayesian networks. In Jordan, M. (ed.), *Learning in Graphical Models*. MIT Press, Cambridge, MA.
- Ideker, T., Ozier, O., Schwikowski, B. and Siegel, A.F. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**, 233–240.
- Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S. and Miyano, S. (2003) Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Comput. Syst. Bioinformatics*, **2**, 104–113.
- Ishii, T., Yoshida, K., Terai, G., Fujita, Y. and Nakai, K. (2001) DBTBS: a database of *Bacillus subtilis* promoters and transcription factors. *Nucleic Acids Res.*, **29**, 278–280.
- Khodursky, A.B., Peter, B.J., Cozzarelli, N.R., Botstein, D., Brown, P.O. and Yanofsky, C. (2000a) DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **97**, 12170–12175.
- Khodursky, A.B., Peter, B.J., Schmid, M.B., DeRisi, J., Botstein, D. and Brown, P.O. (2000b) Analysis of topoisomerase function in bacterial replication fork movement: use of DNA microarrays. *Proc. Natl Acad. Sci. USA*, **97**, 9419–9424.
- Makita, Y., Nakao, M., Ogasawara, N. and Nakai, K. (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its

- contribution to comparative genomics. *Nucleic Acids Res.*, **32**, D75–D77.
- Martinez-Yamout,M., Legge,G.B., Zhang,O., Wright,P.E. and Dyson,H.J. (2000) Solution structure of the cysteine-rich domain of the *Escherichia coli* chaperone protein DnaJ. *J. Mol. Biol.*, **300**, 805–818.
- Milo,R., Shen-Orr,S., Itzkovitz,S., Kashtan,N., Chklovskii,D.D. and Alon,U. (2002) Network motifs: simple building blocks of complex networks. *Science*, **298**, 824–827.
- Nariai,N., Kim,S.-Y., Imoto,S. and Miyano,S. (2004) Using protein–protein interactions for refining gene networks estimated from microarray data by Bayesian networks. *Pac. Symp. Biocomput.*, **9**, 336–347.
- Oleksiak,M.F., Churchill,G.A. and Crawford,D.L. (2002) Variation in gene expression within and among natural populations. *Nat. Genet.*, **32**, 261–266.
- Ong,I.M., Glasner,J.D. and Page,D. (2002) Modelling regulatory pathways in *E. coli* from time series expression profiles. *Bioinformatics*, **18**, 241–248.
- Ott,S. (2004) Finding optimal models for gene networks. PhD thesis, University of Tokyo, Tokyo, Japan.
- Ott,S. and Miyano,S. (2003) Finding optimal gene networks using biological constraints. *Genome Informatics*, **14**, 124–133.
- Ott,S., Imoto,S. and Miyano,S. (2004) Finding optimal models for small gene networks. *Pac. Symp. Biocomput.*, **9**, 557–567.
- Pe’er,D., Regev,A., Elidan,G. and Friedman,N. (2001) Inferring sub-networks from perturbed expression profiles. *Bioinformatics*, **17**, 215–224.
- Robinson,R.W. (1973) Counting labeled acyclic digraphs. In Harary,F. (ed.), *New Directions in the Theory of Graphs*, Academic Press, New York, pp. 239–273.
- Rung,J., Schlitt,T., Brazma,A., Freivalds,K. and Vilo,J. (2002) Building and analysing genome-wide gene disruption networks. *Bioinformatics*, **18**, 202–210.
- Salgado,H., Santos-Zavaleta,A., Gama-Castro,S., Millán-Zárate,D., Díaz-Peredo,E., Sánchez-Solano,F., Pérez-Rueda,E., Bonavides-Martínez,C. and Collado-Vides,J. (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.*, **29**, 72–74.
- Sonenshein,A.L., Hoch,J.A. and Losick,R. (2001) *Bacillus subtilis and its Closest Relatives: From Genes to Cells* ASM Press, Washington, DC.
- Spellman,P., Sherlock,G., Zhang,M., Iyer,V., Anders,K., Eisen,M., Brown,P., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
- Szabo,A., Korszun,R., Hartl,F.U. and Flanagan,J. (1996) A zinc finger-like domain of the molecular chaperone DnaJ is involved in binding to denatured protein substrates. *EMBO J.* **15**, 408–417.
- Tamada,Y., Kim,S.-Y., Bannai,H., Imoto,S., Tashiro,K., Kuhara,S. and Miyano,S. (2003) Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection. *Bioinformatics*, **19**, 227–236.
- Valbuzzi,A. and Yanofsky,C. (2001) Inhibition of the *B. subtilis* regulatory protein TRAP by the TRAP-inhibitory protein, AT. *Science*, **293**, 2057–2059.
- van Someren,E.P., Wessels,L.F.A., Backer,E. and Reinders,M.J.T. (2002) Genetic network modeling. *Pharmacogen.*, **3**, 507–525.