Enumeration of Likely Gene Networks and Network Motif Extraction for Large Gene Networks

Sascha Ott ott@ims.u-tokyo.ac.jp
Satoru Miyano miyano@ims.u-tokyo.ac.jp

Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Keywords: gene networks, network motifs

1 Introduction

The reliable estimation of gene networks from gene expression measurements is a major challenge in the field of Bioinformatics. Recently, an algorithm for the optimal estimation of small gene networks within the Bayesian network framework was found [3]. This algorithm was further extended to allow the enumeration of all optimal networks and also suboptimal networks in the order of their likelihood [2]. In this work, we show how this result can be applied to the enumeration of likely gene networks for a large number of genes.

Enumerating a number of the most likely gene network models instead of just focusing on the single most likely network model allows to evaluate the reliability of the estimations. If we can find a partial network that is common to most of the likely network models, we can expect this part to be the most reliable part. We denote such common parts as gene network motifs.[1]

2 Method

Let us start with defining a class of subsets of the set of acyclic directed graphs. We use $P^N(g)$ to denote the set of parents of a gene g in a network N.

Definition 1:
Let G denote a set of genes. Let $C_g \subseteq G - \{g\}$ for all $g \in G$. We define the *subspace induced by* C_g, $g \in G$, as $\{N \subseteq G \times G | N \text{ acyclic}, \forall g \in G : P^N(g) \subseteq C_g\}$. •

By the definition, a subspace is given by the selection of candidate parents for each gene. As described in [4], the strongly connected components of the graph induced by the candidate parents yield a decomposition of the set of genes G. Let us use $D_1, \ldots, D_n$ to denote this decomposition for a given subspace. We make use of a result from [4] that shows that optimal network models can be found within subspaces, if the number of candidate parents of a gene and $|D_i|$, $i = 1, \ldots, n$, are bound by a constant not larger than about 30. This result was achieved by repeatedly applying the algorithm from [3] to the components D_i. The following algorithm applies the extended algorithm from [2] instead. It takes one parameter $m \in \mathbb{N}$.

[1] Therefore, our definition of a gene network motif is, at first, different from the notion used, for example, in [1], but might turn out to be closely related.

Algorithm 1:

Step 1:	Select a subspace. The subspace induces a decomposition of G: $D_1, \ldots, D_n$.
Step 2:	Apply the enumeration algorithm[2] from [2] to all D_i to find the best m networks $N_{i,k} = (G, E_{i,k})$, $k \leq m$, for D_i.
Step 3:	Set $\beta^{(1)} = (1, \ldots, 1) \in \mathbb{N}^n$.
Step 4:	For all $j = 2, \ldots, m$, do the following two steps:
Step 4a:	Select $i \leq n$ and $k < j$ minimising $score(N_{i,\beta_i^{(k)}}) - score(N_{i,\beta_i^{(k)}+1})$ among all i, k such that $(\beta_1^{(k)}, \ldots, \beta_{i-1}^{(k)}, \beta_i^{(k)} + 1, \beta_{i+1}^{(k)}, \ldots, \beta_n^{(k)})$ does not match $\beta^{(p)}$ for $k \neq p < j$.
Step 4b:	Set $\beta_l^{(j)} = \beta_l^{(k)}$ for all $l \neq i$ and $\beta_i^{(j)} = \beta_i^{(k)} + 1$.
Step 5:	For all $j \leq m$, return $N_j = (G, \bigcup_{i=1}^{n} E_{i,\beta_i^{(j)}})$.

Theorem 1 *(proof omitted)*

Algorithm 1 finds the best m networks within the subspace selected in Step 1.

We note that it may not be necessary to compute the m best networks for each component in order to find the best m networks within the subspace in practical applications. The combinatorial effect when forming a network from n partial networks will in practice allow more than m optimal combinations, even if less than m optimal networks for the single components are known. In order to utilize this effect, Algorithm 1 can be used with a slight modification: the loop in Step 4 must be terminated when one element of a computed vector $\beta^{(j)}$ is the rank of the last solution known for the particular component.

The size of the subspace as defined above is huge as we illustrate with one example. Let us assume dividing G in groups $D_1, \ldots, D_n$ of equal size 20 and select as candidate parents for a gene $g \in D_i$ the set $D_i - g$. The number of networks in the subspace induced by this selection of candidate parents is roughly $(2.34 \cdot 10^{72})^n$. Therefore, if the biologically interesting region of the search space can be roughly identified, the huge size of the subspace as defined in this work should allow to capture this region well.

3 Conclusion

We have shown that the gene network estimation approaches of [4], which are based on selecting a part of the search space and searching the selected search space optimally, can be further extended to enumerate a number of most likely networks in the subspace. This also allows the use of gene network motif extraction techniques [2] in order to identify reliable parts of gene network models.

References

[1] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U., Network motifs: simple building blocks of complex networks, *Science*, 298: 824–827, 2002.

[2] Ott, S., Kim, S.-Y., Hansen, A., and Miyano, S., Gene networks that are better than optimal, submitted.

[3] Ott, S., Imoto, S., and Miyano, S., Finding optimal models for small gene networks, *Pacific Symposium on Biocomputing*, World Scientific, in press.

[4] Ott, S. and Miyano, S., Finding optimal gene networks using biological constraints, *Genome Informatics*, 14:124–133, 2003.

[2]The algorithm has to be slightly modified in order to allow for interactions between genes in different components.