Comparing Scale-free Priors over Graph Structures for Bayesian Inference of Gene Networks

Paul Sheridan
sherida6@is.titech.ac.jp

Takeshi Kamimura
kamimur1@is.titech.ac.jp

Hidetoshi Shimodaira
shimo@is.titech.ac.jp

Department of Mathematical and Computing Sciences
Tokyo Institute of Technology

1. Introduction

In recent years, a large amount of gene expression data has been collected and estimating gene networks has become a central topic in the field of bioinformatics. Several methodologies have been proposed for constructing a gene network based on gene expression data and the Gaussian graphical model approach is one of the most effective methods. When we look at this method from a Bayesian perspective, questions of the nature and consistency of prior probability specification (prior probabilities over graphical structures etc) have yet to be definitively determined, though a lot of ideas have been suggested [1,2].

Recent studies of networks such as the Internet or the WWW have revealed that the probability that a node has k edges or equivalently k adjacent nodes, follows a power law $P(k) \propto k^{-\gamma}$ over a large range of k, with an exponent γ that ranges between 1 and 3 depending on the system. Such networks are called scale-free and this property is suggested to be appropriate for gene networks as well [3].

Many methods have been proposed on how to generate a scale-free network [3], and they can, roughly, be classified into two types of models. In the first, a static model, each node in the network is assigned a weight. Edges are added to the network over time in such a manner that nodes with larger weights have a higher propensity to acquire edges. The second, a dynamic model, generates a scale-free network via the continuous addition of nodes to an existing network. As each node is added, edges connect it preferentially to the previous nodes in the network that are already well connected.

Previous attempts to confgure scale-free prior distributions over graphs focused on a static model [4,5]. However, that model was chosen for its mathematical simplicity alone; while recent evidence has shown that a dynamic model approach is relevant for describing gene networks [6].

In this study, we propose a new scale-free prior based on a biologically relevan dynamic model. This method is applied to S. cerevisiae gene expression data.

2. Gaussian Graphical Models

Graphical models provide representations of the conditional independence structure of a multivariate distribution as well as access to algorithms for the computation of conditional and marginal densities. Multivariate Gaussian graphical models are defined in terms of Markov properties, i.e. conditional independencies associated with the underlying graph. Thus, model selection can be performed by testing these conditional independencies, which are equivalent to specified zeros among certain (partial) correlation coefficients. The graph G consists of a set of nodes V and a set of edges E. Two nodes v_i and v_j are conditionally independent given the remaining variables if, and only if, $e_{ij} \notin E$. The details of Gaussian graphical models are described in [1].

Formal inference is inherently structured by composition; from a Bayesian perspective, we are interested in the posterior distribution

$$P(G \mid Y) \propto P(Y \mid G)P(G)$$

For the first term $P(Y|G)$, we refer to [1] and, in our study, we concentrate on the prior, $P(G)$.

3. A Dynamic Scale-free Network Model

As mentioned above, recent work has established that a dynamic model approach, based on growth and preferential attachment, is relevant for describing the evolution of gene networks. In more detail, for the S. cerevisiae gene network it has been established that *(i)* there is correlation between the age of a gene and its connectivity, and *(ii)* the number of interactions a gene gains during its evolution is proportional to its connectivity (i.e. a linear preferential attachment scheme) [6].

The model employed here is a linear preferential attachment model and is a mild variant of the highly influential model of Barabasi and Albert [7]. The algorithm for generating a scale-free graph G starts by taking a single node, and at every time step a new node is added along with m edges that link the new node to m different nodes already present in the network. With k_i the connectivity of node i, the probability the new node n will be connected to node i with edge e_{in} satisfies

$$P(e_{in}) \propto k_i + Bn^a$$

where n is the number of nodes present in the network, and B is a parameter that is determined by specifying the other model parameter, a. The term in addition to the connectivity k_i is a consequence of allowing the number of edges added with each new node to increase over time. This effect increases with the magnitude of a. The resulting network, at least asymptotically in the number of nodes, has power law exponent $\gamma = 2 + B(1-a)/(1-Ba)$. Unlike the model of Barabasi and Albert, which always yields a network with $\gamma = 3$, the above modification enables us to choose a power law exponent to match that observed in a gene network.

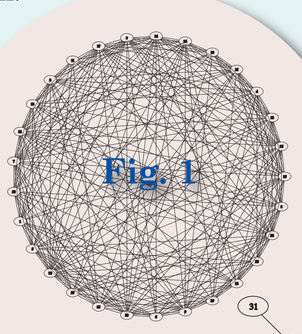

Fig. 1

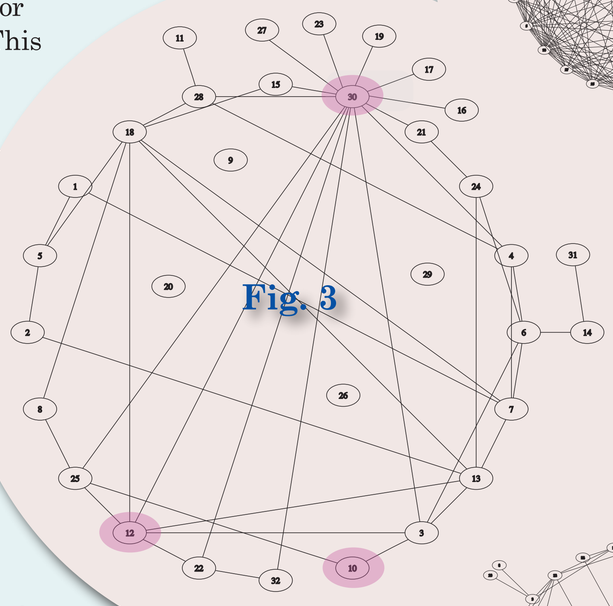

Fig. 3

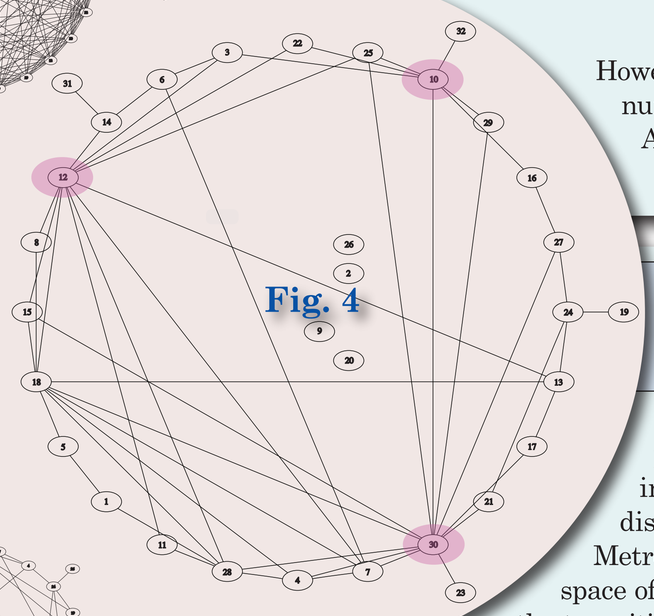

Fig. 4

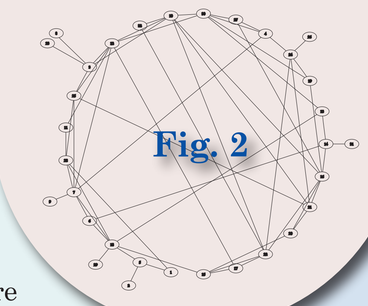

Fig. 2

4. A New Biologically Motivated Scale-free Prior over Graphs

As discussed previously, it has been observed that many gene networks have a heavy-tailed degree sequence k (the number of edges per node) called a power law distribution $P(k) \propto k^{-\gamma}$. And previous attempts to render a prior distribution based on this property have focused solely upon the static model introduced in [5]. We would like to construct a prior using the more biologically relevant dynamic model described in the Part 3. The algorithm to assign a prior probability to any given graph G with nodes $V = \{v_1,\ldots,v_N\}$ is described now.

Let $\sigma = \{\sigma_1,\ldots,\sigma_N\}$ be a permutation of $\{1,\ldots,N\}$. Then each permutation $\sigma_1,\ldots,\sigma_N$, when assigned to the nodes $v_1,\ldots,v_N$, respectively, defines one possible ordering in which $v_1,\ldots,v_N$ were added sequentially to generate a network G under our model. The conditional probability $P(G \mid \sigma)$ can be obtained using the recursive formula

$$P(G_n \mid \sigma) = P(G_{n-1} \mid \sigma) \prod_{e_{in} \in G_n} P(e_{in}) \prod_{e_{in} \notin G_n} \left(1 - P(e_{in})\right)$$

where G_n denotes the network realized at the n'th time step so that $G_N = G$ would be the entire network.

A formula for the prior $P(G)$ is obtained by averaging $P(G \mid \sigma)$ over all possible permutations

$$P(G) = \frac{1}{N!} \sum_{\sigma\,:\,\text{all permutations}} P(G \mid \sigma)$$

However, as the number of nodes increases, the number of permutations grows dramatically. An approximation used in our implementation is described below (see extra notes).

5. MCMC

MCMC is a widely used tool for exploring the space of graphical structures; or in other words, approximating the posterior distribution $P(G \mid Y)$. We implemented the Metropolis-Hastings sampler for searching the space of non-decomposable graphs. Details of how the transition step from G to G' is implemented are given below (see extra notes).

6. A Numerical Example

We applied the new prior to S. cerevisiae gene expression data. We focused on 32 genes which are shown in the table below. The Metropolis-Hasting sampler was run for 250,000 steps (we used Transition type-1 for the first 10,000 steps for fast convergence to the stationary distribution and Transition type-2 thereafter) we took $\gamma = 2.2$. Figures 1,2,3, and 4 are the resulting networks using different priors; they had the highest log posterior probabilities in each chain.

Figure 1: Estimated gene network using the uniform prior over all graphs. This network is very dense and the number of edges a node has is almost uniform, which is inconsistent with biological observation [8,9].
Figure 2: Estimated gene network using a Bernoulli prior on each edge inclusion probability. This approach to prior specification penalizes only the number of edges, so the estimated network is sparser, but the number of edges per node is almost uniform and is inconsistent, again, with biological observation [8,9].
Figure 3: Estimated gene network using a static model based scale-free prior. This time the network is sparse and it has hubs (i.e. node 30), which is consistent with biological observation [8.9].
Figure 4: Estimated gene network using the dynamic model based scale-free prior. It also exhibits a scale-free structure with node 30 a major hub. This network distinguishes itself from the scale-free one in Figure 3 as nodes 10 and 12 are hubs, which agrees with biological evidence [9].
Figure 5: This plot shows that nodes 10 and 12 were the only nodes that differed by more than 2 edges in the estimated networks using the static and dynamic priors.

	gene		gene		gene		gene
1	RAD51	9	SPO16	17	MCM1	25	SIP3
2	CLN1	10	FKH2	18	ACE2	26	SMK1
3	CLB2	11	MBP1	19	GAT3	27	UGA3
4	BUD9	12	SWI6	20	ACA1	28	UME6
5	TSL1	13	NDD1	21	KRE33	29	WAR1
6	JIP1	14	STE12	22	RCO1	30	YER184C
7	EGT2	15	SWI4	23	RFX1	31	YGR067C
8	SWI5	16	FKH1	24	SFL1	32	YRR1

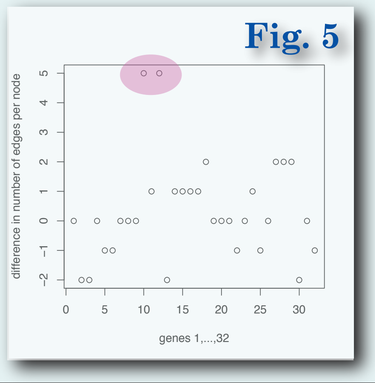

Fig. 5

References

[1] Jones, B. et al., Experimnets in stochastic computation for high-dimensional graphical models. *SAMSI Tech. Report*, 2004-1, 2004.
[2] Imoto, S. et al., Combining microarrays and biological knowledge for estimating gene networks via Bayesian networks. *Journal of bioinformatics and computational biology*, 2:77-98, 2003.
[3] Newman, M., The structure and function of complex networks. *SIAM Rev.*, 45:167-256, 2003.
[4] Lee, D. et al., Scale-free random graphs and Potts model. *Pramana J. Phjy.*, 64:1149-1159, 2005.
[5] Kamimura, T. and Shimodaira, H., A scale-free prior over graph structures for Bayesian inference of gene networks. *PSB Poster Presentation*, 2006.
[6] Eisenberg, E. and Levanon, Y., Preferential attachment in protein network evolution. *Phys. Rev. Letters 91*, 138701, 2003.
[7] Dorogovtsev, S.N. and Mendes, J.F.F., Effect of the accelerating growth of communications networks on their structure. *Phys. Rev. E 63*, 025101, 2001.
[8] van Noort, V. et al., The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Reports*, 5:280-284, 2004.
[9] http://db.yeastgenome.org