Stability of Inferring Gene Regulatory Structure with Dynamic Bayesian Networks

Jagath C. Rajapakse[1,2,3,⋆] and Iti Chaturvedi[1]

[1] Bioinformatics Research Center, School of Computer Engineering, Nanyang Technological University, Singapore 639798
[2] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02142, USA
[3] Singapore-MIT Alliance, Singapore

Abstract. Though a plethora of techniques have been used to build gene regulatory networks (GRN) from time-series gene expression data, stabilities of such techniques have not been studied. This paper investigates the stability of GRN built using dynamic Bayesian networks (DBN) by synthetically generating gene expression time-series. Assuming scale-free topologies, sample datasets are drawn from DBN to evaluate the stability of estimating the structure of GRN. Our experiments indicate although high accuracy can be achieved with equal number of time points to the number of genes in the network, the presence of large numbers of false positives and false negatives deteriorate the stability of building GRN. The stability could be improved by gathering gene expression at more time points. Interestingly, large networks required less number of time points (normalized to the size of the network) than small networks to achieve the same level stability.

Keywords: Dynamic Bayesian networks, gene regulatory networks, Markov chain Monte Carlo simulation, scale-free networks, stability.

1 Introduction

Biological activities are due to interactions between genes and/or gene products (mainly proteins). These interactions are causal in the sense that the expression of one gene causes another gene to be up- or down-regulated. Transcription factors are a group of master proteins that bind to promoter regions and initiate transcription of genes. They positively or negatively mediate ensuing complex activities and enable control of several pathways, simultaneously. A set of genes that coherently interact to achieve a specific biological activity constitute to gene regulatory networks (GRN). Genetic and biochemical networks of cells are constantly subjected to random perturbations and must withstand substantial random perturbations. Therefore, the principles of efficiency and stability of such networks through feedback mechanisms are inherent in biological networks.

[⋆] Corresponding author.

M. Loog et al. (Eds.): PRIB 2011, LNBI 7036, pp. 237–246, 2011.

GRN represents regulations among genes in a directed graph where nodes denote genes and edges regulatory connections. Microarrays are able to gather expression patterns of thousands of genes, simultaneously, and if collected over time or many experiments, underlying GRN can be constructed and inferences on gene regulations can be made. Reconstruction of genetic networks from gene expression data has many applications including inferring underlying biological mechanisms and identification of key biomarkers for drug discovery. A plethora of computational approaches have been introduced in the literature to infer GRN from microarray data, using boolean networks [1], [2], differential equations [3], concept networks, [4], and Bayesian networks [5], [6], [7], [8], [9], [10]. However, investigation of their stability of building GRN has evaded the research community.

Dynamic Bayesian networks (DBN) have recently become increasingly popular in building GRN from gene expression data because of their ability to model causal interactions [11]. Genes are represented at the nodes and regulatory interactions between two genes are represented by conditional probabilities of expressions of the two genes. Sensitivity of GRN in the Bayesian framework has been studied previously by sampling time-series from the posterior distribution [10]. However, the stability of networks of different sizes and its dependence on the number of time points in constructing GRN have not been evaluated. In this paper, we investigate the stability of building GRN with DBN by extensive simulations. Gene expression datasets of varying time points were generated by GRN, using scale-free topologies. Stability measures of estimating the structure are obtained by using bootstrapped time-series datasets. The effect of the number of time points and genes on the stability are also investigated.

2 Gene Regulatory Networks

A Bayesian network (BN) models the likelihood of a set of random variables by taking into account their conditional independences. When gene regulatory networks (GRN) are modeled by BN, genes are represented at the nodes and regulatory interactions are parameterized over the connections by conditional probabilities of gene expressions. Suppose that expressions $x = (x_i)_{i=1}^{I}$ of I number of genes indexed by set $\{i\}_{i=1}^{I}$ are given. If s denotes the structure and θ the parameters of the network, the likelihood of gene expressions is given by

$$p(x|s, \theta) = \prod_{i=1}^{I} p(x_i|a_i, \theta_i) \tag{1}$$

where $p(x_i|a_i, \theta_i)$ is the conditional probability of gene expression x_i of gene i, given the set of gene expressions a_i of its parents; and $\theta = (\theta_i)_{i=1}^{I}$ where θ_i denotes the parameters of the conditional probability. The BN decomposes the likelihood of gene expressions into a product of conditional probabilities conditioned on the expressions of parent genes.

2.1 Dynamic Bayesian Networks

Dynamic Bayesian networks (DBN) extends the BN framework to better capture the temporal characteristics of gene expressions by assuming a first-order stationary Markov chain. Suppose that the time-course of gene expression of gene i is given by $x_i = (x_i(t))_{t=1}^{T}$ where $x_i(t)$ denotes the expression of gene i at time t and T is the total number of time points at which the expressions of genes are gathered. The time points are assumed to be equally spaced. The graph structure of DBN represents regulations from genes in the previous time point to the genes at the present point.

Suppose that gene expressions are discretized into K levels (or states). The parent gene expressions a_i will have $q_i = K^{|a_i|}$ number of states where $|a_i|$ denotes the number of parents of gene i. Let $\theta_{ijk} = p(x_i(t) = k|a_i(t-1) = j)$. By using the property of decomposability, the likelihood (1) can be written as

$$p(x|s,\theta) = \prod_{i=1}^{I} \prod_{j=1}^{J_i} \prod_{k=1}^{K} \theta_{ijk}^{N_{ijk}}. \tag{2}$$

where $N_{ijk} = \sum_{t=2}^{T} \delta(x_i(t) = k)\delta(a_i(t-1) = j)$ denotes the total number of counts of parents' state $a(i)$ being j when the gene i takes the value k in the preceding time point. Note that conditional probabilities are given by the conditional probability distribution (cpd) table: $\theta = \{\theta_{ijk}\}_{I \times J_i \times K}$.

2.2 Maximum Likelihood Estimate of the Structure

Marginalizing (2) over the parameters:

$$p(x|s) \propto \int p(x|s,\theta)p(\theta|s)d\theta \tag{3}$$

where $p(\theta|s)$ denotes the prior densities of the parameters. Assuming that the densities of conditional probabilities are independent:

$$p(\theta|s) = \prod_{i=1}^{I} \prod_{j=1}^{J_i} \prod_{k=1}^{K} p(\theta_{ijk}) \tag{4}$$

and substituting (2) and (4) in (3), we get

$$p(x|s) = \prod_{i=1}^{I} \prod_{j=1}^{J_i} \int \prod_{k=1}^{K} \theta_{ijk}^{N_{ijk}} p(\theta_{ijk})d\theta_{ijk} \tag{5}$$

Assuming Dirichlet priors for conditional densities:

$$p(\theta_{ijk}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_{ijk})}{\prod_{k=1}^{K} \Gamma(\alpha_{ijk})} \prod_{k=1}^{K} \theta_{ijk}^{(\alpha_{ijk}-1)} \tag{6}$$

where hyperparameters α_{ijk} correspond to parameters of Dirichlet prior of θ_{ijk}. Substituting (6) in (5), the likelihood is given by

$$p(x|s) = \prod_{i=1}^{n} \prod_{j=1}^{J_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{K} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{7}$$

where $N_{ij} = \sum_{k=1}^{K} N_{ijk}$ and Γ denotes the gamma function. Here we take $\alpha_{ij} = 1.0/(\sqrt{q_i})$. The Dirichlet priors enable modeling of complex interactions among the genes in regulatory networks [5].

Using Lagrange theory, the maximum likelihood (ML) estimates of the parameters can be derived from (7):

$$\hat{\theta}_{ijk} = \frac{\alpha_{ijk} + N_{ijk} - 1}{\alpha_{ij} + N_{ij} - K} \tag{8}$$

2.3 Structure Learning

The structure of GRN is obtained by the maximum a posteriori (MAP) estimation:

$$s^* = \arg\max_s p(s|x) \tag{9}$$

In order to find the optimal structure, Markov chain of structures is formed with Markov chain Monte Carlo (MCMC) simulation which converges to the optimal structure at the equilibrium. The Metropolis-Hastings method of MCMC is adopted, which associates an acceptance mechanism when new sample structures are drawn. The acceptance of new structure s^{new} is given by

$$\min\left\{1, \frac{p(s^{\text{new}}|x)}{p(s|x)} . \frac{Q(s^{\text{new}}|s)}{Q(s|s^{\text{new}})}\right\} \tag{10}$$

where the Metropolis-Hastings acceptance ratio:

$$\frac{p(s^{\text{new}}|x)}{p(s|x)} . \frac{Q(s^{\text{new}}|s)}{Q(s|s^{\text{new}})} . \tag{11}$$

Sampling new structures by using the above procedure generates a Markov chain converging in distribution to the true posterior distribution. In practice, a new network structure is proposed by applying one of the elementary operations such as deleting, reversing, or adding an edge, and then discarding those structures that violate the acyclic condition. The first term of the acceptance ratio, the ratio of likelihoods, is computed using (7). The second term or Hastings ratio is obtained by

$$\frac{Q(s^{\text{new}}|s)}{Q(s|s^{\text{new}})} = \frac{N^{\text{new}}}{N} \tag{12}$$

where N denotes the size of the neighborhood obtained by elementary operations on structure s and counting of the valid structures.

3 Stability

The stability of building GRN was evaluated by drawing time-series samples of gene expressions from a given network topology. The structure s is defined by a connectivity matrix $c = \{c_{ii'}\}_{n \times n}$ where $c_{ii'} = 0$ or $\neq 0$ depending on the presence or absence of a regulatory connection between two genes, i and i'. Stability was evaluated by sampling gene expression time-series from GRN represented by a DBN with known structure and parameters. The distance based similarity criteria were used to evaluate the stability of estimating structure and parameters of GRN.

3.1 Simulation of Gene Expression Time-Series

The likelihood of gene expressions $x(t)$ of genes at t is given by

$$p(x(t)|s, \theta) = \prod_{i=1}^{I} p(x_i(t)|x(t-1), \theta_i) \tag{13}$$

and the gene expression value of gene i is estimated by

$$\hat{x}_i(t) = \arg\max_k p(x_i(t) = k|\theta_{ijk}, x(t-1) = j) \tag{14}$$

If the gene expressions at $t = 1$ is initialized, $\{x_i(1)\}_{i=1}^{I}$, subsequent time points are generated by sampling from GRN by using (14) and adding Gaussian noise $N(0, \sigma^2)$.

Algorithm 1. Sampling gene expression data

```
Given: structure $s$, parameters $\theta$, and Gaussian noise s.d. $\sigma$
Let $x(1) = (x_i(1) = \text{rand}[1, K])_{i=1}^{I}$
for $t = 2$ to $T$ do
  for $i = 1$ to $I$ do
    if $x(t-1) = j$ and $j \in J_i$ then
      $\hat{x}_i(t) = \arg\max_k p(x_i(t) = k|\theta_{ijk}, x(t-1) = j)$
    else
      $\hat{x}_i(t) = 0$
    end if
    $\epsilon_i \sim N(0, \sigma^2)$
    $x_i(t) = \hat{x}_i(t) + \epsilon_i$
  end for
end for
Return: dataset $(x(t))_{t=1}^{T}$
```

3.2 Stability of Estimation of Structure

We propose a similarity based stability criterion for evaluating GRN building algorithms, which is measured by the mean over all pairwise Hamming distances among all detected connections over different datasets.

Let $\{x^b\}_{b=1}^{B}$ be a set of B samples of gene expression datasets and s^b be the GRN derived from b-th dataset x^b. Consider two GRNs with connectivity matrices c^b and $c^{b'}$ derived from datasets x^b and $x^{b'}$, respectively; the similarity of the two networks is obtained by the average Hamming distances over all regulatory connections:

$$\rho\left(c^b, c^{b'}\right) = 1 - \frac{1}{N^b + N^{b'}} \sum_{i=1}^{I}, \sum_{i'=1}^{I} d\left(c^b_{i,i'}, c^{b'}_{i,i'}\right). \tag{15}$$

where d denotes the Hamming distance between the two structures and N^b denotes the number of connections in the network s^b having the connectivity matrix c^b. The Hamming distance takes into account both the presence (1) and the absence (0) of regulatory connections between two networks. The stability is in the range of $[0, 1]$ where a lower value indicates a higher stability of GRN inference algorithm.

Using the stability of two networks in (15), the stability of GRN structure is obtained by averaging over B number of structures. The average of pairwise stability given in (15) over the bootstrapped samples is then used as the stability of the network structure:

$$\rho_{\text{structure}} = \frac{2}{B(B-1)} \sum_{b=1}^{B} \sum_{b'=b+1}^{B} \rho\left(c^b, c^{b'}\right). \tag{16}$$

The stability of an independent connection is defined as the average of bootstrap samples

$$\rho_{Connection}\left(i, i'\right) = \frac{1}{B} \sum_{b=1}^{B} c^b_{i,i'} \tag{17}$$

4 Experiments and Results

4.1 Network Topology

Multiple time-series datasets of gene expressions were generated for a given network topology. In this study, various scale-free networks with varying number of genes (nodes) were generated using Barabasi-Albert model [12] as most real world networks, such as biological networks, are scale-free. Assuming that large GRNs adhere to the topology of scale-free networks, synthetic structures of GRN were built by initiating a small number of nodes. New nodes and edges were added with preferential attachment because probability of addition of new nodes to the existing node is not uniform. A node with high number of edges attracts higher number of new nodes compared to a node with no connection. This phenomena in fact leads to power-law distribution where probability $P(i)$ of preferential attachment of a new gene to existing gene i is given by

$$P(i) \sim d_i^{\gamma} + b \tag{18}$$

where d_i denotes the number of adjacent edges not initiated by gene i itself (which approximates to the in-degree of gene i). The parameters γ denotes the power of preferential attachment and b the attractiveness of the gene with no adjacent edges. New edges are added until a pre-specified number of edges are achieved. Since, GRN are causal we randomly flipped the direction of edges to create feedback loops while generating the network.

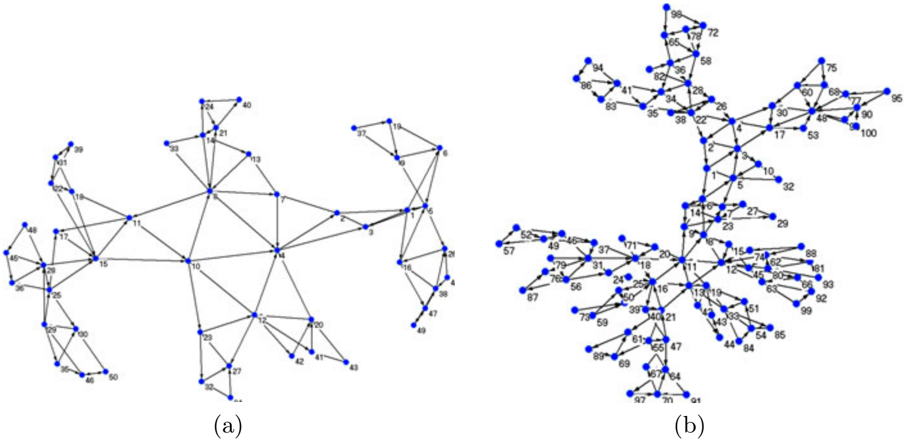


Fig. 1. Synthetic scale free networks with feedback for (a) 50 genes (b) 100 genes

4.2 Performance Evaluation

Scale-free network topologies were generated with in-degree set to a maximum of 5 genes. Figure 1 illustrates scale-free networks of 50 and 100 genes. GRNs defined by scale-free networks of 50, 100 and 250 genes were simulated and the parameters, or the cumulative probability distribution (cpd) tables $\{\theta_{ijk}\}_{i=1,j=1,k=1}^{I,J_i,K}$, were randomly initialized. The genes were assumed to take three states $\{\ -1,\ 0,\ +1\}$ representing down-, no-, and up-regulation of genes. Genes with no parents were set to up-regulation with corresponding CPD $= \{0.1, 0.1, 0.8\}$, for all remaining genes the state was determined by the dominant state in parent set.

Accuracy. In order to study the effects of the number of time points on the size of the network being built, gene expression values were sampled at $T \in \{10, 50, 100, 200, 250\}$ time points from GRN by using Algorithm 1. First, gene expression at the first time point were initialized randomly; second, gene expressions at other time points were sampled from DBN, using (13); and third, Gaussian noise with variance $\sigma^2 = 0.2$ were added. Structural learning was done using MCMC simulations until convergence. For a given I number of genes and T number of time points, $B = 100$ bootstrap time-series datasets were generated. Performance measures such as precision[1], recall[2], accuracy [3], and F-measure[4] were evaluated using true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), respectively.

Table 1 shows the performance of each method with varying numbers of genes and time points. As seen, the number of time points equaling the number of genes produced over 90% of accuracy of networks consisting of more than 50 genes. However, precision and recall measures were low, indicating that the performance was adversely affected by the presence of large numbers FN and FP (more by FP). Increase in the number of time points improved overall performance; note that the improvement was higher for larger networks.

Table 1. Accuracy of building DBN model with Dynamic Bayesian Networks

#Genes	#Samples	Precision	Recall	Accuracy	F-measure
	10	0.05 ± 0.01	0.05 ± 0.01	0.91 ± 0.09	0.05 ± 0.01
	50	**0.21 ± 0.05**	**0.41 ± 0.07**	**0.90 ± 0.09**	**0.28 ± 0.06**
50	100	0.59 ± 0.10	0.68 ± 0.09	0.96 ± 0.09	0.63 ± 0.09
	200	0.95 ± 0.09	0.98 ± 0.09	0.95 ± 0.09	0.98 ± 0.09
	250	0.96 ± 0.09	0.96 ± 0.09	0.98 ± 0.09	0.98 ± 0.09
	10	0.05 ± 0.01	0.05 ± 0.01	0.94 ± 0.11	0.05 ± 0.01
	50	0.23 ± 0.04	0.45 ± 0.07	0.94 ± 0.11	0.31 ± 0.05
100	**100**	**0.37 ± 0.06**	**0.58 ± 0.08**	**0.97 ± 0.11**	**0.46 ± 0.06**
	200	0.83 ± 0.10	0.83 ± 0.10	0.97 ± 0.10	0.83 ± 0.10
	250	0.89 ± 0.10	0.84 ± 0.10	0.98 ± 0.10	0.86 ± 0.10
	10	0.02 ± 0.01	0.02 ± 0.01	0.97 ± 0.09	0.02 ± 0.01
	50	0.18 ± 0.02	0.39 ± 0.05	0.97 ± 0.09	0.25 ± 0.02
250	100	0.31 ± 0.03	0.54 ± 0.05	0.97 ± 0.09	0.40 ± 0.04
	200	0.49 ± 0.06	0.63 ± 0.06	0.98 ± 0.10	0.55 ± 0.06
	250	**0.63 ± 0.06**	**0.71 ± 0.06**	**0.98 ± 0.09**	**0.67 ± 0.06**

Stability. The heatmap of figure 2 shows that the stability of learning GRN structure by using DBN decreases with the number of time points relative to the network size. Interestingly, small networks required more number of time points, relative to the number of nodes in the networks, than the large networks. Figure 3 shows stability of individual edges in networks consisting of different numbers of genes. Increase in the number of time points results in an increase in the number of edges that can be constructed stably. The stability deteriorates when

[1] Precision = TP/(TP+FP).

[2] Recall = TP/(TP + FN).

[3] Accuracy = TP+TN/(TP+TN+FP+FN).

[4] F-measure $=2\ \frac{Precision \times Recall}{Precision + Recall}$.

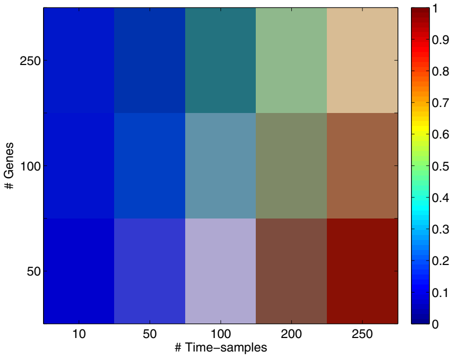
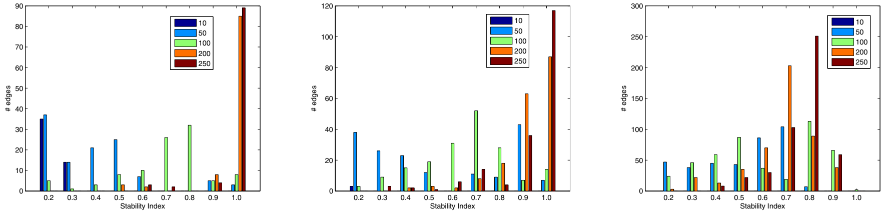


Fig. 2. Heatmaps illustrating the stability of estimating the structure of GRN by DBN. The number of discrete levels of gene expressions is three ($K = 3$).



(a) (b) (c)

Fig. 3. Distribution of stability of edges with the number of time points of gene regulatory networks with (a) 50 genes, (b) 100 genes, and (c) 250 genes

the number of genes relative to the networks size are increased with respective to the number time points.

5 Conclusions

Biological networks are constantly subjected to random perturbations and noise, and efficient feedback and compensatory mechanisms naturally provide stability. Gene expressions gathered from microarray experiments are confounded not only by biological noise and variations but also by measurement errors and artifacts. As seen, although the accuracy is high, GRN built from time-series of gene expressions are often unstable or non-reproducible.

In this paper, we investigated stability of building GRN in DBN framework by using simulations. By bootstrapping gene expression time-series datasets synthetically from scale-free networks (taken as topology closer to biological networks), stability of building GRN of different sizes, with varying number of time points were studied. Although a number of time points equal to the number of genes in the network provide accuracy as high as 90%, the presence of false negatives and false positives were high, especially when the number of time points are relatively less than the size of the network. This could only be reduced by increasing the number of time points at which gene expressions are gathered.

This work with synthetic data provides a foundation for future studies of stability of building GRN with real data. In practice, although gene expressions of a large number of genes are gathered simultaneously, repetition over many time points is limited due to high cost and time involved, and the nature of experiments. In such scenarios, either bootstrapping or regularization of time samples of gene expressions is needed to study the stability and reproducibility of computationally derived GRN. Investigation on real data is beyond the scope of the present manuscript but the application to real data remains an important and challenging problem in functional genomics and systems biology.

References

1. Li, P., Zhang, C., Perkins, E.J., Gong, P., Deng, Y.: Comparison of probabilistic boolean network and dynamic bayesian network approaches for inferring gene regulatorynetworks. BMC Bioinformatics 8, S13–S20 (2007)
2. Akutsu, T., Miyano, S., Kuhara, S.: Algorithms for identifying boolean networks and related biological networks based on matrix multiplication and fingerprint function. Journal of Computational Biology 7(3-4), 331–343 (2000)
3. Liu, B., Thiagarajan, P., Hsu, D.: Probabilistic approximations of signaling pathway dynamics. In: Computational Methods in Systems Biology, pp. 251–265 (2009)
4. Gebert, J., Motameny, S., Faigle, U., Forst, C.V., Schrader, R.: Identifying genes of gene regulatory networks using formal concept analysis. Journal of Computational Biology 15(2), 185–194 (2008)
5. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using bayesian networks to analyze expression data. Journal of Computational Biology 7(3-4), 601–620 (2000)
6. Imoto, S., Kim, S., Goto, T., Miyano, S., Aburatani, S., Tashiro, K., Kuhara, S.: Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network. J. Bioinform. Comput. Biol. 1(2), 231–252 (2003)
7. Nariai, N., Kim, S., Imoto, S., Miyano, S.: Using protein-protein interactions for refining gene networks estimated from microarray data by bayesian networks. In: Pac. Symp. Biocomput., pp. 336–347 (2004)
8. Ota, K., Yamada, T., Yamanishi, Y., Goto, S., Kanehisa, M.: Comprehensive analysis of delay in transcriptional regulation using expression profiles. Genome Informatics 14, 302–303 (2003)
9. Perrin, E.D., Liva, R., Mazurie, A., Bottani, S., Mallet, J., dAlche-Buc, F.: Gene network inference using dynamic bayesian networks. Bioinformatics 12(2), 138–148 (2003)
10. Husmeier, D.: Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. Bioinformatics 19(17), 2271–2282 (2003)
11. Friedman, N., Murphy, K., Russell, S.: Learning the structure of dynamic probabilistic networks. In: Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence (UAI 1998), pp. 139–140 (1998)
12. Albert-Laszlo, B., Reka, A.: Emergence of scaling in random networks. Science 286(5439), 509 (1999)