Identifying Disease Candidate Genes via Large-scale Gene Network Analysis

Haseong Kim

Department of Electrical and Electronic Engineering,
Imperial College London, UK
E-mail: hk308@imperial.ac.uk

Taesung Park

Department of Statistics,
Seoul National University,
South Korea
E-mail: tspark@snu.ac.kr

Erol Gelenbe*

Department of Electrical and Electronic Engineering,
Imperial College London, UK
E-mail: e.gelenbe@imperial.ac.uk
*Corresponding author

Abstract: Since gene regulatory networks provide a systematic view of a complex living system, it is important to develop tools which are not only able to build reliable and large-scale gene regulatory networks but also able to identify disease candidate genes using the estimated networks. In this work, we introduce a reverse engineering technique, Bayesian model averaging based networks (BMAnet), which ensembles all appropriate linear models to tackle the uncertainty of model selection and integrates heterogeneous biological datasets. Then various network evaluation measures are used for the comparison of estimated networks and one of the measures called random walk with restart (Rwr) is utilized to search for disease candidate genes.

In the simulation study, our reverse engineering method shows better performance than Elastic-net and Gaussian graphical models but the topological quantities are varying among the three methods. In the real data analysis, brain tumor gene expression samples consisting of non-tumor, grade III, and grade IV were analyzed to estimate their gene networks with total 4422 genes. Based on these estimated networks, 169 brain tumor related candidate genes were identified and some of them were found to be related with "wound", "apoptosis", and "cell death" processes.

Keywords: Large-scale gene regulatory networks; Data integration; Network comparison; Candidate gene identification.

Reference to this paper should be made as follows: Kim, H., Park, TS., and Gelenbe, E. (xxxx) 'Mining Disease Candidate Genes via Large-scale Gene Network Analysis', *Int. J. Data Mining and Bioinformatics*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Haseong Kim Taesung Park Erol Gelenbe

1 Introduction

Rapid growth of high-throughput biological data has led to the development of tools for gene regulatory networks (GRN) analysis which is capable of capturing not only the evidences of physical molecular interactions but also the perspective of their functional roles in a complex system. For example, GRNs can help to identify transcription factors of a specific disease marker genes Carro et al. (2009) and be used to develop candidate drug development Schadt et al. (2009). Also they were investigated to find evolutional clues in developmental processes which affect certain disease development Erwin & Davidson (2009). In order to build these regulatory networks, various mathematical/statistical methods have been introduced Friedman et al. (2000), Zou & Conzen (2005), Margolin et al. (2006), Martin et al. (2007), Opgen-Rhein & Strimmer (2007*b*), Hirose et al. (2008), Kim et al. (2009), De Smet & Marchal (2010) which are usually based on mRNA expression intensities.

However, still many challenges remain in genome scale network analyses. First of all, it is necessary to develop a large-scale network construction algorithm that tackles dimensionality problem and integrates various biological information sources. The dimensionality problem is caused by high-throughput data consisting of at most hundreds of samples while the number of genes could be up to tens of thousands. So models based on ordinary differential equations or linear regression with the ordinary least square employ regularization approaches which make the solution unique and shrinks predictors towards zero Tibshirani (1996), Gustafsson et al. (2005). Elastic-net Zou & Hastie (2005), Friedman et al. (2010) and Graphical Gaussian model Opgen-Rhein & Strimmer (2007*a*), Schfer & Strimmer (2005) are such regularization approaches available for the large-scale GRN construction. Elastic-net imposes on a combination of Lasso and Ridge penalties, which is useful for high dimensional data even when the variables are highly correlated. GGM is also available for the large dataset by the use of Stein-type shrinkage estimator. On the other hand, due to the lack of information of mRNA expression, the reverse engineering method should be able to integrate heterogeneous biological data such as ChIP-chip/seq and DNA/protein sequence information Zhu et al. (2008), Hecker et al. (2009).

Another issue in our study is the network evaluation. Most of the reverse engineering techniques have been evaluated in terms of their sensitivity and specificity based on simulation data. However, when it comes to 'large-scale GRNs', topological properties could be important as much as the individual component connectivity evaluation. It is mainly due to the meaningful genes could

be linked to each other as a module (motif) in the sparse and large-scale GRNs Burda et al. (2011), Milo et al. (2002), Berg & Lässig (2004)). So number of graph alignment techniques based on graph theory have been introduced Flannick et al. (2006), Berg & Lässig (2006), Kuchaiev et al. (2010), Xulvi-Brunet & Li (2010) along with structural measurements Barabási & Oltvai (2004).

The aim of our study is to find disease candidate genes using large-scale GRNs estimated from various source of biological information including mRNA expression. To this end, we introduce a reverse engineering technique based on Bayesian model averaging (BMA) technique Raftery (1995) which ensembles all appropriate linear models to tackle the uncertainty of model selection and integrate heterogeneous biological datasets using Gibbs prior distribution Imoto et al. (2002), Werhli & Husmeier (2008). Then the estimated networks are evaluated using 10 different quantities and one of the quantities called "random walk with restart" is utilized to identify the disease candidate genes by measuring the distances from a gene to a set of genes that are known to be related with the target disease.

2 Bayesian model average for large-scale GRNs

The basic idea of our GRN construction is to search for every possible linear models with a fixed in-degree k and aggregate appropriate models using BMA. Suppose a standardized gene expression dataset $X = x_1, ..., x_p$ where x_i indicates the ith gene then we can define a regression model (denoted by M_{il}) with k genes

$$x_i = \sum_{j \in S_{lk}} b_{ji} x_j + e_i \tag{1}$$

where S_{lk} represents a set of genes belonging to lth combination among all possible combinations each of which consists of k genes, b_{ji} is a coefficient representing the effect of the jth gene on the ith gene, and e_i is an error term. For example, when $X = \{x_1, x_2, x_3, x_4\}$, $k = 2$ and $i = 1$, all possible combinations of the independent variables $(j = 2, 3, 4)$ are $S_{12} = \{x_2, x_3\}, S_{22} = \{x_2, x_4\}$, and $S_{32} = \{x_3, x_4\}$. So M_{11} is $x_1 = b_{21}x_2 + b_{32}x_3 + e_1$. Let θ_{ji} be the true coefficient of b_{ji}. Then our aim is to find $p(\Theta|X)$ where $\Theta = (\theta_1, ..., \theta_i)$ and $\theta_i = (\theta_{1i}, ..., \theta_{ji})(i, j = \{1, ..., p\})$. By Bayes' rule and the law of total probability,

$$p(\theta_{ji}|X) = \sum_{l=1}^{L} p(\theta_{ji}|X, M_{il}) p(M_{il}|X) \tag{2}$$

$$\text{where } p(M_{il}|X) = \frac{p(X|M_{il})p(M_{il})}{\sum_{h=1}^{L} p(X|M_{ih})p(M_{ih})} \tag{3}$$

and L is the number of all possible models. These equations mean that the full posterior distribution of θ_{ji} is a weighted average of its posterior distributions, $p(\theta_{ji}|X, M_{il})$ and the weight is the posterior model probability, $p(M_{il}|X)$. Raftery obtained the posterior model probability approximation which is called Bayesian information criteria (BIC) Raftery (1995). BIC of M_{il} is $n \log(1 - R^2) + k \log n$ where R^2 is the coefficient of determination of the model M_{il}. In our study,

maximum k value is fixed three for computation efficiency so, in an exhaustive searching, total number of models $L = \sum_{k=1}^{K} \binom{n}{k}$. Now, we can obtain the posterior mean approximation of the true coefficient as follows Raftery (1995)

$$E(\theta_{ji}|X, \theta_{ji} \neq 0) \approx \sum_{l=1}^{L} b_{ji} p(M_{il}|X) I(j \in S_{il}) \tag{4}$$

where $I(C)$ is an indicator function which is 1 when C is true or 0 other wise.

In order to reduce false positives, we can prune the edges whose effect size is likely to be zero by deleting the edges having the smallest value of Equation (4) until its edge rato in Equation (5) reaches a criterion α_{Er}.

$$Er = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(P(\theta_{ji} \neq 0|X) \neq 0)}{n(n-1)/2} \tag{5}$$

Note that, in our real data example, we use $\alpha_{Er} = 0.002$ which is in agreement with $Er \approx 0.002$ observed in DNA-protein/protein-protein networks from BIOGRID Stark et al. (2006).

In order to integrate heterogeneous data and enhance the biological meaning of the estimated GRNs, we used the model prior, $p(M_{il})$, having the form of Gibbs distribution which was successfully applied in other integration studies Imoto et al. (2002), Werhli & Husmeier (2008). Assume a set of genes V_i affects the ith gene, which is known from other sources of biological data. Then we can define an 'energy' function of M_{il} as follows.

$$E(M_{il}) = \frac{1}{k} \sum_{j=1}^{k} (1 - I(x_j \in V_i)) \tag{6}$$

where I is an indicator function and k is the number of in-degrees. The meaning of this energy is a gap between observed and expected degrees that the model M_{il} is true. Equation (6) explains that the energy is zero when the dependent variables of the model M_{il} are all appeared in V_i while it increases as no evidences are found in the source. In this way, other biological information can also be taken to enhance the biological significance of the inferred GRNs. If we have two energy functions, E_1 and E_2, from different sources, then the prior distribution will be

$$P(M_{il}|\beta_1, \beta_2) = G \cdot e^{-(\beta_1 E_1(M_{il}) + \beta_2 E_2(M_{il}))} \tag{7}$$

where G is a normalization constant and β_1 and β_2 are hyperparameters which indicates the strength of the influence of the prior knowledge. In our study, $\beta_1 = \beta_2 = 5$ based on the simulation study but Werhli *et. al.* showed appropriate values could be chosen automatically Werhli & Husmeier (2008).

3 Network Evaluation

A network structure estimated by reverse engineering methods has been commonly evaluated in terms of sensitivity and specificity by comparing it to a reference

network structure. But we used following 10 measures (In Figure 1) including (a) total number of edges and (b) maximum degree of a network graph.

(c) Difference of degree (Dd): The degree of a node means the number of connections to the other nodes. Let $d_g(i)$ be the degree of the ith gene in a network structure g ($i = \{1, 2, ..., n\}, g = \{A, B\}$). Then

$$\text{Dd} = \frac{1}{n}\sum_{i=1}^{n}\{d_A(i) - d_B(i)\}^2 \tag{8}$$

(d) Sensitivity (Sens), (e) specificity (Spec), and (f) positive predictive value (Ppv) are defined as fallows,

$$\text{Sens} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{Spec} = \frac{TN}{FP + TN} \tag{10}$$

$$\text{Ppv} = \frac{TP}{TP + FP} \tag{11}$$

where TP, FP, FN, and TN are the number of true positive, false positive, false negative, and true negative, respectively.

(g) Similarity (Sim): It measures the similarity of two networks, A and B. Let E_g be the set of edges of a network structure g ($g = \{A, B\}$) then Sim can be defined

$$\text{Sim} = \frac{N(E_A \cap E_B)}{N(E_A \cup E_B)} \tag{12}$$

where $N(S)$ indicates the number of elements of a set S. Sim has the value 1 when both networks are the same, and decreases as two networks share less edges Xulvi-Brunet & Li (2010).

(h) Clustering coefficient (Cc): (Global) Clustering coefficient is a ratio between the number of connections existing among three nodes in its neighbering triplet structure and the maximal number of edges that can exist among them. If we denote the clustering coefficient of the ith node in a network g by $C_g(i)$ then Cc can be defined as

$$\text{Cc} = \frac{1}{n}\sum_{i=1}^{n}\{C_A(i) - C_B(i)\}^2 \tag{13}$$

(i) Average shortest path length (Asp): Shortest path length is the number of edges between two nodes in a (undirected) network so Asp is the mean distance of two nodes when these two are reachable.

$$\text{Asp} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1}^{n} l(i, j) \tag{14}$$

where $l(i,j)$ is the shortest length between two nodes, $i,j(=1,2,...,n)$. $l(i,j)=0$ when $i=j$ or the i node cannot reach to the jth node.

(j) Random walk with restart (Rwr): Random walk on an undirected graph can be used to capture the relationship of a group of nodes in the given network structure Köhler et al. (2008). As a variant of the discrete-time random walk process, random walk with restart allows the walk to start again from the initial node s with probability r in every time step. Let $\mathbf{p}^t$ be a length n vector where the ith element is the probability that the walker is at node i at time t. Then the next probability of Rwr is defined as follows.

$$\mathbf{p}^{t+1} = (1+r)\mathbf{A}\mathbf{p}^t + r\mathbf{p}^0 \tag{15}$$

where $\mathbf{A}$ is the column-normalized adjacency matrix of the graph and $\mathbf{p}^0$ is the initial probability vector which includes a set of non-zero elements (a group of genes that we are interested in, denoted by Z). The steady state probability vector of $\mathbf{p}^t$ can be obtained by iterating Equation (15) until the difference between $\mathbf{p}^t$ and $\mathbf{p}^{t+1}$ is less than a criterion (10^{-6} in this study). Therefore, the Rwr of the group of genes is

$$\text{Rwr} = \sum_{j \in Z} \mathbf{p}_j \tag{16}$$

where Z is the set of genes which are known to be related to a target disease.

4 Disease Candidate Gene Detection

Conventionally, disease candidate genes have been found using differentially expressed gene (DEG) approaches such as t-test or ANOVA. But in this research, we search for the candidate genes that closely linked to a group of disease related genes by measuring the distance from the group of genes to the candidate gene. Given the set of genes Z which are disease related genes, the distance between Z and the ith gene (Rwr_i) can be defined as

$$\text{Rwr}_i = \sum_{j \in Z'} \mathbf{p}_j \tag{17}$$

where $Z' = \{Z, i\}$ and p_j is the jth element of vector $\mathbf{p}$. This Rwr has high value if the genes in Z are closely connected to the target gene i. Note that as an alternative of Rwr, Asp in Equation (14) could be used as a similarity measure among a group of genes but it is known that Rwr outperforms Asp which does not consider the topology of the group members Köhler et al. (2008).

5 Simulation study

To evaluate the proposed Bayesian model averaging network (BMAnet), we generated large-scale gene expression data as follows. Firstly large-scale networks

whose in-degree distribution follows scale-free Barabási & Oltvai (2004) was obtained using R package "igraph". Each of these network structures was converted into a covariance matrix Geiger & Heckerman (1994) with appropriate edge weights which were randomly chosen among $\{-2, -1, -0.5, 0.5, 1, 2\}$. Then the expression data were generated from the multivariate normal distribution using R package "mvtnorm". In this way, we prepared 40 datasets consisting of 20 replications for 1000 and 2000 nodes, respectively. Each dataset has 50 samples. We used these simulated datasets to estimate their network structure using three reverse engineering methods, BMAnet, elastic-net (Enet) Zou & Hastie (2005) and Gaussian graphical model with shrinkage estimator (GGM) Opgen-Rhein & Strimmer (2007*a*). In addition to the 8 quantities described in Method section, two more quantities, the number of edges and max degree, were measured from the inferred networks.

On the other hand, in order to evaluate Rwr used to detect disease candidate genes, we inserted a module network consisting of 10 nodes into each simulated network structure. A node in the module network is reachable directly or indirectly to any of the other 9 nodes. These 10 nodes are considered as the disease related genes (Z in Equation (17)). So high Rwr means that a walker tends to be stay in any of the 10 nodes with a high probability, which reflects the 10 nodes are closely linked together.

Figure 1 shows the mean and standard deviation of the 10 quantities including the total number of edges and max degrees in each of 1000 and 2000 node datasets. Note that we fixed the specificity by 0.99 in estimating network structures. In Figure 1, (a) the total number of edges shows similar amongst the three methods but is getting larger as the number of node increases despite the tight control of specificity (0.99). (b) The max degrees of the true network structures reached around 300 to 400 with large variances. Though BMAnet limited the maximum in-degree of a model by three, its estimated network structures have more than 50 degrees, which is almost the same as Enet. (c) In the difference of degrees, Enet shows the most closest to the true networks. However, BMAnet has the best performance in terms of (d) sensitivity, (f) positive predictive value and (g) similarity. (h) The clustering coefficient of BMAnet is also the most similar to that of true networks. (i) Average shortest path values of BMAnet and Enet are comparable to that of the true networks. But variance of BMAnet is smaller than the others. (j) Random walk with restart (The higher, the better) of BMAnet is relatively lower than Enet and GGM but BMAnet shows the most robust results.

6 Human Brain tumor GRNs

In order to apply our approach to real data, human brain tumor dataset GSE4290 was collected from Gene Expression Omnibus (GEO) and its large-scale GRNs were constructed. This dataset has total 107 samples which consist of 23 non-tumor, 76 grade III tumor (26 astrocytomas + 50 oligodendrogliomas), and 81 grade IV tumor (glioblastomas) samples. We normalized the expression data of each grade and selected 4422 genes based on the significant DEGs from ANOVA. Then BMAnet was applied to build the network structures for each of the three groups. Note that the samples of this dataset were not measured in time course,

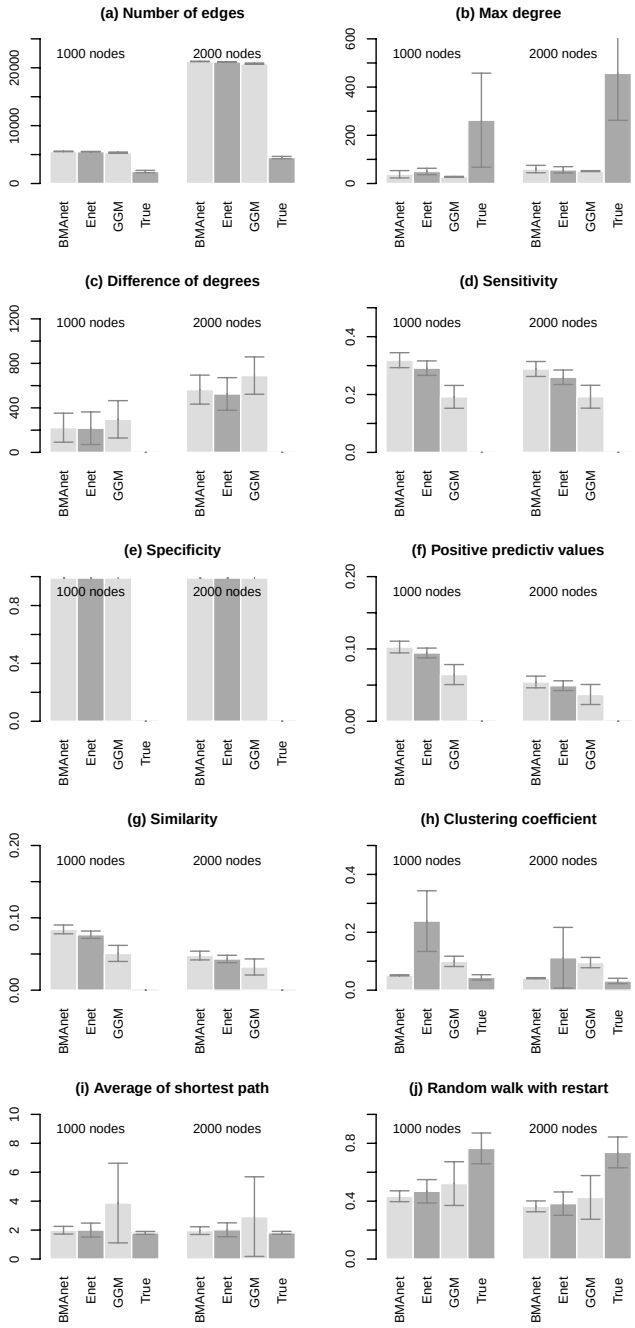

Figure 1 Comparison of 10 measurements of the three reverse engineering methods (BMAnet, Enet, GGM) along with the true networks (True). Please see the context for detailed description.

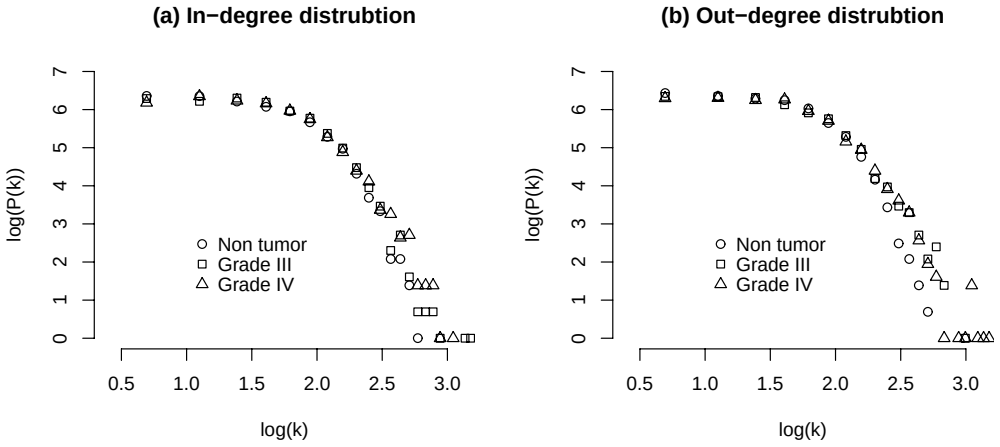

Figure 2 (a) In-degree and (b) out-degree distributions of nontumor, grade III, and IV tumor network structures.

the estimated networks are more likely to be "co-expression networks" rather than "gene regulatory networks". Along with the expression datasets, DNA-protein binding affinity Ernst et al. (2010) is used to update the prior information of BMAnet.

Figure 2 shows the in/out-degree distributions of the estimated networks. Though the distributions are not exactly following the power-law distribution, they show a reasonable range of in/out-degrees covering more than 20 (e^3) despite a model in BMAnet has the maximum three in-degree limitations.

In order to find candidate genes affecting brain tumor, we found 107 tumor related genes from NCBI Gene database. Then the Rwr_i of Equation (17) were measured for all 4422 genes and identified 169 genes that are up-regulated in grade III and IV samples while they are not expressed in non-tumor samples. Figure (3) (a) $\sim$ (c) show the biplots of Rwr values among the three tumor grades; (a) non-tumor vs. grade III, (b) non-tumor vs. grade IV, (c) grade III vs. grade IV. In each plot, we divided four groups of genes based on their expression patterns, for example, U_{N3} (D_{N3}) indicates the genes having higher (lower) Rwr values in grade III while it is low (high) in non-tumor group. Genes in B_{N3} group have high Rwr values in both grade III and non-tumor groups. Therefore, our interested genes in this study are belonging to $\{(U_{N3} \cup U_{N4}) \cap B_{34}\}$

In Figure (3), (d), (e) and (f) show the non-tumor, grade III, and grade IV gene networks, respectively. Light gray dots indicate the known 107 tumor related genes and dark gray ones denote the identified 169 genes via our approach. As we expected, the 169 genes are more closely linked with the known 107 genes in the grade III and IV networks than the non-tumor network. In order to estimate the functional properties of these 169 genes, we obtained the over-represented GO terms of these genes using DAVID Da Wei Huang et al. (2008) which performs Fisher exact test for evaluating the association between two independent groups (observed proportion in a group vs. background proportion in the other group). Table 1 shows GO terms whose p-values of the Exact test are less than an

Table 1 Significantly over-represented GO terms (Biological Process) with p-value < 0.01.

GO term names	p-values
response to wounding	0.00005
negative regulation of gene expression	0.0003
negative regulation of transcription	0.0004
inflammatory response	0.0007
negative regulation of nucleobase-containing compound metabolic proc.	0.0012
negative regulation of nitrogen compound metabolic process	0.0014
positive regulation of macromolecule metabolic process	0.002
negative regulation of macromolecule biosynthetic process	0.0022
regulation of apoptotic process	0.0024
regulation of programmed cell death	0.0026
regulation of cell death	0.0028
negative regulation of cellular biosynthetic process	0.0028
response to hypoxia	0.0032
negative regulation of biosynthetic process	0.0034
regulation of transcription from RNA polymerase III promoter	0.0039
response to oxygen levels	0.041
wound healing	0.045
negative regulation of macromolecule metabolic process	0.054
positive regulation of apoptotic process	0.066
positive regulation of programmed cell death	0.069
positive regulation of cell death	0.072
positive regulation of nitrogen compound metabolic process	0.093
regulation of cell cycle	0.093

empirically chosen criterion (0.1). It shows that many of the 169 genes are related with regulations of "wounding", "cell death", "cell cycle", and "apoptosis".

7 Discussion

In this study, we showed a series of large-scale gene network analyses that cover network construction, network evaluation, and disease candidate gene detection. BMA technique used in our approach has been usually employed to take into account the uncertainty of the model selection problem. So it could be one of the best solutions for explaining complex GRNs where their genes/proteins can alternatively interact with each other depending on their environmental conditions. Moreover, this Bayesian approach enables us to integrate different source of biological information such as the protein-DNA binding affinity. In order to evaluate the estimated networks, 10 quantities of the networks were compared using three different reverse engineering methods where BMAnet showed better or similar performance among the three methods. Finally, the disease candidate genes were found by one of the network evaluation measures, Rwr_i, which is a distance from the ith gene to a group of genes known to be related with the target disease. Utilizing Rwr is not a new approach but we hope we can identify more reliable candidate genes by cooperating with a proper large-scale reverse engineering method. For example, in the brain tumor analysis, ANOVA and FDR adjustment detected too many DEGs (For example, 3023 out of 4422 genes have less FDR values than 0.01. These genes are indicated with black colored edges in Figure 3 (d), (e), and (f)) so it is not possible to pick up proper candidates without the estimated network structures.

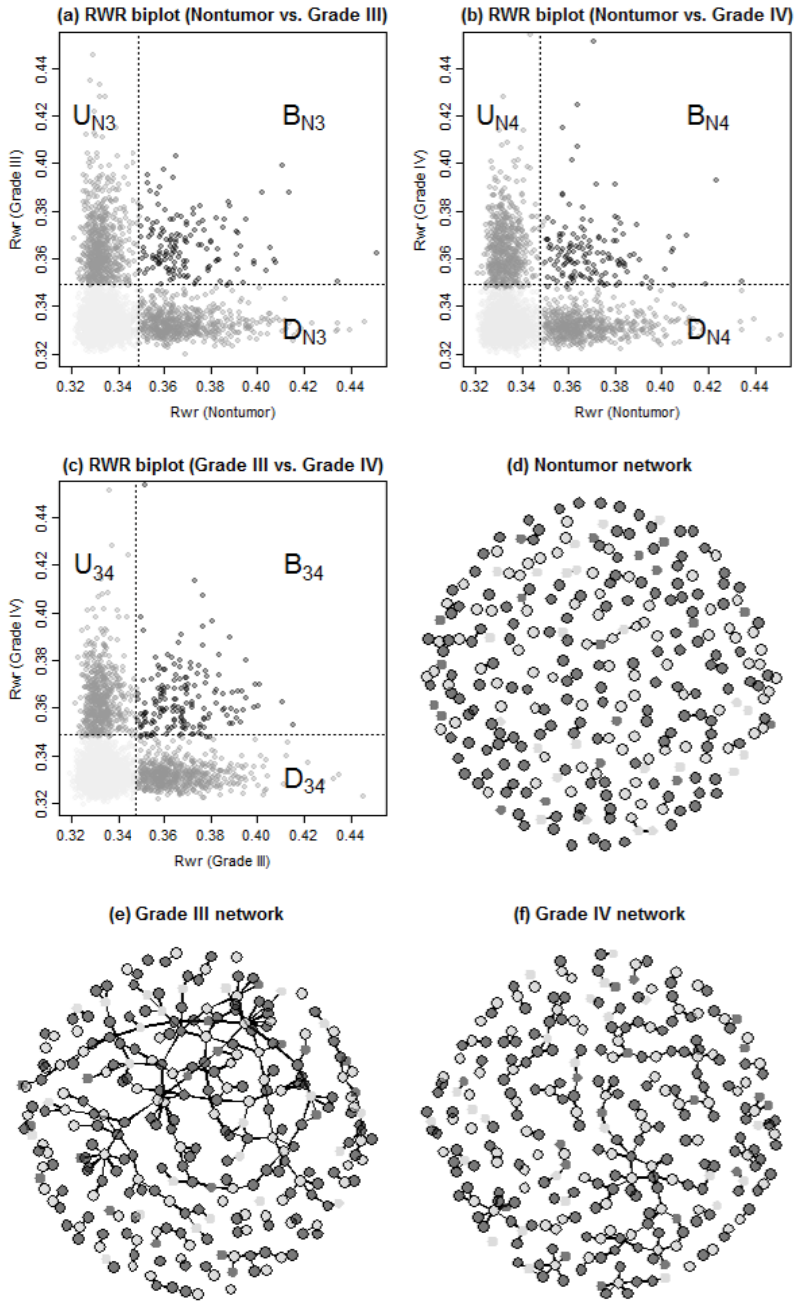

Figure 3 Biplots of Rwr (a) non-tumor vs. grade III, (b) non-tumor vs. grade IV, (c) grade III vs. grade IV. In each plot, we divided genes in three groups, U_{AB}: genes having high Rwr in B group but low in A group, D_{AB}: genes having low Rwr in B group but high in A group, B_{AB}: genes having high Rwr values in both A and B group. (d) co-expression networks of Non-tumor, (e) grade III, and (f) grade IV samples. Light gray dots represent the known tumor related genes and dark gray nodes indicate the identified 169 genes

Via further investigation of the candidate genes, we might be able to identify the disease triggering pathways which will provide more variety of drug targets in industry. But it is also necessary to understand more detailed regulatory relationships among the candidate genes including their proteins/metabolites' dynamic behaviors.

Acknowledgements

References

Barabási, A. & Oltvai, Z. (2004), 'Network biology: understanding the cell's functional organization', *Nature Reviews Genetics* **5**(2), 101–113.

Berg, J. & Lässig, M. (2004), 'Local graph alignment and motif search in biological networks', *Proceedings of the National Academy of Sciences of the United States of America* **101**(41), 14689.

Berg, J. & Lässig, M. (2006), 'Cross-species analysis of biological networks by bayesian alignment', *Proceedings of the National Academy of Sciences* **103**(29), 10967–10972.

Burda, Z., Krzywicki, A., Martin, O. & Zagorski, M. (2011), 'Motifs emerge from function in model gene regulatory networks', *Proceedings of the National Academy of Sciences* **108**(42), 17263–17268.

Carro, M., Lim, W., Alvarez, M., Bollo, R., Zhao, X., Snyder, E., Sulman, E., Anne, S., Doetsch, F., Colman, H. et al. (2009), 'The transcriptional network for mesenchymal transformation of brain tumours', *Nature* **463**(7279), 318–325.

Da Wei Huang, B., Lempicki, R. et al. (2008), 'Systematic and integrative analysis of large gene lists using david bioinformatics resources', *Nature protocols* **4**(1), 44–57.

De Smet, R. & Marchal, K. (2010), 'Advantages and limitations of current network inference methods', *Nature Reviews Microbiology* **8**(10), 717–729.

Ernst, J., Plasterer, H., Simon, I. & Bar-Joseph, Z. (2010), 'Integrating multiple evidence sources to predict transcription factor binding in the human genome', *Genome research* **20**(4), 526.

Erwin, D. & Davidson, E. (2009), 'The evolution of hierarchical gene regulatory networks', *Nature Reviews Genetics* **10**(2), 141–148.

Flannick, J., Novak, A., Srinivasan, B., McAdams, H. & Batzoglou, S. (2006), 'Graemlin: general and robust alignment of multiple large interaction networks', *Genome research* **16**(9), 1169–1181.

Friedman, J., Hastie, T. & Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software* **33**(1), 1.

Friedman, N., Linial, M., Nachman, I. & Pe'er, D. (2000), 'Using Bayesian networks to analyze expression data', *Journal of computational biology* **7**(3-4), 601–620.

Geiger, D. & Heckerman, D. (1994), Learning gaussian networks, Technical report, Technical report, Microsoft Re.

Gustafsson, M., Hornquist, M. & Lombardi, A. (2005), 'Constructing and analyzing a large-scale gene-to-gene regulatory network-lasso-constrained inference and biological validation', *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* **2**(3), 254–261.

Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E. & Guthke, R. (2009), 'Gene regulatory network inference: Data integration in dynamic models–A review', *Biosystems* **96**(1), 86–103.

Hirose, O., Yoshida, R., Imoto, S., Yamaguchi, R., Higuchi, T., Charnock-Jones, D. et al. (2008), 'Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models', *Bioinformatics* **24**(7), 932.

Imoto, S., Goto, T. & Miyano, S. (2002), Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression, *in* 'Pacific Symposium on Biocomputing', Vol. 7, pp. 175–186.

Kim, H., Lee, J. & Park, T. (2009), 'Inference of Large Scale Gene Regulatory Networks Using Regressionbased Approach', *Journal of Bioinformatics and Computational Biology* **7**(4), 717–35.

Köhler, S., Bauer, S., Horn, D. & Robinson, P. (2008), 'Walking the interactome for prioritization of candidate disease genes', *The American Journal of Human Genetics* **82**(4), 949–958.

Kuchaiev, O., Milenković, T., Memišević, V., Hayes, W. & Pržulj, N. (2010), 'Topological network alignment uncovers biological function and phylogeny', *Journal of the Royal Society Interface* **7**(50), 1341–1354.

Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. & Califano, A. (2006), 'Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context', *BMC bioinformatics* **7**(Suppl 1), S7.

Martin, S., Zhang, Z., Martino, A. & Faulon, J. (2007), 'Boolean dynamics of genetic regulatory networks inferred from microarray time series data', *Bioinformatics* **23**(7), 866.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002), 'Network motifs: simple building blocks of complex networks', *Science's STKE* **298**(5594), 824.

Opgen-Rhein, R. & Strimmer, K. (2007*a*), 'From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data', *BMC Systems Biology* **1**(1), 37.

Opgen-Rhein, R. & Strimmer, K. (2007*b*), 'Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process', *BMC bioinformatics* **8**(Suppl 2), S3.

Raftery, A. (1995), 'Bayesian model selection in social research', *Sociological methodology* **25**, 111–163.

Schadt, E., Friend, S. & Shaywitz, D. (2009), 'A network view of disease and compound screening', *Nature Reviews Drug Discovery* **8**(4), 286–295.

Schfer, J. & Strimmer, K. (2005), 'A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics', *Statistical applications in genetics and molecular biology* **4**(1), 32.

Stark, C., Breitkreutz, B., Reguly, T., Boucher, L., Breitkreutz, A. & Tyers, M. (2006), 'Biogrid: a general repository for interaction datasets', *Nucleic acids research* **34**(suppl 1), D535.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

Werhli, A. & Husmeier, D. (2008), 'Gene regulatory network reconstruction by Bayesian integration of prior knowledge and/or different experimental conditions', *Journal of bioinformatics and computational biology* **6**(3), 543–572.

Xulvi-Brunet, R. & Li, H. (2010), 'Co-expression networks: graph properties and topological comparisons', *Bioinformatics* **26**(2), 205–214.

Zhu, J., Zhang, B., Smith, E., Drees, B., Brem, R., Kruglyak, L., Bumgarner, R. & Schadt, E. (2008), 'Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks', *Nature genetics* **40**(7), 854–861.

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.

Zou, M. & Conzen, S. (2005), 'A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data', *Bioinformatics* **21**(1), 71.