Elastic Net Enabled Sparse-Aware Maximum Likelihood for Structural Equation Models in Inferring Gene Regulatory Networks

Anhui Huang

June 13, 2023

Contents

Summary

Understanding the multiple levels of gene regulations is the key for the prediction of complex cellular behavior. Integrating genetic perturbations with gene expression data has been proven to be more accurate in learning of causal regulatory relationships among genes comparing to treating each gene expression level as an individual quantitative trait. The previously designed sparse-aware maximum likelihood method for structural equation models (SEM-SML) has been shown to be able to integrate such information to infer gene regulatory networks (GRN) systematically and offer significant better performance than state-of-the-art algorithms. We extended the SEM-SML to incorporate adaptive elastic net (EN) penalty for the likelihood function of the SEMs, and implemented the SEM-EN software in efficient C/C++ with parallel computational capability by Message Passing Interface (MPI). The parallel design is capable of scaling up the network structure inference in a computer cluster, and enables SEM-EN to infer a network structure with thousands of nodes. Simulation studies demonstrated that SEM-EN was capable of inferring a large network within affordable computational time while achieved more accurate power of detection than SEM-SML. The software was further applied to infer the GRN in budding yeast systematically, in which two set of experimental perturbations with co-regulated gene set information were available on the AMN1 and LEU2

genes. The SEM-EN identified GRN had two clusters with hubs and members in line with the experimental perturbations, corroborating the strength of SEM-EN. While the parallel version of the SEM-EN software for computer cluster is implemented with command line interface, the SEM-EN method is also implemented in C/C++ with a user-friendly R interface for personal computers. An R software package sparseSEM with both SEM-SML and SEM-EN features is available on the Comprehensive R Archive Network (CRAN), and the command-line software is freely available upon request.

Key Words:

- SEM: structural equation models
- EN: elastic net
- GRN: gen regulatory network
- sparseML: sparse-aware maximum likelihood

Back to Top

Introduction

Understanding biological network at system-level is crucial to gaining insights into gene functions and cellular dynamics. To elucidate the complexity of gene regulatory networks (GRNs) and uncover the mechanism of gene regulations that lead to complex biologically diversified phenotypes, a large number of studies have been conducted at genetic, transcriptomic, proteomic and metabolomic level [1]. Experimental approaches in deducing physical interactions of individual genes are time consuming and labor intensive, whereas computational methods that exploit genome-wide expression data and genetic perturbations from high-throughput technologies are efficient and cost-effective [2].

A number of computational methods have been developed to leverage different intermediate phenotypes, such as transcript, protein or metabolite level, to understand cell regulation process comprehensively. For example, a co-expression or relevance network infers network structures through measuring the similarity level in gene expression [3, 4], a Bayesian network evaluates the dependence structure among genes [5], and a Gaussian graphic model approach evaluates the presence of an edge if the pair of genes are conditionally dependent given expression levels of all other genes [6, 7]. Another approach employs regularized linear regression models to find the co-occurrence among genes to infer the gene network [8-11]. Recently, a powerful structural equation models (SEMs) for the GRN modeling was developed [12], which systematically integrated both genetic perturbation and gene expression data, and inferred the GRN through a sparse-aware maximum likelihood (SML) method. The SEM-SML applied an adaptive l1-norm regularization term on the likelihood function, which was then optimized via an efficient block coordinate ascent algorithm. Simulations of the SML algorithm demonstrated that it accurately inferred regulatory relations among genes and offered significant better performance than state-of-the-art algorithms [12].

The SEM-SML algorithm was motivated by the fact that the gene networks are sparse [13-15]. While this is the first study that infers sparse SEM systematically, other penalty functions, especially the penalty function of the elastic net (EN) [16, 17] may improve the inference accuracy for GRNs. This is based on the following observations. Although the Lasso-based methods achieve good performance in the inference of GRNs and are ranked top on the list of a number of methods for GRN inference [18], they tend to miss interactions in feed-forward loops, fan-in motifs and fan-out motifs. This is likely due to the fact that Lasso typically chooses only one variable among several highly correlated variables. On the other hand, it has

been known through experimentation that a gene regulator in GRN can typically shape the expression profile of a set of genes, meaning that the expression of the set of co-regulated genes can be highly correlated [1, 2]. For example, in gene expression microarray analysis, researchers aim at finding a group of up and down-regulated gene expression patterns under different experimental treatments, and discover novel and unexpected functional relationships among genes [19]. Such observations make the elastic net the right fit since it retains correlated variables while still yielding a sparse model [16].

In this paper, we developed an SEM-based method for the inference of GRNs that maximizes the l1/l2-regularized likelihood function similar to the one used in the adaptive EN [16, 17]. The SEM with adaptive elastic net penalty algorithm (SEM-EN) was maximized through a parallelized efficient block coordinate ascent algorithm, which inferred the network structure on each node in parallel. Considering that the MATLAB implementation of the SEM-SML algorithm in [12] was time consuming and not applicable to large network with thousands of nodes, here we further developed a software tool in C/C++ with message passing interface (MPI) to accelerate the computation through parallelization. Thanks to the elastic net penalty, the SEM-EN algorithm encourages a grouping effect that not only predicts causal regulatory genes, but also elucidates the complexity of cell regulation at system level. Computer simulation demonstrated that SEM-EN outperformed SEM-SML with higher power of detection (PD) and similar false discovery rate (FDR). The SEM-EN algorithm was further applied to infer the GRN of a previously described budding yeast (Saccharomyces cerevisiae) dataset.

Back to Top

Methods

Sparse SEM model for gene regulatory networks

Consider the expression levels of N_g genes from N individuals measured in microarray or RNAseq experiments. Following the design of [12], the gene regulatory network is postulated to obey the SEM:

$$\mathbf{y}_i = \mathbf{B}\mathbf{y}_i + \mathbf{F}\mathbf{x}_i + \boldsymbol{\mu} + \varepsilon_i, \quad i = 1, \cdots, N \tag{1}$$

where $\mathbf{y}_i := \left[y_{i1}, \cdots, y_{iN_g}\right]^T$ is the expression levels of N_g genes from N individuals, and $\mathbf{x}_i := \left[x_{i1}, \cdots, x_{iN_q}\right]^T$ denotes the genotype of $N_q \geqslant N_g$ eQTLs of individual $i, i = 1, \ldots, N$. In this paper, we focus on the genetic variations observed at expression quantitative trait loci (eQTLs), and the developed methods can be applied to other genetic variations such as single nucleotide polymorphisms (SNPs), copy number variations (CNVs) and gene knockdown by RNA interference (RNAi) or controlled gene overexpression. B is an $N_g \times N_g$ matrix contains unknown parameters defining the network structure, and is assumed to be sparse; $\mathbf{F}$ is an $N_g \times N_q$ matrix captures the effect of N_q eQTLs for N_g genes; $\boldsymbol{\mu}$ is an $N_g \times 1$ vector accounts for possible model bias; and ε_i is an $N_g \times 1$ vector captures the residual error. Typically, ε_i is modeled as a zero-mean Gaussian vector with covariance σ^2, where I denotes the $N_g \times N_g$ identity matrix.

We assume that locations of N_q eQTLs have been determined using existing eQTL mapping methods, thus $\mathbf{F}$ has N_q entries with known locations but unknown effect size, and $N_g N_q - N_q$ of zero entries. Note that there are two structural properties in the GRN. First, as noted in [12], GRN and other general biochemical networks are sparse, meaning that only a relative smaller number of genes can be regulators of a given gene, thus matrix $\mathbf{B}$ is sparse. Second, a regulator typically can shape the expression profile of a set of genes, meaning that the expression of the set of co-regulated genes can be highly correlated [1, 2]. Based on the SEM-SML algorithm designed in [12], we developed a network inference algorithm that exploits the aforementioned structural properties.

Structural equation models with adaptive elastic net penalty (SEM-EN)

Let us define $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_N], \mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$ and $\mathbf{E} = [\varepsilon_1, \cdots, \varepsilon_N]$, then we can write the SEM in (1) as $\mathbf{Y} = \mathbf{BY} + \mathbf{FX} + \boldsymbol{\mu}\mathbf{1}^T + \mathbf{E}$. The SEM-EN algorithm applied the I_1/I_2-norm penalty to the log-likelihood function of SEM in eq (2) of [12]. Let $\tilde{\boldsymbol{y}}_i = \boldsymbol{y}_i - \overline{\boldsymbol{y}}_i$ and $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \overline{\mathbf{x}}, i = 1, \cdots, N$, where $\overline{\boldsymbol{y}} = \sum_{i=1}^N \boldsymbol{y}_i/N$ and $\overline{\mathbf{x}} = \sum_{i=1}^N \mathbf{x}_i/N$, and collect them into matrices $\tilde{\mathbf{Y}} = [\tilde{\mathbf{y}}_1, \cdots, \tilde{\mathbf{y}}_N], \tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_N]$. Then SEM-EN infers the model parameters through I_1/I_2 penalized ML estimation:

$$(\hat{\mathbf{B}}, \hat{\mathbf{F}})_{EN} = \arg\max_{\mathbf{B},\mathbf{F}} N\sigma^2 \log|\det(\mathbf{I} - \mathbf{B})| - \frac{1}{2}\|\tilde{\mathbf{Y}} - \mathbf{B}\tilde{\mathbf{Y}} - \mathbf{F}\tilde{\mathbf{X}}\|_F^2 - \lambda\alpha\|\mathbf{B}\|_{1,W} - \frac{1}{2}(1-\alpha)\lambda\|\mathbf{B}\|_2 \tag{2}$$

$$\text{subject to } B_{ii} = 0, \forall i = 1, \cdots, N_g, F_{jk} = 0, \forall (j,k) \in S_q$$

where $\|\mathbf{B}\|_{1,W} := \sum_{i=1}^{Ng}\sum_{j=1}^{Ng} w_{ij}|B_{ij}|$ with B_{ij} denotes the (i,j) th entry of $\mathbf{B}$, $\|\cdot\|_F$ denotes the Frobenius norm, and S_q denotes the set of row and column indices of the entries of $\mathbf{F}$ know to be zero. Parameters $\lambda > 0$ and $\alpha \in [0,1]$ are the penalty parameters following the design of the elastic net for linear regression [16]. Weights w_{ij} in the I_1 penalty term are introduced in line with the adaptive the elastic net (EN) [17] and are selected as the estimated coefficients of the ridge regression same as the one described in [12]. Then the SML algorithm is modified to obtain the estimated parameters. Specifically, with the I_1/I_2 penalty $-\lambda\alpha w_{ij}|B_{ij}| - (1-\alpha)\lambda B_{ij}^2/2$ applied to eq(10) in [12], we have the following objective function:

$$g_{ij}(B_{ij}) := N\hat{\sigma}^2 \log|\alpha_0 - c_{ij}B_{ij}| + \alpha_1 B_{ij} - \left[(1-\alpha)\lambda + \frac{1}{2}\alpha_2\right] B_{ij}^2 - \lambda w_{ij}|B_{ij}| \tag{3}$$

where $\hat{\sigma}^2$ is the variance estimate, c_{ij} denotes the (i,j) th co-factor of matrix $(\mathbf{I} - \hat{\mathbf{B}})$ with $\hat{\mathbf{B}}$ being the estimate of matrix $\mathbf{B}, \alpha_0 := \det(\mathbf{I} - \hat{\mathbf{B}}) + c_{ij}\hat{B}_{ij}$ with $\hat{B}_{ij}$ being the estimate of B_{ij}, $\alpha_1 := \left[\left(\mathbf{I} - \hat{\mathbf{B}} + \mathbf{e}_i\mathbf{e}_j^T\hat{B}_{ij}\right)\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T - \hat{\mathbf{F}}^{new}\tilde{\mathbf{X}}\tilde{\mathbf{Y}}^T\right]_{ij}$ with $\mathbf{e}_i$ and $\mathbf{e}_j$ being the ith and j th canonical vectors in $\mathfrak{R}^{Ng}, \hat{\mathbf{F}}^{\text{new}}$ being the estimate of $\mathbf{F}$, and $\alpha_2 := \left\|\tilde{\mathbf{Y}}^T\mathbf{e}_j\right\|_2^2$. Let us define $\tilde{\alpha}_2 = 2(1-\alpha)\lambda + \alpha_2$, then the solution to the objective function in (3) is identical to that of [12] with α_2 replaced by $\tilde{\alpha}_2$, and eqs(1216) in [12] are applied to inter the network parameters. Furthermore, let $\tilde{Q}_{ij}(\lambda)$ denote the derivative of the differential part of (2) :

$$\tilde{Q}_{ij}(\lambda) = \frac{N\sigma^2 c_{ij}(\lambda)}{\det(\mathbf{I} - \hat{\mathbf{B}}(\lambda))} + \left[\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T - \hat{\mathbf{B}}(\lambda)\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^T - \hat{\mathbf{F}}(\lambda)\tilde{\mathbf{X}}\tilde{\mathbf{Y}}^T\right]_{ij} - (1-\alpha)\lambda\hat{\mathbf{B}}(\lambda) \tag{4}$$

where $\hat{\mathbf{B}}(\lambda)$ and $\hat{\mathbf{F}}(\lambda)$ denote the optimal estimate of (2) for a given λ with fixed α, and σ^2 can be estimated as $\hat{\sigma}^2 = \frac{1}{NN_g}\|\tilde{\mathbf{Y}} - \hat{\mathbf{B}}(\lambda)\tilde{\mathbf{Y}} - \hat{\mathbf{F}}(\lambda)\tilde{\mathbf{X}}\|_F^2$. Then the strong rule [39] for SEM-EN is available as following. Let $\lambda_{\max}$ denote the smallest λ that yields $\hat{B}_{ij} = 0, \forall i,j$, and $\lambda_{\max} > \lambda_1 > \cdots > \lambda_{\min}$ is a decreasing set of values, the following discarding rule can be applied to find solution of (2):

$$\begin{aligned} &\left|\tilde{Q}_{ij}(\lambda_{\max})\right| < w_{ij}(2\lambda_l - \lambda_{\max}) \Rightarrow \hat{B}_{ij}(\lambda_l) = 0 \\ &\left|\tilde{Q}_{ij}(\lambda_{l-1})\right| < w_{ij}(2\lambda_l - \lambda_{l-1}) \quad \Rightarrow \quad \hat{B}_{ij}(\lambda_l) = 0 \end{aligned} \tag{5}$$

where $\lambda_1 < \lambda_{\max}$ is a value in the path of λ. Note that for any $\lambda > \lambda_{\max}, \mathbf{B} = 0$, and $\hat{\mathbf{F}}(\lambda)$ is fixed, thus $\tilde{Q}_{ij}(\lambda)$ is also fixed. Therefore, $\lambda_{\max}$ can be obtained from the following equation:

$$\lambda_{\max} = \max_{i,j=1,\cdots,Ng}\left|\frac{Q_{ij}(\lambda_{\max})}{w_{ij}}\right| \tag{6}$$

Software implementation

The block coordinate ascent algorithm in Algorithm 1 of [12] is updated with equations (3) and (4) along with the discarding rules in equations (5) and (6), and is parallelized to reduce execution time.

Specifically, a master computer node is designated to compute and check to convergence criterion, which is determined as err $= \left\|\hat{\mathbf{B}} - \hat{\mathbf{B}}^{\text{new}}\right\|_F^2 / \|\hat{\mathbf{B}}\|_F^2 + \left\|\hat{\mathbf{F}} - \hat{\mathbf{F}}^{new}\right\|_F^2 / \|\hat{\mathbf{F}}\|_F^2$ being smaller than a prespecified small value. And several slavery nodes will be assigned to compute the solution for each row of $\hat{\mathbf{B}}$. The parallelized computation is achieved in a high performance computing (HPC) clusters, and the software is implemented in C/C++ utilizing open MPI. On the other hand, when the scale and degree of a network to be inferred is small and computation is less demanding, we also provide the serial version of the $C/C++$ program with a user friendly R interface. To achieve fast computation, BLAS and LAPACK [37] were utilized in implementation of both software packages.

Simulation study and real data analysis

The network simulation method described in [12] was followed to simulate networks for this study, except that networks were simulated with larger scale and degree. While N_g and number of expected edges (N_e) in simulations of [12] were small ($N_g = 10$ or $30, N_e = 1$ or 3) due to the large amount of computation, we simulated a large network with $N_g = 300$ and $N_e = 3$ to examine the scalability of SEM-EN algorithm. Specifically, a random DAG of $N_g = 300$ and an expected $N_e = 3$ edges per node was first generated by creating directed edges between two randomly picked nodes. Then matrix $\mathbf{F}$ was set as the $N_g \times N_g$ identity matrix, and B_{ij} was generated randomly from uniform distribution in $(0.5, 1)$ or $(-1, -0.5)$ if there was an edge from node j to node i. We simulated two sets of DAGs with different noise levels. The first one had E_{ij} sampled from a Gaussian distribution with zero mean and variance 0.01 , while the second one having variance 0.05 . For each network, 100 replicates were generated and analyzed by SEM-EN for each sample size ranging from 100 to 1000 by step of 100 . We considered 20 values of α ranging from 1 to 0.05 by step size of 0.05 and $20\lambda s$ from $\lambda_{\max}$ to $0.0001\lambda_{\max}$ with even step size at log scale. Ten-folds CV were used to determine the optimal parameters. The PD and FDR of SEM-EN were compared with that of SEM-SML.

We also applied the SEM-EN algorithms to infer a GRN in budding yeast

(Saccharomyces cerevisiae) [40]. The data contains expression levels of 6,126 yeast ORFs and 2,956 genetic markers from 112 yeast segregants from the cross of a laboratory strain (BY4716) and a wild strain (RM11-1a) [30]. We only kept ORFs accepted in the Yeast Comparative Genomics database [41] and those with less than 5% of missing expression data, resulted in 3,380 ORFs. To obtain the set of cis-eQTLs out of the 2,957 genetic markers, we first associated gene markers with an ORF if they were in distance ≤ 20 kb representing the QTL resolution for this cross [40]. The set of cis-eQTLs was then obtained by testing gene expression levels with their associated gene markers through Wilcoxon rank-sum test following the procedure described in [40]. Totally, 1,162 ORFs were found having cis-eQTLs. For simplicity, we only kept one of the most significant cis-eQTLs with the smallest p-value if multiple cis-eQTLs were found for an ORF. The gene expression profile of 3,380 ORFs and genetic marker of 1,162 eQTLs were then used for network inference by SEM-EN.

Back to Top

Results

Scalability of the parallel network structure inference

To achieve feasible computation of inferring a GRN with hundreds or thousands of nodes, the parallel block coordinate ascent algorithm was implemented with a master-slavery paradigm. While one master node as designated for the computation of program initialization, decomposing the problem into small tasks, assign tasks for multiple slave nodes, gathering the results for determining convergence and generate the final result

[20], number of slavery nodes can be supplied by users upon available system resources. To gain insights into the scaling property of the parallel computation, we also implemented the SEM-SML algorithms [12] with C/C++ and paralleled the block coordinate ascent algorithm with Open MPI. In fact, the SEM-SML algorithm is a special case of SEM-EN with shrinkage parameter $\alpha = 1$ (see Methods section for details). The scaling property of the parallel computation was shown in Figure 1, where the performance was obtained from inference of a sparse DAG having $N = 300$ samples, $N_g = 300$ genes, $N_e = 3$ edges, and $\sigma^2 = 0.05$. The computational time was the mean of 5 replicates. From Figure 1, a strong scaling pattern [21] was observed for both SEM-SML and SEM-EN.

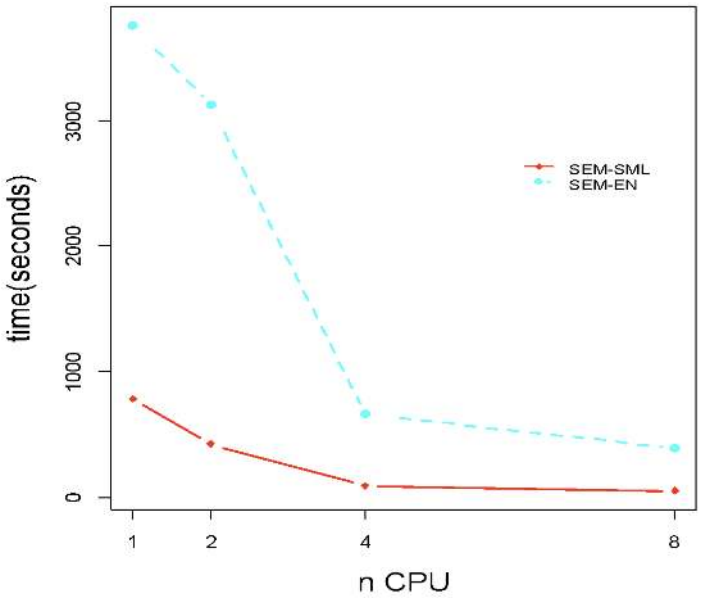

Figure 1: Scaling pattern for SEM-SML and SEM-EN

Simulation study

To evaluate the performance of SEM-EN, we compared PD and FDR with that of SEM-SML. If $\hat{B}_{ij} \neq 0$, then we consider there is an edge from gene j to gene i. the PD and FDR of the SEM-SML and SEM-EN for different sample sizes are depicted in Figure 2 and Figure 3 corresponding to two directed acyclic graphs (DAGs) with number of genes $N_g = 300$, expected number of edges per node $N_e = 3$ and residual variance $\sigma^2 = 0.01$ and $\sigma^2 = 0.05$, respectively. The result of Figure 2 and 3 represents mean PD and FDR for 100 replicates for each for the 10 different sample sizes. It is observed that SEM-EN achieves higher PD and similar FDR comparing with that of SEM-SML for both DAGs despite of different sample sizes. Moreover, it can be seen that the performance gain of SEM-EN is more significant for the DAG with larger noise level (Figure 3). Take $N = 500$ for example, the PD/FDR of SEM-EN for $\sigma^2 = 0.01$ are 0.9537/0.0433, comparing to 0.9005/0.0375 of SEM-SML. For $\sigma^2 = 0.05$ with the same sample size, the numbers are 0.9008/0.1300, and 0.7663/0.1182 for the two methods, respectively.

Inference of the yeast GRN

The gene expression profile of 3,380 ORFs and genetic marker of 1,162 cis-eQTLs were used for network inference by SEM-EN. Two parameters controlling degree of sparseness in the GRN need to be learnt from data. Cross validation (CV) identified the optimal shrinkage parameters as $(\alpha, \lambda) = (0.45, 0.0063)$ (see Methods section for the definitions of shrinkage parameters) for the SEM-EN method. With the pair of parameters, SEM-EN inferred a sparse GRN with 159 open reading frames (ORFs) involving 267 edges. The

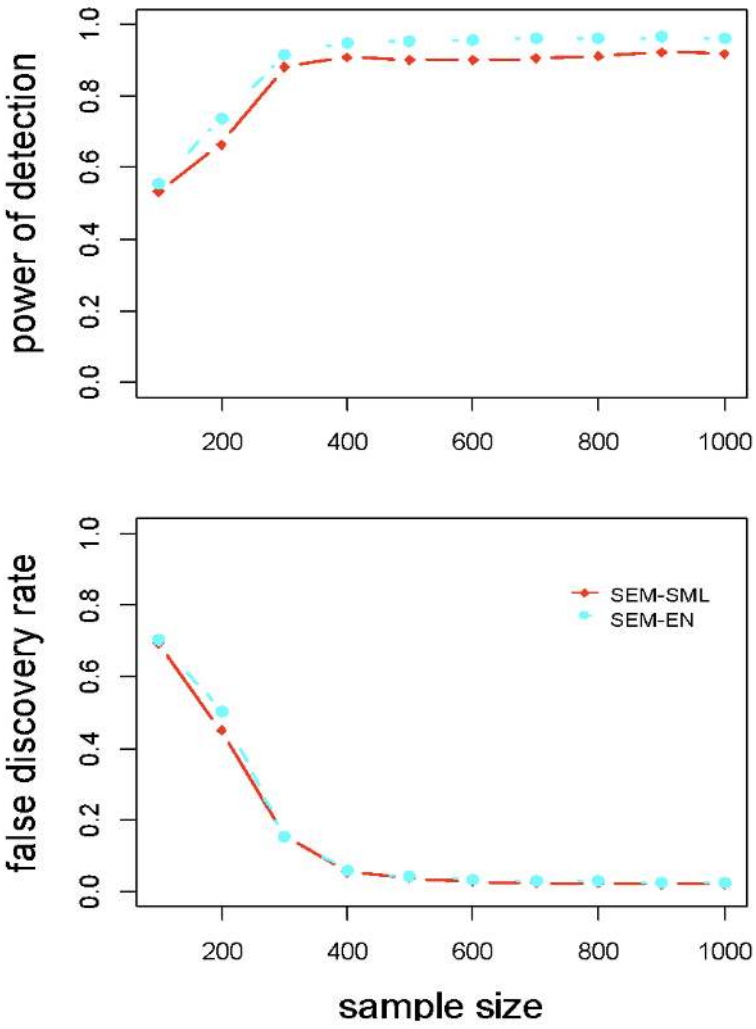

Figure 2: Performance of SEM-SML and SEM-EN for DAG simulation $\sigma^2 = 0.01$

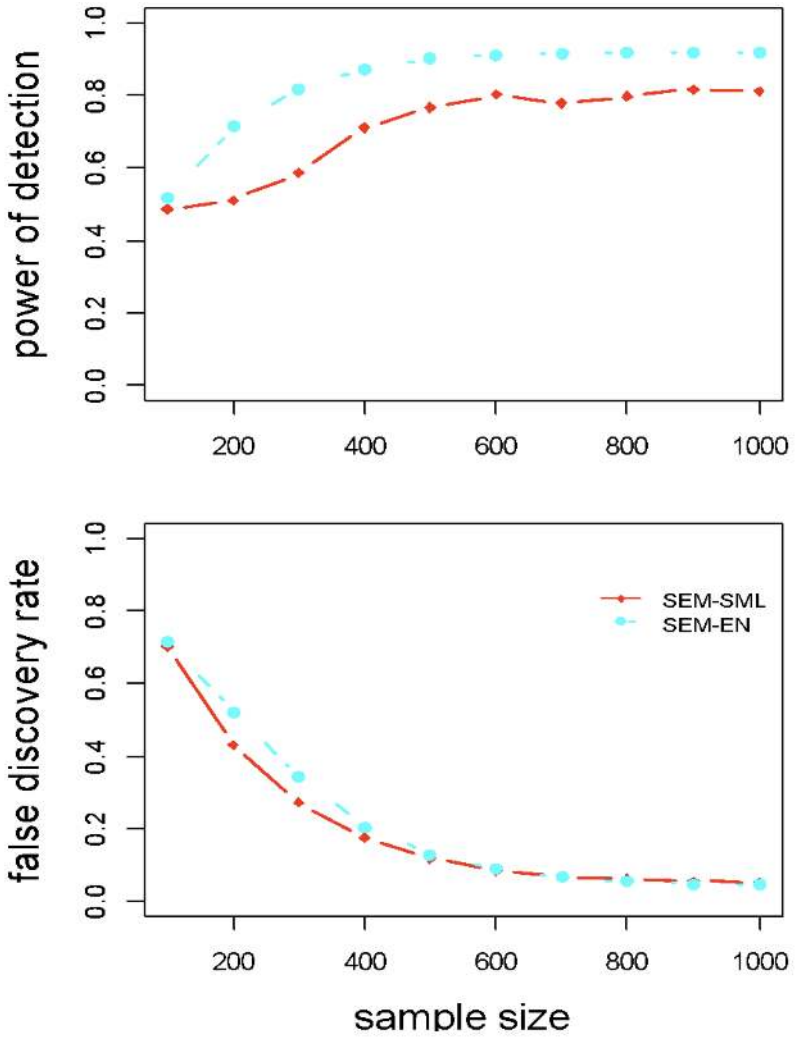

Figure 3: Performance of SEM-SML and SEM-EN for DAG simulation $\sigma^2 = 0.05$

Network parameters[a]	$N_g = 300, N_e = 3, \sigma^2 = 0.01$				$N_g = 300, N_e = 3, \sigma^2 = 0.05$			
	PD		FDR		PD		FDR	
Sample size	SML	EN	SML	EN	SML	EN	SML	EN
100	0.5333	0.5558	0.6927	0.7026	0.4871	0.5164	0.7019	0.7146
200	0.6642	0.7387	0.4507	0.5033	0.5098	0.7141	0.4317	0.5222
300	0.8811	0.9136	0.1562	0.1537	0.5875	0.8190	0.2734	0.3433
400	0.9068	0.9474	0.0558	0.0614	0.7120	0.8713	0.1751	0.2053
500	0.9005	0.9537	0.0375	0.0433	0.7663	0.9008	0.1182	0.1300
600	0.9007	0.9571	0.0292	0.0350	0.8028	0.9112	0.0859	0.0883
700	0.9049	0.9623	0.0244	0.0311	0.7795	0.9133	0.0685	0.0666
800	0.9117	0.9616	0.0230	0.0290	0.7979	0.9185	0.0602	0.0544
900	0.9224	0.9635	0.0213	0.0267	0.816	0.9203	0.0553	0.0482
1000	0.9170	0.9621	0.0203	0.0250	0.8119	0.9195	0.0519	0.0465

[a] PD and FDR were obtained from 100 replicates of the network analysis.

Figure 4: PD and FDR for SEM-SML and SEM-EN in analyzing the two simulated DAGs

network was visualized via Cytoscape [22] shown in Figure 5 and the positional information of the ORFs involved in the GRN was depicted in Figure 7 by Circos software [23].

It has been known that in GRN, sub-network are typically associated with particular biological functions [24]. In our study, five major clusters of the GRN was identified via Cytoscape [22] shown in different colors in Figure 5, and gene ontology (GO) term enrichment analysis was performed for each clusters by Gorilla [25] (Table 1/Figure 4). Specific to molecular functions, Cluster 1 (lime color in Figure 5) includes 28 ORFs and is enriched with aldehyde dehydrogenase (NAD) activity and transferase activity that transferring aldehyde or ketonic groups. Cluster 2 (teal color) includes 10 ORFs and is enriched with asparaginase activity and iron ion transmembrane transporter activity. Cluster 3 (slate blue color) contains 30 ORFs and is enriched with carbamoyl-phosphate synthase activity, oxidoreductase activity, NAD binding, ad dicarboxylic acid transmembrane transporter activity. Cluster 4 (olive color) contains 13 ORFs and is enriched with mating pheromone activity, DNA binding and bending, and RNA polymerase II transcription. Cluster 5 (red color) contains 16 ORFs and is enriched with glucosidase activity. The corresponding significant terms were shown in Figure 5.

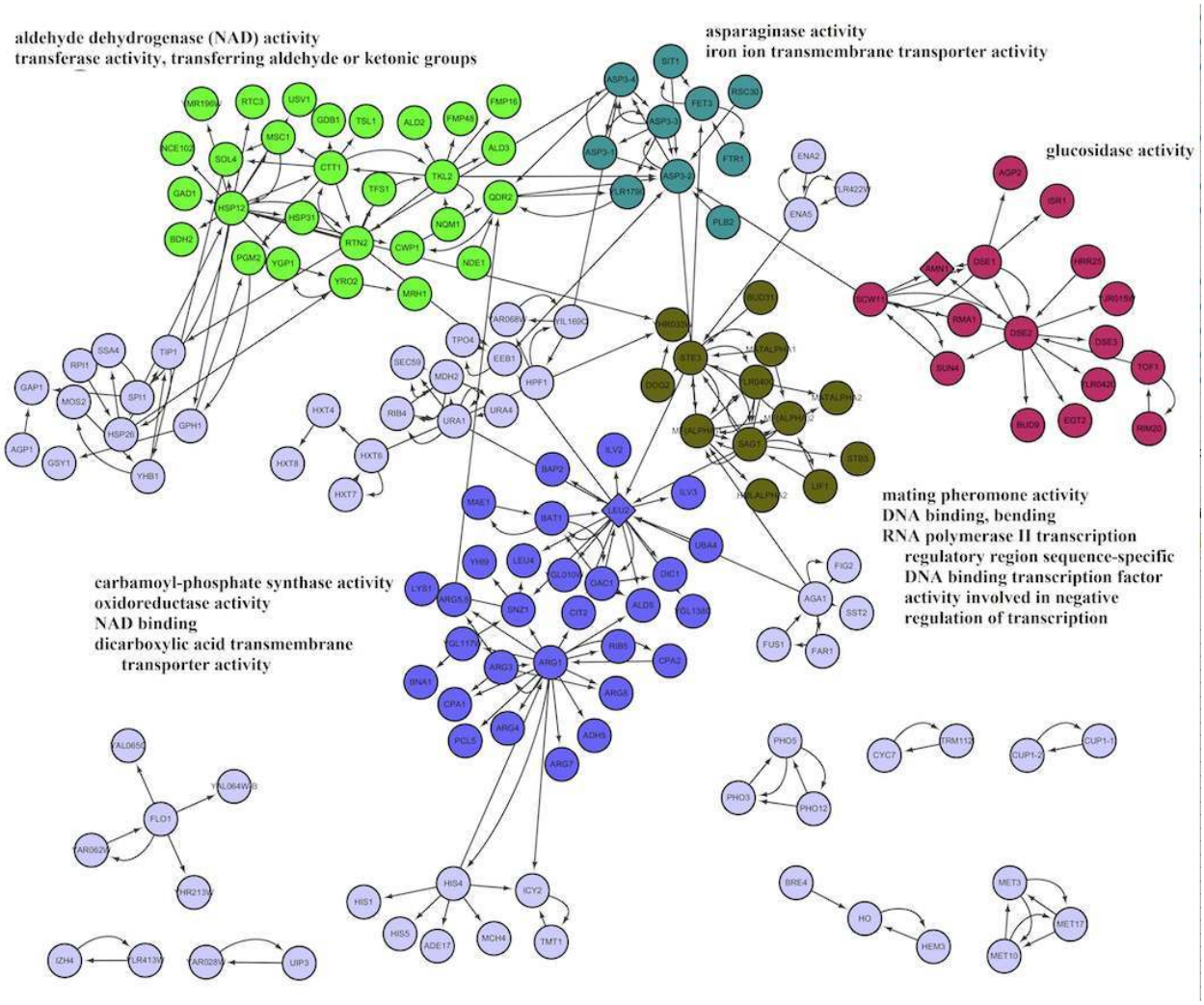

Figure 5: Sparse budding yeast GRN inferred by SEM-EN

Within a network cluster, nodes that are most relevant for the corresponding cluster function often have higher degree, meaning that there are more edges connected to them than other nodes [26]. Encoding the leucine biosynthetic enzyme, LEU2 is deleted in the RM parents in the segregants [27], and has been

Cluster	GO category (p-value)	Type
1	polyamine catabolic process (1.44×10-4)	biological processes
	beta-alanine biosynthetic process (2.41×10-5)	
	trehalose biosynthetic process (4.98×10-4)	
	pentose-phosphate shunt (4.88×10-5)	
	plasma membrane organization (1.44×10-4)	
	aldehyde dehydrogenase (NAD) activity (2.38×10-4)	molecular functions
	transferase activity, transferring aldehyde or ketonic groups (3.57×10-4)	
2	cellular response to nitrogen starvation (7.65×10-11)	biological processes
	asparagine catabolic process (2.55×10-11)	
	iron assimilation by reduction and transport (2.87×10-6)	
	high-affinity iron ion transport (8.59×10-6)	
	iron ion transmembrane transporter activity (6.70×10-7)	molecular functions
	asparaginase activity (2.55×10-11)	
	cell wall-bounded periplasmic space (3.57×10-10)	cellular components
	extracellular region (2.52×10-7)	
	high affinity iron permease complex (2.87×10-6)	
3	rginine biosynthetic process (1.09×10-17)	biological processes
	ornithine biosynthetic process (5.52×10-7)	
	leucine biosynthetic process (4.77×10-6)	
	isoleucine biosynthetic process (1.62×10-5)	
	valine biosynthetic process (2.74×10-6)	
	glyoxylate cycle (9.74×10-4)	
	dicarboxylic acid transport (5.72×10-4)	
	carbamoyl-phosphate synthase (glutamine-hydrolyzing) activity (8.28×10-5)	molecular functions
	oxidoreductase activity (2.47×10-4)	
	NAD binding (3.31×10-4)	
	dicarboxylic acid transmembrane transporter activity (9.74×10-4)	
	carbamoyl-phosphate synthase complex (2.77×10-5)	cellular components
	mitochondrial matrix (2.24×10-5)	
4	negative regulation of mating-type specific transcription from RNA polymerase II promoter (2.52×10-5)	biological processes
	pheromone-dependent signal transduction involved in conjugation with cellular fusion (1.48×10-5)	
	mating (4.20×10-6)	
	sexual reproduction (5.61×10-4)	molecular functions
	mating pheromone activity (1.49×10-7)	
	DNA binding, bending (1.30×10-5)	
	RNA polymerase II transcription regulatory region sequence-specific DNA binding transcription factor activity involved in negative regulation of transcription (6.31×10-4)	
	extracellular region (8.46×10-4)	cellular components
5	cytokinesis, completion of separation (4.13×10-6)	biological processes
	glucosidase activity (5.36×10-5)	molecular functions
	fungal-type cell wall (4.27×10-6)	cellular components
	cellular bud neck (8.49×10-4)	
	anchored component of membrane (6.21×10-4)	

[a] Among significant GO categories, only the most specific terms (leaf node of the GO hierarchies) are listed.

Figure 6: Enriched GO terms of the five clusters in yeast GRN

previously predicted as a causal regulator for the expression of a subset of genes by several independent studies [27-29]. In our GO term enrichment analysis shown in Table 1/Figure 4, leucine biosynthetic process is significant for Cluster 3 (p-value $= 4.77 \times 10^{-6}$), and LEU2 is identified as regional hub in for the cluster with 20 genes directly connected to it (Figure 5). The parent lab strain BY4716 has AMN1 gene deleted. $AMN1$ gene encodes a protein required for daughter cell separation, multiple mitotic checkpoints, and chromosome stability [27]. It has been verified in previous study that $AMN1$ gene is responsible for the regulation of a set of genes (SCW11, DSE1, DSE2, DSE3, DSE4, ISR1, PRY3, EGT2, SUN4, BUD9) [27]. Among them, SCW11, DSE1, DSE2, DSE3, ISR1, EGT2, SUN4, and BUD9 are in Cluster 5 in Figure 5. These results from previous biological experiments and independent research studies corroborate the strength of the SEM-EN algorithm.

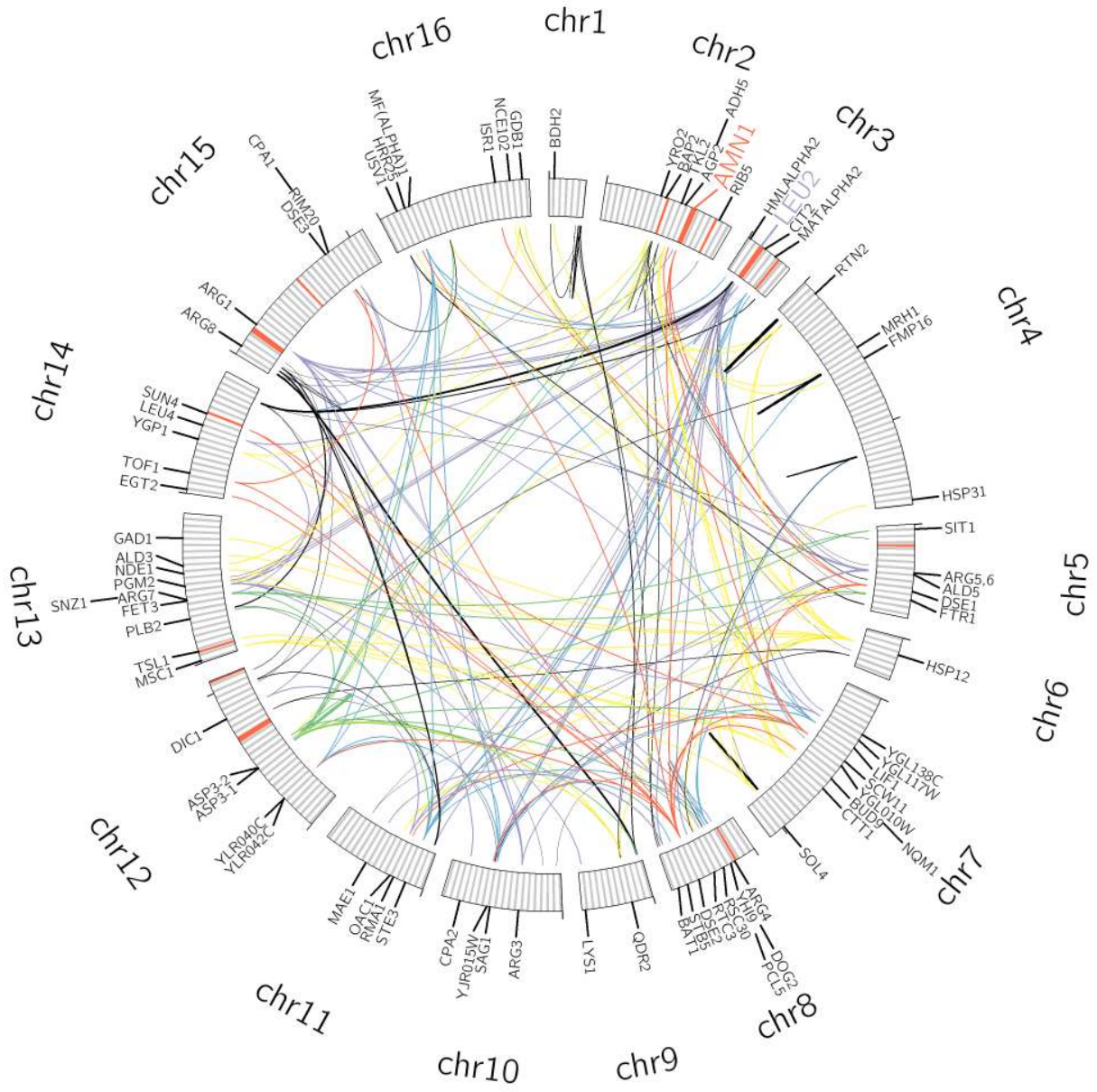

Figure 7: Budding yeast gene interaction pattern with link edges identified by SEM-EN

While the genetic engineering of $AMN1$ and LEU2 genes serves as a direct positive control for the GRN

eQTL hotspot	Position (chr: bp)	Common functions	Yvert et al. prediction[a]	SEM-EN regulators [b]
1	(2: 390,000)	Ribosome protein	None	*BAP2, TIP1, HSP26*
2	(2: 560,000)	DSE rRNA processing	*AMN1, MAK5*	*AMN1,ADH5*
3	(2: 710,000)	Unknown	None	*RIB5*
4	(3: 100,000)	Amino acid catabolism	*LEU2*	*LEU2,CIT2,FUS1*
5	(3: 230,000)	Mating	*MATALPHA1*	*MATALPHA1, MATALPHA2*
6	(5: 130,000)	Uracil catabolism	*URA3*	None
7	(8: 130,000)	Pheromone response	*GPA1*	*ARG4*
8	(12: 680,000)	Heme ferrochelatase	*HAP1*	None
9	(12: 1,070,000)	Subtelomeric	*SIR3*	*SST2*
10	(13: 70,000)	Unknown	None	*TSL1*
11	(14: 503,000)	Mitochondria	None	*SUN4*
12	(15: 180,000)	Msn2/4 targets	None	None
13	(15: 590,000)	Respiration	*CAT5*	None

[a] "None" indicates that no regulators were identified for the hot spot, *cis*-eQTL hotspots information are available from the supplemental information of [27].

Figure 8: eQTL hotspots identified in the original publication of the budding yeast dataset and predicted using SEM-EN

inference in budding yeast [27, 29, 30], the network structure inferred by SEM-EN shed light into the gene regulatory relationships. Previous study on single cis-eQTL mapping revealed 13 eQTL hot spots (Table 3/Figure 8), where eQTLs have pleiotropic effects on a number of expression traits [27]. The positions of eQTL hot spots can be visualized as the red ticks on the yeast chromosomes in Figure 7, and the regulator genes in the GRN inferred from SEM-EN are listed in Table 3/Figure 8. While the original study analyzed all 6,126 yeast ORFs and identified 8 regulators for the 13 eQTL hot spots, our study considered 3,380 ORFs that passed data quality control (see the Methods section for details) and identified regulators for 9 of the hot spots. Regulators such as $AMN1, LEU2$, and MATALPHA1 are consistent with previous study (Table 3/Figure 8). Particularly, SEM-EN identified regulators that are missed in previous study including RIB5 for hot spot 3 , TSL1 for hot spot 10 and SUN4 for hotspot 11. Among them, RIB5 is a gene encodes the riboflavin synthase that catalyzes the last step of the riboflavin biosynthesis pathway and involves in amino acid biosynthesis. RIB5 interacts with ARG1 in the SEM-EN identified GRN. Association of RIB5 and $ARG1$ has been characterized in previous study [31], and both are key proteins for cellular growth. TSL1 interacts with CTT1, both of which have been identified to be associated with cellular growth under stress [32]. SUN4 is a gene involved in cell wall separation [33], and interacts with DSE2 and SCW11 in the SEM-EN inferred GRN (Figure 5). DSE2 is a daughter cell-specific secreted protein that plays a key role in daughter cell separation. During the separation process, DSE2 degrades cell wall from the daughter side and causes daughter cell to separate from the mother cell [34]. SCW11 is a cell wall protein that plays a key role in conjugation during mating [35], and its functional association with DSE2 have been experimentally characterized [36]. Given the genetic perturbation background in daughter cell separation and amino acid biosynthesis, previous molecular experiment results corroborate the predictive power of the inferred GRN and the strength of the SEM-EN method.

Back to Top

Discussion

Understanding the multiple levels of gene regulations is the key for the prediction of complex cellular behavior. Integrating genetic perturbations with gene expression data has been proven to be more accurate in learning causal regulatory relation among genes comparing to treating each gene expression level as an individual quantitative trait [12]. The SEM-SML has been shown to be able to integrate such information to infer GRN systematically and offer significant better performance than state-of-the-art algorithms. We extended the SEM-SML to incorporate adaptive elastic net penalty [16] for the likelihood function of the SEMs, and implemented the SEM-EN software in efficient C/C++ with parallel computational capability by MPI. Simulation studies demonstrated that SEM-EN is capable of inferring a large network within affordable computational time while achieving high power of detection than SEM-SML. The software is further applied to systematically infer the GRN in budding yeast.

The work in this paper improved the SEM-SML algorithm [12] from two directions. First, while previous work was implemented in MATLAB, the SEM-SML algorithm and the new SEM-EN algorithm were implemented in C/C++ with the fast basic linear algebra subprograms (BLAS) and linear algebra package (LAPACK) [37] in this paper. Simulation demonstrated that computational time for SEM-SML in C/C++ was reduced by more than 10 times compared to the one implemented in MATLAB using a single CPU node. To achieve computational feasibility for inferring large network with thousands of nodes, the block coordinate ascent technique in Algorithm 1 of [12] is parallelized with Open MPI [38]. A strong scaling pattern for the parallel implementation was observed and depicted in Figure 1, and the new software is capable of inferring a network structure more than 100 times faster. For example, while the MATLAB implementation takes several days to infer a network structure with 300 of nodes, and it is not realistic for it to infer a network structure with thousands of node, in the yeast GRN analysis, we demonstrated that SEM-EN is able to analyze a network with more than 3,000 nodes.

Second, with the adaptive elastic net [16, 17] in place of l_l-norm penalty for the SEM likelihood function, we showed that SEM-EN achieves higher PD while controlling a similar FDR to that of SEM-SML (Figures 2, 3). Superior performance of SEM-EN over SEM-SML is expected due to two reasons. First, it is known that l_1-norm penalty in regularized linear regression typically keeps only one out a group of correlated effects [16]. However, in cellular metabolisms, a gene regulator typically can shape the expression profile of a set of genes, resulting in highly correlated gene expression patterns [1, 2]. Mathematically, it can be shown from SEM that gene expression levels have covariance $\text{cov}(\mathbf{Y}) = \left[(\mathbf{I}-\mathbf{B})^{-1}\right]^T \sigma^2\mathbf{I}(\mathbf{I}-\mathbf{B})^{-1}$, which is not diagonal. Therefore, the performance gain of SEM-EN over SEM-SML is expected given strong correlation among gene expression levels. Second, SEM-EN improves inference accuracy by taking the grouping effect into account, since it has been known that elastic net penalty enjoys a strong grouping effect than the l_l-norm penalty in linear regression. In fact, SEM-SML becomes a special case of SEM-EN with parameter $\alpha = 1$.

In inferring the GRN of budding yeast, SEM-EN integrates genetic perturbations with genome-wide expression data. On the one hand, for the set of regulation that have been verified in previous studies such as the set of genes regulated by $AMN1$ and $LEU2$, the GRN obtained by SEM-EN is in line with previous findings, which corroborates the strength of the SEM-EN method. On the other hand, with the grouping effect encouraged by SEM-EN, we were also able to identify other ORFs interacting with known regulators. For example, seven ORFs in Custer 5 was not reported in the list of [27] that linked to AMN1. Among them, $RMA1$ interacts with $DSE2$ and $SCW11$, both are directly linked to $AMN1$; $AGP2$ directly interacts with DSE 1 and has a distance of 2 to $AMN1$, and TOF1 and HRR25 both interacts with DSE2. The genes in the cluster were known to specifically expressed in daughter cells during budding [36], and it may be worthy of experimental investigation to further study their roles in gene regulation affecting daughter-cell separation after budding.

Back to Top

References

1. Civelek M, Lusis AJ: Systems genetics approaches to understand complex traits. Nat Rev Genet 2014, 15:34-48.

2. Nuzhdin SV, Friesen ML, McIntyre LM: Genotype-phenotype mapping in a post-GWAS world. Trends in Genetics 2012, $28 : 421 - 426$

3. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS: Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci USA 2000, 97:12182-12186

4. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: Reverse engineering of regulatory networks in human B cells. Nat Genet 2005, 37:382-390.

5. Friedman N, Linial M, Nachman I, Pe'er D: Using Bayesian networks to analyze expression data. *J* Comput Biol 2000, 7:601-620

6. Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M: Sparse graphical models for exploring gene expression data. *J* Multivar Anal 2004, 90:196-212

7. Schafer J, Strimmer K: An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics $2005, 21 : 754 - 764$

8. Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V: The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome Biol 2006, 7:R36.

9. di Bernardo D, Thompson MJ, Gardner TS, Chobot SE, Eastwood EL, Wojtovich AP, Elliott SJ, Schaus SE, Collins JJ Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. Nat Biotechnol 2005, 23:377-383

10. Gardner TS, Di Bernardo D, Lorenz D, Collins JJ: Inferring genetic networks and identifying compound mode of action via expression profiling. Science 2003, 301:102-105

11. Schafer J, Strimmer K: A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol $2005, \mathbf{4} : 32$.

12. Cai X, Bazerque JA, Giannakis GB: Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. PLoS Comput Biol 2013, 9:e1003068.

13. Jeong H, Mason SP, Barabasi A-L, Oltvai ZN: Lethality and centrality in protein networks. Nature 2001, 411:41-42.

14. Tegner J, Yeung MS, Hasty J, Collins JJ: Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. Proc Natl Acad Sci USA 2003, 100:5944-5949.

15. Thieffry D, Huerta AM, Perez-Rueda E, Collado-Vides J: From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. Bioessays 1998, 20:433-440.

16. Zou H, Hastie T: Regularization and variable selection via the elastic net. J Roy Stat Soc B Met 2005, 67:301-320.

17. Zou H, Zhang HH: On the adaptive elastic-net with a diverging number of parameters. Ann Stat 2009 , 37:1733.

18. Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G: Wisdom of crowds for robust gene network inference. Nat Meth 2012, 9:796-804

19. Slonim DK, Yanai I: Getting started in gene expression microarray analysis. PLoS Comput Biol 2009, 5:e1000543

20. Quinn MJ: Parallel Programming. TMH CSE; 2003.

21. Akl SG: Parallel computation: models and methods. Prentice-Hall, Inc.; 1997.

22. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res 2003, 13:2498-2504.

23. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: Circos: an information aesthetic for comparative genomics. Genome Res 2009, $19 : 1639 - 1645$.

24. Langfelder P, Mischel PS, Horvath S: Whenis hub gene selection better than standard meta-analysis? PloS one 2013, $8 : e61505$

25. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z: GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 2009, $\mathbf{10} : 48$

26. Almaas E: Biological impacts and context of network theory. Journal of Experimental Biology 2007, 210:1548-1558.

27. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L: Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat Genet 2003, 35:57-64

28. Zhu J, Sova P, Xu Q, Dombek KM, Xu EY, Vu H, Tu Z, Brem RB, Bumgarner RE, Schadt EE: Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. *PLoS* Biol 2012,10:e1001301.

29. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, Bumgarner RE, Schadt EE: Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat Genet 2008, 40:854-861.

30. Brem RB, Yvert G, Clinton R, Kruglyak L: Genetic dissection of transcriptional regulation in budding yeast. Science 2002 296:752-755

31. Creighton C, Hanash S: Mining gene expression databases for association rules. Bioinformatics 2003, 19:79-86.

32. Meadows R: Yeast survive by hedging their bets. PLoS Biol 2012, 10:e1001327.

33. Mouassite M, Camougrand N, Schwob E, Demaison G, Laclau M, Guerin M: The $\boldsymbol{SUN}$, family: yeast $\boldsymbol{SUN}$, $4/SCW3$ is involved in cell septation. Yeast $2000, 16 : 905 - 919$

34. Klis FM, Mol P, Hellingwerf K, Brul S: Dynamics of cell wall structure in Saccharomyces cerevisiae. FEMS microbiology reviews $2002, \mathbf{26 : 239 - 256}$

35. Zeitlinger J, Simon I, Harbison CT, Hannett NM, Volkert TL, Fink GR, Young RA: Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. Cell 2003, 113:395-404.

36. Colman-Lerner A, Chin TE, Brent R: Yeast $\boldsymbol{Cbk1}$ and $\boldsymbol{Mob2}$ activate daughter-specific genetic programs to induce asymmetric cell fates. Cell $2001, 107 : 739 - 750$

37. Anderson E: LAPACK Users' Guide. Society for Industrial and Applied Mathematics; 1999.

38. Gabriel E, Fagg G, Bosilca G, Angskun T, Dongarra J, Squyres J, Sahay V, Kambadur P, Barrett B, Lumsdaine A, et al: Open MPI: goals, concept, and design of a next generation MPI implementation. In Recent Advances in Parallel Virtual Machine and Message Passing Interface. Volume 3241. Edited by Kranzlmuller D, Kacsuk P, Dongarra J: Springer Berlin Heidelberg; 2004: 97-104: Lecture Notes in Computer Science].

39. Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J, Tibshirani RJ: Strong rules for discarding predictors in lasso-type problems. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2012, 74:245-266.

40. Brem RB, Kruglyak L: The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proceedings of the National Academy of Sciences of the United States of America 2005, 102:1572-1577

41. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: Sequencing and comparison of yeast species to identify genes and regulatory elements. Nature 2003, **423** : 241 − 254

Back to Top