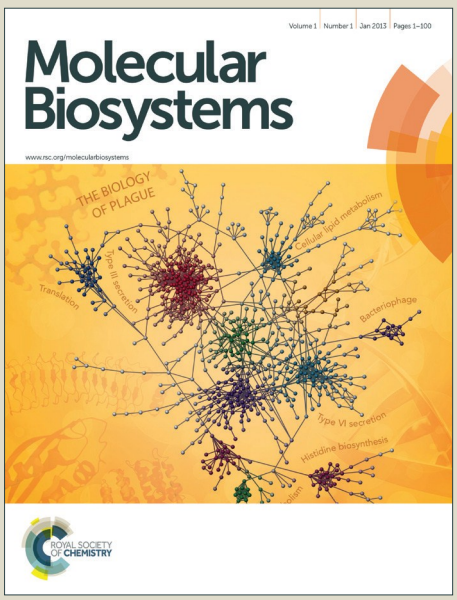
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Journal Name

ARTICLE TYPE

Cite this: DOI: 10.1039/xxxxxxxxxx

Large scale gene regulatory network inference with a multi-level strategy†

Jun Wu,[a] Xiaodong Zhao,[b] Zongli Lin *[c] and Zhifeng Shao[d]

Received Date
Accepted Date

DOI: 10.1039/xxxxxxxxxx

www.rsc.org/journalname

Transcriptional regulation is a basis of many crucial molecular process and an accurate inference of the gene regulatory network is a helpful and essential task to understand cell functions and to gain insights into biological process of interest in System Biology. Inspired by the Dialogue for Reverse Engineering Assessments and Methods (DREAM) projects, many excellent gene regulatory network inference algorithms have been proposed. However, it is still a challenging problem to infer gene regulatory network from gene expression data on a large scale. In this paper, we propose a gene regulatory network inference method based on a multi-levels strategy (GENIMS), which can give results that are more accurate and robust than the state-of-the-art methods. The proposed method mainly consists of three levels, which are original feature selection step based on guided regularized random forest, normalization of individually feature selection and the final refinement step according to the topological property of gene regulatory network. To prove the accuracy and robustness of our method, we compare our method with the state-of-the-art methods on the DREAM4 and DREAM5 benchmark networks and the results indicate that the proposed method can significantly improve the performance of gene regulatory network inference. Additionally, we also discuss the influence of the selection of different parameters in our method. The source code of the proposed method is provided as a supplementary material to the paper.

Introduction

Transcriptional regulation is a basis of many crucial molecular process and it can control the response of the cell or organism to various intra- and extracellular signals through orchestrating the expression of genes[1]. To understand cell functions and to gain insights into biological processes of interest, an accurate gene regulatory network inference is a helpful and essential task in System Biology. Recently, with the rapid development of high-throughput technologies, such as RNA-seq and DNA microarrays, the availability of mounts of gene expression collections make it possible to infer the gene regulatory network topology on a large-scale. On the other hand, various computational methods have been proposed to facilitate the development of gene regulatory network inference.

However, inferring the gene regulatory network from gene expression data remains a challenging problem, as the data are typically noisy and high dimensional, and the information on the topology of transcriptional networks is incomplete[2]. Moreover, because the amount of potential interaction largely exceeds the number of available observations, inferring the large scale gene regulatory network from gene expression data is a badly underdetermined problem [3,4]. To inspire the development of gene regulatory network inference algorithm, the DREAM projects provide a set of benchmark networks that can be used to compare the performance of various gene regulatory network algorithms[5–9]. Besides simulating the benchmark network and gene expression characteristics from two well-studied systems, *E. coli* and *S. cerevisiae*, the DREAM projects provide real gene expression data for large scale gene regulatory network inference competition. Many computational methods have been introduced to solve the problem of gene regulatory network inference, such as those based on probabilistic graphical model, ordinary differential equation and information theory. Most existing methods for gene regulatory network inference are based on the assumption that the expression of a target gene can be modeled as a function of the expressions of the potential regulator genes. These methods focus on discovering the potential interactions which can explain the functional relationship of the expression data and the information provided by these methods can give us crucial mechanism

[a] *Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing of Ministry of Education, Shanghai, China; E-mail: junwu302@gmail.com.*
[b] *School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China; E-Mail: xiaodong122@yahoo.com.*
[c] *Charles L. Brown Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, Virginia, United States of America; Email: zl5y@virginia.edu.*
[d] *School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China; E-Mail: zs9q@virginia.edu.*

details of transcriptional regulation[10].

The probabilistic graphical model based methods, such as Bayesian network[11–13] and graphical Gaussian model[14,15], employ the graphical modeling techniques and the probability knowledge to infer the gene network. These methods mainly describe the gene network by a graph and then learn the parameters of the models through analyzing the multivariate joint probability distributions over the observations. However, learning the parameters from the expression data is computationally intensive, especially when the number of parameters is large. As a result, the methods based on probabilistic graphical model are unsuitable for the large scale gene regulatory network inference. The ordinary differential equation based methods[16–18] are mainly designed for the time series gene expression data and model the dynamics of a regulatory network by a set of ordinary differential equations. The ordinary differential equation based methods are the best analysis approaches for non-linear systems to analyze network dynamics, such as to locate limit cycles and to investigate bifurcation behaviour. However, as the ordinary differential equation based methods have high-dimensional parameter spaces, a large amount of experiment data is required. Without the disadvantages of the methods described above, such as unsuitability for large-scale network inference and computational intensiveness, the information theory based methods have led to several robust and reliable algorithms for gene network inference and have emerged as a standard in this field[19–22]. The information theory based methods mainly depend on the mutual information between the expressions of all pairs of genes to infer the interaction. Several improved methods, such as ARACNE[21], CLR[23] and MRNET[22,24], have also been proposed to remove the indirect and reduce redundancy interactions. The ARACNE algorithm utilizes the data processing inequality to remove the indirect connections from triplets of genes, while the CLR algorithm utilizes the adaptive background correction algorithm to modify the value of mutual information to improve the precision. The MRNET algorithm uses a forward selection strategy to identify maximally independent set of neighbors for each gene.

Recently, machine learning theory based methods, such as the TIGRESS and GENIE3 methods, developed for supervised feature selection, have been introduced to solve the problem of gene regulatory network inference. Both the TIGRESS[25] and GENIE3 methods[26] decompose the gene network inference problem into separate regression problems with respect to the target genes. The main difference between the two methods lies in the algorithm to solve the regression problems. The TIGRESS method solves the regression problems with the least angle regression algorithm combined with stability selection, while the GENIE3 method uses the random forest to solve the regression problems. Most strikingly, the GENIE3 method won the best performance in both the DREAM4 *in silico* 100 multifactorial challenge and the subsequent DREAM5 network inference challenge. Inspired by the idea of decomposing the gene network inference problem into separate regression problems, several other effective methods also proposed in recent years[3,10,27,28].

The method proposed in this paper also inherits the idea of decomposing the gene network inference problem into individual regression problems and the algorithm used to solve the regression problems is the guided regularized random forest algorithm. To further improve the performance of inference, we propose a multi-level strategy which consists of three levels. In the first level, the guided regularized random forest algorithm is used to solve the individual regression problems. In the second level, the results returned in the previous level are normalized. In the last level, the results are refined according to the topology property of the large scale gene regulatory network. The benchmark networks provided by the DREAM4 and DREAM5 projects are used to evaluate the performance of our proposed method. Through comparison with the state-of-the-art methods, our method is proven to perform more accurately and robustly.

Method

Problem statement

In this paper, we focus on recovering the network solely from multifactorial perturbation data. The multifactorial perturbation data can be obtained from the steady states of series of different perturbation experiments. A typical source of the perturbation data is gene expression profiles obtained from different patients. Let the recovered gene regulatory network is a directed graph with p nodes, each of which represents a gene. An edge directed form one gene i to another gene j indicates that gene i directly regulates the expression of gene j.

Suppose that we have a set of gene expression data, including p genes and n samples, which can be expressed by a $p \times n$ matrix: $D = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]^{\mathrm{T}}$, where $\mathbf{x}_i \in \Re^n, i = 1, 2, \ldots, p$, is the expression vector of genes i across all the samples: $\mathbf{x}_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n})^{\mathrm{T}}$. We need to design a gene regulatory network inference algorithm that utilizes the gene expression data to predict the regulatory interaction graph, which is represented by an adjacency matrix $W = \{w_{i,j} : w_{i,j} \geq 0\}, i, j = 1, 2, \ldots, p$. The value of $w_{i,j}$ is used to assess the confidence of the regulation relationship between gene i and gene j. A larger $w_{i,j}$ indicates that the predicted regulatory interaction between gene i and gene j is more reliable.

We decompose the inference of a gene regulatory network, which contains p target genes, into p independent regression problems. In each regression, the interaction between the regulator genes, such as transcription factors, and the target gene will be inferred. For the jth regression, the inputs are the expression vector of target gene j and the expression vectors of potential regulator genes across all the samples (we suppose that the potential regulator genes belong to the set of the p target genes). The aim of the regression analysis is to infer an unknown function f_j that discovers the relationship between the expression of the target gene and the expressions of the potential regulatory genes as follows,

$$f_j : \mathbf{x}_j = f_j(\mathbf{x}_i) + \varepsilon_j, \quad \forall j \in \{1, 2, \ldots, p\}, \tag{1}$$

where $i \in \{1, 2, \ldots, p : i \neq j\}$ and ε_j is a random noise. We desire to select a small subset of genes, from which f_j can give an optimum regression result and the selected genes are supposed to be true regulators of gene j.

Random forest and regularized random forest algorithms

Random forest is an ensemble of multiple decision trees and is one of the most successful supervised learning models for classification and regression. The gene regulatory network inference method GENIE3, which won the best performance in both the DREAM4 *in silico* 100 multifactorial challenge and the subsequent DREAM5 network inference challenge, was developed based on the random forest algorithm. The random forest algorithm can provide the information gain for each variable selection. Suppose that in one tree, variable x is used to split the tree at node v, the information gain of x at node v can be calculated as follows,

$$Gain(x,v) = H(x,v) - \alpha^l H(x,v^l) - \alpha^r H(x,v^r), \tag{2}$$

where $H(x,v)$ is the Shannon entropy of x at node v, v^l (v^r) is the left (right) child node of v, and α^l (α^r) is the proportion of observations assigned to the left (right) child node in the input observation of node v. The importance score of variable x_i can be obtained through averaging all the information gains of x_i in each decision tree,

$$I_x = \frac{1}{n_{\text{tree}}} \sum_{v \in V_x} Gain(x,v), \tag{3}$$

where n_{tree} is the number of decision trees used for the ensemble and V_x refers to the set of all the nodes split by variable x in each decision tree.

The regularized random forest algorithm introduces a regularized constraint to penalize the redundant and unimportant selections which are the disadvantages of the traditional random forest algorithm. The regularized constraint is employed in the information gain calculation step. In particular, the regularized information gain is computed as follows,

$$Gain_{\text{R}}(x,v) = \begin{cases} \lambda Gain(x,v) & x \in F, \\ Gain(x,v) & x \notin F, \end{cases} \tag{4}$$

where F is the set of variables used for splitting in the previous nodes, which is an empty set at the root node in the first tree, and $\lambda \in (0,1]$ is the regularization parameter used for the penalty adjustment, with a larger λ leading to less penalization.

Gene regulatory network inference with the guided regularized random forest algorithm

In inferring the gene regulatory network with regularized random forest algorithm, as described above, a key factor is the selection of the regularization parameter λ. Following the suggestion in [29], we use the preliminary random forest results as a guideline for the λ selection. We penalize the genes with larger importance scores in the preliminary result less and vice versa. For ease of description, we represent the assignment of the regularization parameters with a $p \times p$ matrix $\Lambda = \{\lambda_{i,j}\}$, $i,j = 1,2,\ldots,p$. In the jth regression, suppose that gene j is selected as the target gene and other genes are regarded as potential regulator genes. With the preliminary random forest result, we obtain an importance score $\lambda_{i,j}$, $i \neq j$, for potential regulator gene i. After all regressions, we set the regularization parameter matrix Λ as follows,

$$\lambda_{i,j} = \gamma \frac{\lambda_{i,j} - \min\{\Lambda\}}{\max\{\Lambda\} - \min\{\Lambda\}}, \tag{5}$$

where $0 < \gamma \leq 1$ is a constant used to control the global penalization degree and $\min\{\Lambda\}$ and $\max\{\Lambda\}$ are the minimal and maximal values among elements of matrix Λ. For further details, we explain these procedure as follows.

Algorithm 1 The regularization parameters assignment

Input: $p \times n$ gene expression matrix D; parameter control coefficient γ.
output: $p \times p$ regularization parameter matrix Λ.
Initialize: Initialize Λ with 0; $\gamma = 0.7$.

1. **for** each gene j, $j \in \{1,2,\ldots,p\}$ **do**
   (a) Select all the potential regulator genes.
   (b) Importance score vector is obtained through random forest regression.
   (c) Assign these importance scores to the corresponding positions in the jth column of Λ.
   end for
2. Obtain the final Λ with equation 5.

Once the assignment of regularization parameters is achieved, the regularized random forest algorithm is applied to solve the p independent regression problems. We can construct a $p \times p$ adjacency matrix $W = \{w_{i,j}\}$, which represents the graph of the inferred gene regulatory network after all the regression problems are solved. The value $w_{i,j}$ reflects the confidence level of the edge outgoing from a potential regulator gene i to a target gene j ($i \neq j$). The adjacency matrix W was obtained by stacking all the p independent regression solutions column by column.

Normalization and network refinement

In most published methods, such as the TIGRESS and GENIE3 methods, edges among all gene pairs are ranked according to their confidence level to form the final gene regulatory network. However, as the p regression problems are solved individually, we hold the opinion that the results should be normalized to make them comparable to each other. In this paper, the q-norm based method [27] is introduced to address this issue and the the normalized weight matrix $\hat{W}$ is calculated as follows,

$$\hat{w}_{i,j} = \frac{w_{i,j}}{\left(\sum_{j=1}^{p} w_{i,j}^q\right)^{1/q}}. \tag{6}$$

It is important to note that the diagonal elements are not considered in the normalization. Our experimental studies, which are described in detail in the results section, indicate that the inference performance is improved more significantly if q is selected in the range $[2,4]$.

Additionally, we also apply a refinement procedure to improve the accuracy of our method. The main assumption is that the

gene regulatory network is sparse, which means that only a small subset of the target genes are regulated by a regulator gene. In other words, the confidence levels of the edges outgoing from a regulator gene should be easily distinguishable. Based on this assumption, the edges outgoing from a regulatory gene with a better distinguishability may be more important and reliable. As the value of $\hat{w}_{i,j}$ reflects the confidence level of the edge outgoing from regulator gene i to target gene j, the ith row of $\hat{W}$ contain the confidence levels of the regulation relationships between the regulator gene i and all the target genes. The distinguishability of regulator gene i can be measured by the variance of the ith row of $\hat{W}$, which is then used as a guideline of the refinement procedure. The refined adjacency matrix $\widetilde{w}_{i,j}$ can be obtained as follows,

$$\widetilde{w}_{i,j} = \sigma_i^2 \cdot \hat{w}_{i,j}, \tag{7}$$

where σ_i^2 is the variance of the ith row of $\hat{W}$.

Summary: the GENIMS algorithm

The overall GENIMS algorithm is summarized as follows. First, the gene regulatory network inference problem is divided into several individual regression problems and the guided regularized random forest algorithm is introduced to solve these regression problems. The solutions of these regression problems are combined to form the first level's results of the proposed method. Second, considering the regression problems are solved separately, we use the q-norm method to normalize these solutions to make them comparable. The normalized results can be considered the second level's results. At last, we give a refinement step to further improve the performance based on the sparsity assumption. The variance of the normalized weights of each regulator gene are used as a guide in this step to achieve the third level's results.

Results

Data source and performance evaluation

In this paper, we mainly use two series of benchmark networks, the DREAM4 *in-silico* multifactorial networks and the DREAM5 networks, to evaluated the proposed method and all the methods involved in comparisons. The detailed information of these networks are listed in Table 1.

The results of the inference algorithms are compared with the gold standard structure of networks provided by the DREAM projects organizers. Two evaluation metrics are considered, the area under the precision-recall curve (AUPR) and the area under the receiver operating characteristic curve (AUROC). The p-values of the two metrics measure the probability of the results if a random prediction is equal to or better than the result of the proposed method. For each network, the score of the prediction performance can be defined as,

$$score_k = -0.5\log_{10}\left\{p^k_{\text{AUPR}} \times p^k_{\text{AUROC}}\right\}, k = 1,2,\ldots,N, \tag{8}$$

where N is the number of networks, p^k_{AUPR} and p^k_{AUROC} indicate the p-values of AUPR and AUROC, respectively. Obviously, the overall score $score_{\text{all}}$ of the prediction performance on all the networks can be computed as their mean,

$$score_{\text{all}} = \frac{1}{N}\sum_{k=1}^{N} score_k. \tag{9}$$

Parameter selection

There are two parameters to consider in performing our proposed method. They are the global penalization parameter γ and the normalization parameter q. In what follows, we will discuss the influence of the selection of these two parameters to the final performance of the proposed method. The five DREAM4 networks will be used as the benchmark networks.

We use cross validation to choose the values of γ and q. Ten values of γ are chosen uniformly over the range of $[0.1,1]$ and one hundred values of q are chosen uniformly over the range of $[1,20]$. For each parameter combination and each benchmark network, we perform the GENIMS method and obtain the corresponding prediction score. The results are shown in Fig. 1, in which we can see that the value of the global penalization parameter γ does not significantly influence the network inference performance, while the value of the normalization parameter q is crucial to the performance of the algorithm performance. We observe that our algorithm performs well for each network when the value of normalization parameter q is larger than 2 and less than 4. In this paper, we set $q = 3$ and $\gamma = 0.7$ as the default values.

Performance evaluation on the DREAM Dataset

We first evaluate the proposed method with the five DREAM4 *in-silico* multifactorial networks. Each of the five networks consists of 100 genes and each network has been obtained through extracting some important and typical modules from actual biological networks of *E. coli* and *S. cerevisae*. We compare the proposed method with seven state-of-the-art methods, the ARACNE, CLR, MRNET, TIGRESS, GENIE3, NIMEFI and ENNET methods. The ARACNE, CLR and MRNET methods are implemented by the *minet* R package[22] with the default paramters, while the TIGRESS, GENIE3, NIMEFI and ENNET methods are implemented by their respective authors. As the NIMEFI method is the ensemble of several different methods, we only select the results of the G+E-SVR+E-EL version to represent the performance of the NIMEFI method. The details of the comparison are shown in Table 2. The values of AUROC, AUPR, and the corresponding p-values are given for each network. The overall score for each algorithm is also given in the last column and the best results for each column are typed in bold. In this table we can see that the performance of the proposed method ranks top in the overall score and performances best in all but one instance.

The second evaluation is based on the DREAM5 networks. There are four networks provided in the DREAM5 network inference challenge. AS no verified TF-TG interaction is provided for the second network, we only use the other three benchmark networks to evaluate the performance of the gene regulatory network inference methods. These three networks are different in size and structure. The expression of the first network is simulated using the GeneNetWeaver simulator[30]. The expression data

Table 1 Details of the data source

Name	#Networks	#Samples	#Genes	#Transc Factors	Type
DREAM4 *in-silico* multifactorial	5	100	100	100	Artificial
DREAM5 *in-silico*	1	805	1643	195	Artificial
DREAM5 *E. coli*	1	805	4511	334	Real
DREAM5 *S. cerevisae*	1	536	5950	333	Real

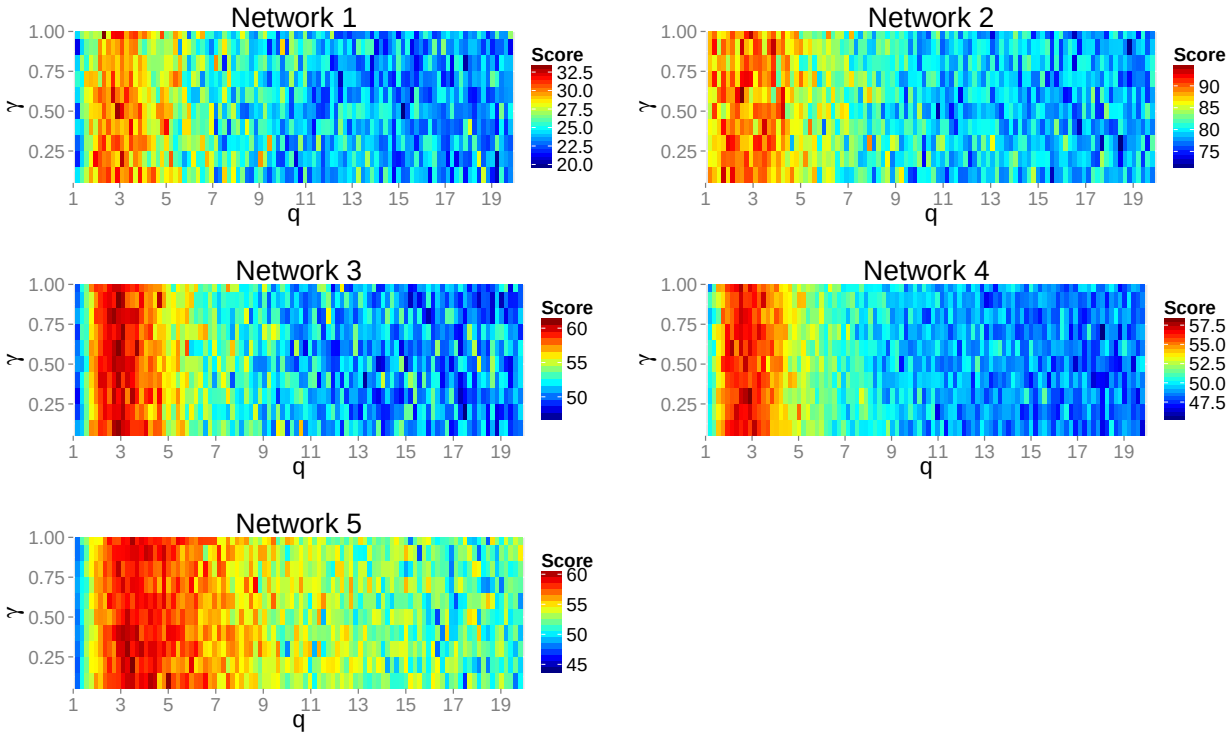


Fig. 1 The influence evaluation of the parameters selection. The x-axis indicates the selection of normalization parameter q, the y-axis indicates the selection of global penalization parameter γ and the color indicates the prediction performance.

Table 2 Results of different methods on DREAM4 *in-silico* multifactorial networks

Method	Metric	Network 1	Network 2	Network 3	Network 4	Network 5	Overall Score
ARACNE	auroc	0.616	0.574	0.664	0.643	0.654	23.554
	p.auroc	$1.67E-6$	$6.99E-5$	$1.06E-13$	$1.02E-10$	$1.46E-11$	
	aupr	0.130	0.107	0.222	0.180	0.190	
	p.aupr	$1.70E-28$	$7.92E-32$	$2.03E-52$	$2.83E-40$	$2.03E-42$	
CLR	auroc	0.685	0.687	0.731	0.713	0.729	26.887
	p.auroc	$4.47E-12$	$7.77E-20$	$8.62E-25$	$7.04E-20$	$4.86E-22$	
	aupr	0.131	0.110	0.174	0.172	0.170	
	p.aupr	$1.36E-28$	$4.32E-33$	$2.85E-40$	$3.97E-38$	$1.99E-37$	
MRNET	auroc	0.675	0.691	0.740	0.716	0.733	28.455
	p.auroc	$3.56E-11$	$1.62E-20$	$1.74E-26$	$2.69E-20$	$1.41E-22$	
	aupr	0.128	0.119	0.194	0.176	0.185	
	p.aupr	$8.21E-28$	$2.31E-37$	$3.23E-45$	$3.86E-39$	$3.12E-41$	
TIGRESS	auroc	**0.769**	0.717	0.781	0.791	0.764	38.848
	p.auroc	$6.81E-21$	$1.92E-25$	$4.00E-35$	$5.68E-33$	$5.65E-28$	
	aupr	0.165	0.161	0.233	0.228	0.234	
	p.aupr	$6.47E-37$	$1.87E56$	$4.16E-55$	$5.08E-52$	$5.51E-53$	
GENIE3	auroc	0.745	0.733	0.775	0.791	0.798	37.482
	p.auroc	$3.3E-18$	$1.1E-28$	$9.7E-34$	$6.7E-33$	$1.9E-34$	
	aupr	0.154	0.155	0.231	0.208	0.197	
	p.aupr	$3.3E-34$	$7.9E-54$	$1.8E-54$	$5.5E-47$	$4.6E-44$	
NIMEFI	auroc	0.76	0.73	0.78	0.81	0.80	40.762
	p.auroc	$4.0E-20$	$3.3E-28$	$1.0E-35$	$9.1E-37$	$2.2E-34$	
	aupr	0.16	0.16	0.25	0.23	0.24	
	p.aupr	$1.2E-37$	$3.7E-57$	$6.0E-60$	$1.9E-51$	$1.8E-53$	
ENNET	auroc	0.731	0.807	0.813	0.822	0.829	52.839
	p.auroc	$1.13E-16$	$1.29E-46$	$1.02E-42$	$5.86E-39$	$6.10E-41$	
	aupr	0.184	0.261	0.289	0.291	0.286	
	p.aupr	$1.49E-41$	$9.39E-106$	$2.62E-69$	$1.44E-67$	$1.38E-65$	
GENIMS	auroc	0.750	**0.826**	**0.832**	**0.842**	**0.860**	**58.212**
	p.auroc	$9.34E-19$	$5.30E-52$	$1.26E-47$	$3.93E-43$	$5.78E-48$	
	aupr	**0.184**	**0.296**	**0.303**	**0.300**	**0.311**	
	p.aupr	$1.67E-41$	$3.72E-124$	$8.78E-73$	$7.31E-70$	$1.35E-71$	

of the other two networks corresponding are real world network for *E. coli* and *S. cerevisiae*. The DREAM5 was the first challenge to infer gene regulatory network on a genomic scale. A comparison similar to the previous comparison is carried out and the results are shown in Table 3. As only the first two DREAM5 network results are provided by the original NIMEFI article[3], we only show these two results for NIMEFI method in this paper. The best results for all the methods are typed in bold. We can see that our proposed method performs best with the metric of AUROC. From the results we also find that all methods achieve better results for the *in-silico* network than for the real world network. One important reason for the poor performance of the inference methods for the real world networks is that the gold standard information may not be complete.

Robustness and generalizability evaluation

The robustness is an important indicator of an inference algorithm. To further illustrate the robustness of our proposed method, we respectively generate ten gene expression datasets for the five DREAM4 benchmark networks for evaluation. We perform the GENIE3, ENNET and our proposed method on these datasets, and the results are shown in Fig. 2 in the form of boxplots. In the figure we can see that our proposed method outperforms these two methods, and the robustness of our method also can be guaranteed.

To further evaluate the performance of our method, we use the t-test to evaluate the statistical significance of the hypothesis that our method outperforms the GENIE3 and ENNET methods. The p values are listed in Table 4, Table 5 and Table 6, in which we can see that our method performs significantly better than the GENIE3 method with all metrics and in all instances. Our method also performs significantly better than the ENNET method with the most evaluation metrics and in most instances.

The above evaluations are all based on the benchmark networks of the DREAM projects. To illustrate the generalizability of the proposed method, we extract ten networks from the real *E.coli* and *Yeast* gene regulatory networks, respectively. Using the GeneNetWeaver simulator[30], we generate a gene expression dataset for each extracted network. We use both the normal and the log-normal noises to model the deviations. The results are shown in Fig. 3, in which we can see that the proposed method gives better and more robust inference results than the other five methods.

Performance improvement of the multi-level strategy

In this paper, we apply a multi-level strategy to improve the performance of the gene regulatory network inference algorithm. There are three levels in our method, which are the original guided regularized random forest result, the normalized result and the final refinement result. We record the results return by all the levels and evaluate the dynamic change of the performance. To ensure high reliability, we respectively generate ten gene expression datasets based on the five DREAM4 benchmark networks, and each level of our method is performed on these datasets. The results are shown in Fig. 4.

In the figure, we can see that the multi-level strategy improves the performance of inference significantly, especially the value of AUPR and the prediction score. For the value of AUROC, the normalization step improves over the original regularized random forest results by 1.4% on average and the refinement step improves over the normalized results by 6.5% on average. For the AUPR value, the normalization step improves over the original results by 13.5% on average and the refinement step improves over the normalized results by 27.4% on average. For the prediction score, the normalization step improves over the original results by 11.4% on average and the refinement step improves over the normalized results by 40.4% on average. The above results indicate that the normalization and refinement plays a large role in achieving the performance of our proposed gene network inference method.

Conclusion

In this paper, we proposed the GENIMS method to solve the problem of gene regulatory network. The GENIMS method adopts a three-level strategy and obtains better performance. The first level is to solve the individual regression problem with the guided regularized random forest algorithm, the second level is to apply q-norm normalization to reduce of the bias among different regression results and the third level is to refine the previous results according to the sparsity property of large scale gene regulatory networks. We first discussed the influence of the selection of different parameters in the GENIMS method. The performance is not sensitive to the selection of the global penalization parameter γ, but we can obtain better performance when the normalization parameter q is selected in the range of $[2,4]$. Then, we evaluated our method with the benchmark networks provided by the DREAM4 and DREAM5 projects. The evaluation results indicate that the GENIMS method can obtain more accurate and robust performance than the state-of-the-art methods. Additionally, we also evaluated the robustness and generalizability of the proposed methods.

Acknowledgements

This work was supported by the State Key Development Program for Basic Research of China (2013CB967402), Longhua Medical Project of the State Clinical Research Center of TCM at the Longhua Hospital (LYTD-21) and National Natural Science Foundation of China under grant No. 61221003, 91019004 and 91229123. The authors would like to thank the reviewers in advance for their comments.

References

1 C. R. Harwood and I. Moszer, *Comparative and functional genomics*, 2002, **3**, 37–41.
2 T. S. Gardner and J. J. Faith, *Physics of life reviews*, 2005, **2**, 65–88.
3 J. Ruyssinck, V. A. Huynh-Thu, P. Geurts, T. Dhaene, P. Demeester and Y. Saeys, *PloS one*, 2014, **9**, e92709.
4 C. Siegenthaler and R. Gunawan, *PloS one*, 2014, **9**, e90481.
5 G. Stolovitzky, D. Monroe and A. Califano, *Annals of the New*

Table 3 Results of different methods on DREAM5 networks

Method	Metric	Network 1		Network 3		Network 4	
ARACNE	auroc / aupr	0.545	0.099	0.512	0.029	0.5	0.017
	p-value	0.999	0.999	0.999	0.999	0.999	0.997
CLR	auroc / aupr	0.666	0.217	0.538	0.05	0.505	0.019
	p-value	0.998	1.25E-29	0.999	0.116	0.999	0.979
MRNET	auroc / aupr	0.668	0.194	0.525	0.041	0.501	0.018
	p-value	0.995	1.02E-15	0.999	0.881	0.999	0.999
TIGRESS	auroc / aupr	0.789	0.32	0.589	0.066	0.514	0.02
	p-value	9.96E-68	3.61E-146	0.0593	4.84E-06	0.24	0.516
GENIE3	auroc / aupr	0.814	0.291	0.619	0.094	0.517	**0.021**
	p-value	7.41E-105	1.55E-104	5.17E-12	1.17E-20	0.021	0.077
NIMEFI	auroc / aupr	0.82	0.32	0.63	**0.11**	*	*
	p-value	3.1E-118	8.7E-151	2.1E-19	4.4E-34	*	*
ENNET	auroc / aupr	0.867	**0.432**	0.642	0.069	0.532	**0.021**
	p-value	2.78E-213	0	1.09E-26	3.28E-07	5.78E-16	0.077
GENIMS	auroc / aupr	**0.897**	0.426	**0.705**	0.052	**0.533**	0.02
	p-value	3.94E-293	0	4.31E-96	0.05	2.4E-17	0.516

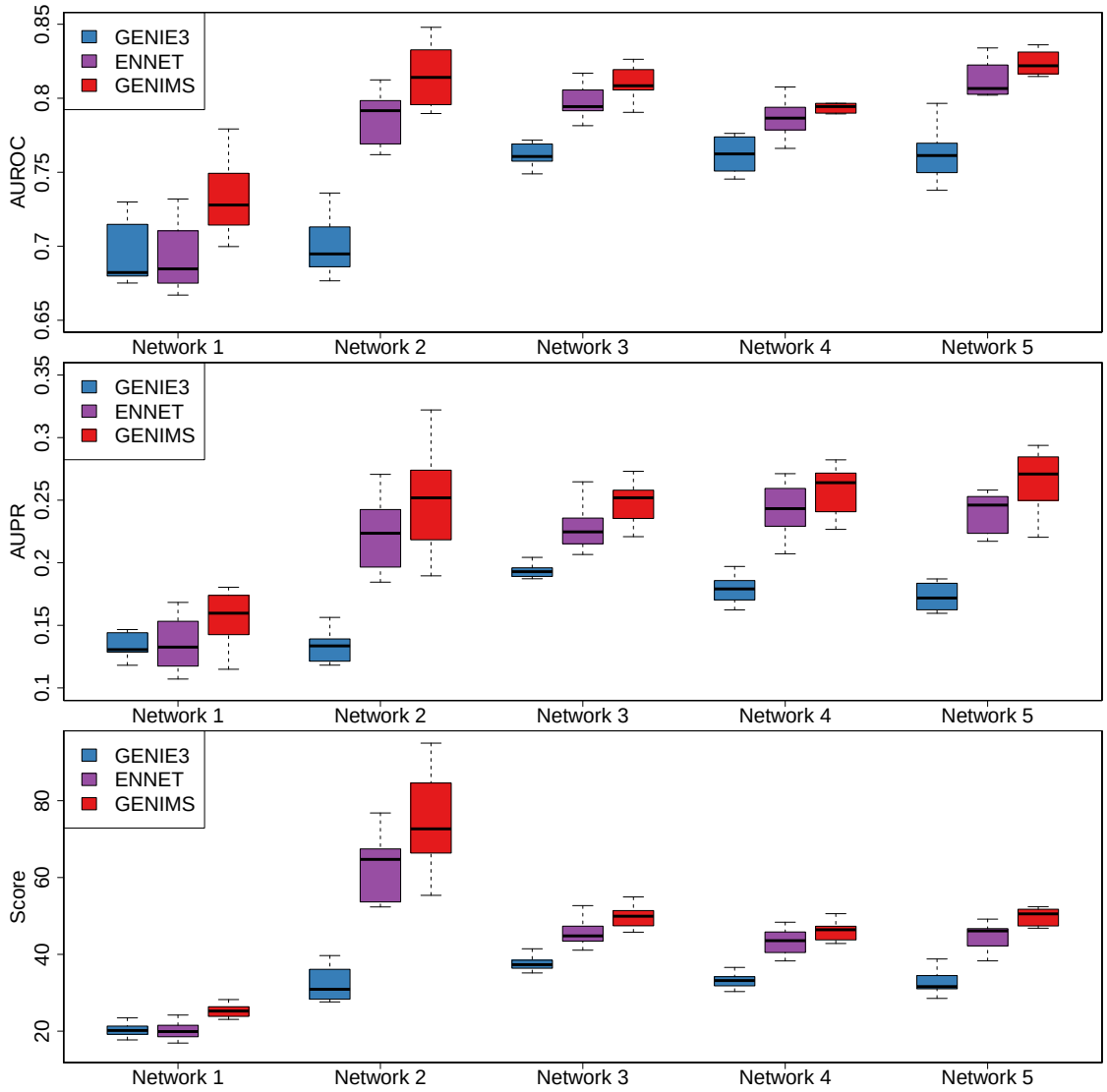


Fig. 2 The robustness evaluation of the GENIE3, ENNET and GENIMS method. The top plot is the comparison of the AUROC values. The middle plot is the comparison of the AUPR values and the bottom is the prediction score comparison.

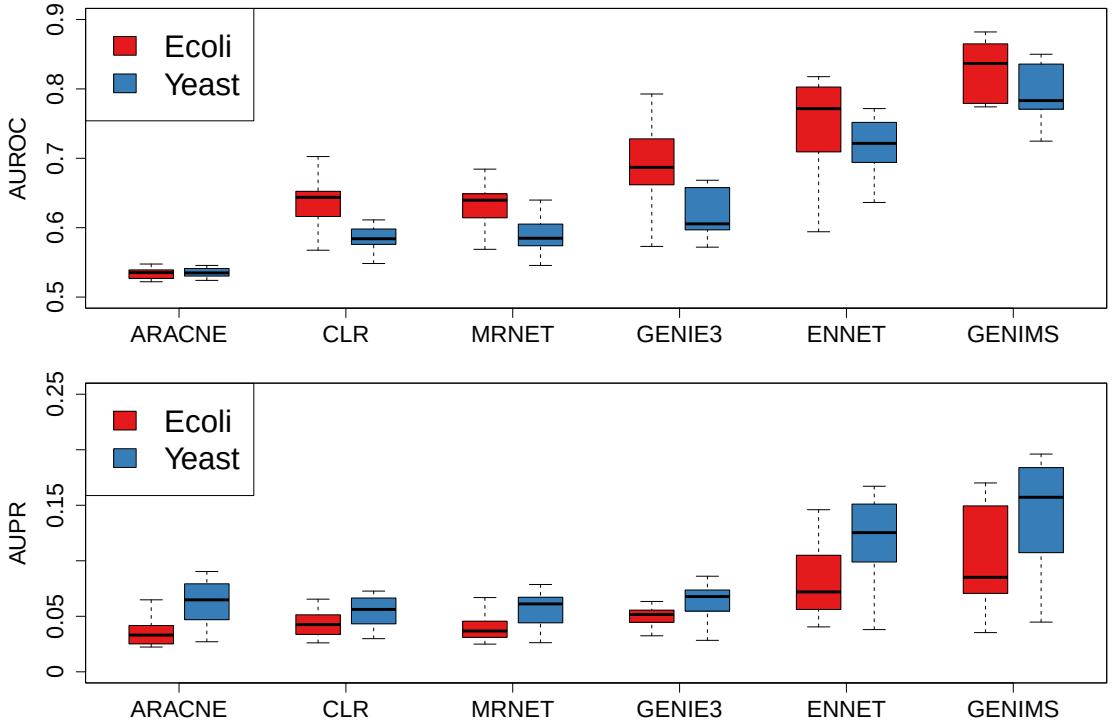


Fig. 3 The generalizability evaluation of the gene regulatory network inference methods. The top plot is the AUROC values of the 6 gene network inference methods and the bottom is the AUPR values of the 6 gene network inference methods.

Table 4 Significance testing of the AUROC values

	Network 1	Network 2	Network 3	Network 4	Network 5
GENIE3	1.14E-03	1.23E-10	5.84E-09	2.01E-06	1.50E-05
ENNET	7.92E-04	2.70E-03	2.36E-02	9.45E-02	1.24E-01

Table 5 Significance testing of the AUPR values

	Network 1	Network 2	Network 3	Network 4	Network 5
GENIE3	6.87E-03	5.52E-06	1.57E-07	1.17E-08	4.79E-08
ENNET	3.68E-02	1.10E-01	1.07E-02	6.40E-02	1.01E-02

Table 6 Significance testing of the prediction scores

	Network 1	Network 2	Network 3	Network 4	Network 5
GENIE3	5.75E-04	7.41E-07	7.82E-09	4.56E-10	1.83E-09
ENNET	2.94E-03	3.73E-02	7.46E-03	3.61E-02	1.23E-02

York Academy of Sciences, 2007, **1115**, 1–22.

6 D. Marbach, T. Schaffter, C. Mattiussi and D. Floreano, *Journal of computational biology*, 2009, **16**, 229–239.

7 G. Stolovitzky, R. J. Prill and A. Califano, *Annals of the New York Academy of Sciences*, 2009, **1158**, 159–195.

8 D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano and G. Stolovitzky, *Proceedings of the National Academy of Sciences*, 2010, **107**, 6286–6291.

9 D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, M. Kellis, J. J. Collins, G. Stolovitzky *et al.*, *Nature methods*, 2012, **9**, 796–804.

10 J. Sławek and T. Arodź, *BMC systems biology*, 2013, **7**, 106.

11 F. Yavari, F. Towhidkhah and S. Gharibzadeh, Biomedical Engineering Conference, 2008. CIBEC 2008. Cairo International, 2008, pp. 1–4.

12 B. Yang, J. Zhang, J. Shang and A. Li, Signal Processing, Communications and Computing (ICSPCC), 2011 IEEE International Conference on, 2011, pp. 1–4.

13 A. K. Tan and M. S. Mohamad, *Jurnal Teknologi*, 2012, **58**, 1–6.

14 N. Krämer, J. Schäfer and A.-L. Boulesteix, *BMC bioinformatics*, 2009, **10**, 384.

15 P. Menéndez, Y. A. Kourmpetis, C. J. ter Braak and F. A. van Eeuwijk, *PloS one*, 2010, **5**, e14147.

16 T. Chen, H. L. He, G. M. Church *et al.*, Pacific symposium on biocomputing, 1999, p. 4.

17 A. Polynikis, S. Hogan and M. di Bernardo, *Journal of theoretical biology*, 2009, **261**, 511–530.

18 J. Cao, X. Qi and H. Zhao, *Next Generation Microarray Bioinformatics*, Springer, 2012, pp. 185–197.

19 W. Zhao, E. Serpedin and E. R. Dougherty, *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 2008, **5**, 262–274.

20 A. Noor, E. Serpedin, M. Nounou, H. Nounou, N. Mohamed and L. Chouchane, Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2012 IEEE Symposium on, 2012, pp. 418–423.

21 A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. D. Favera and A. Califano, *BMC bioinformatics*, 2006, **7**, S7.

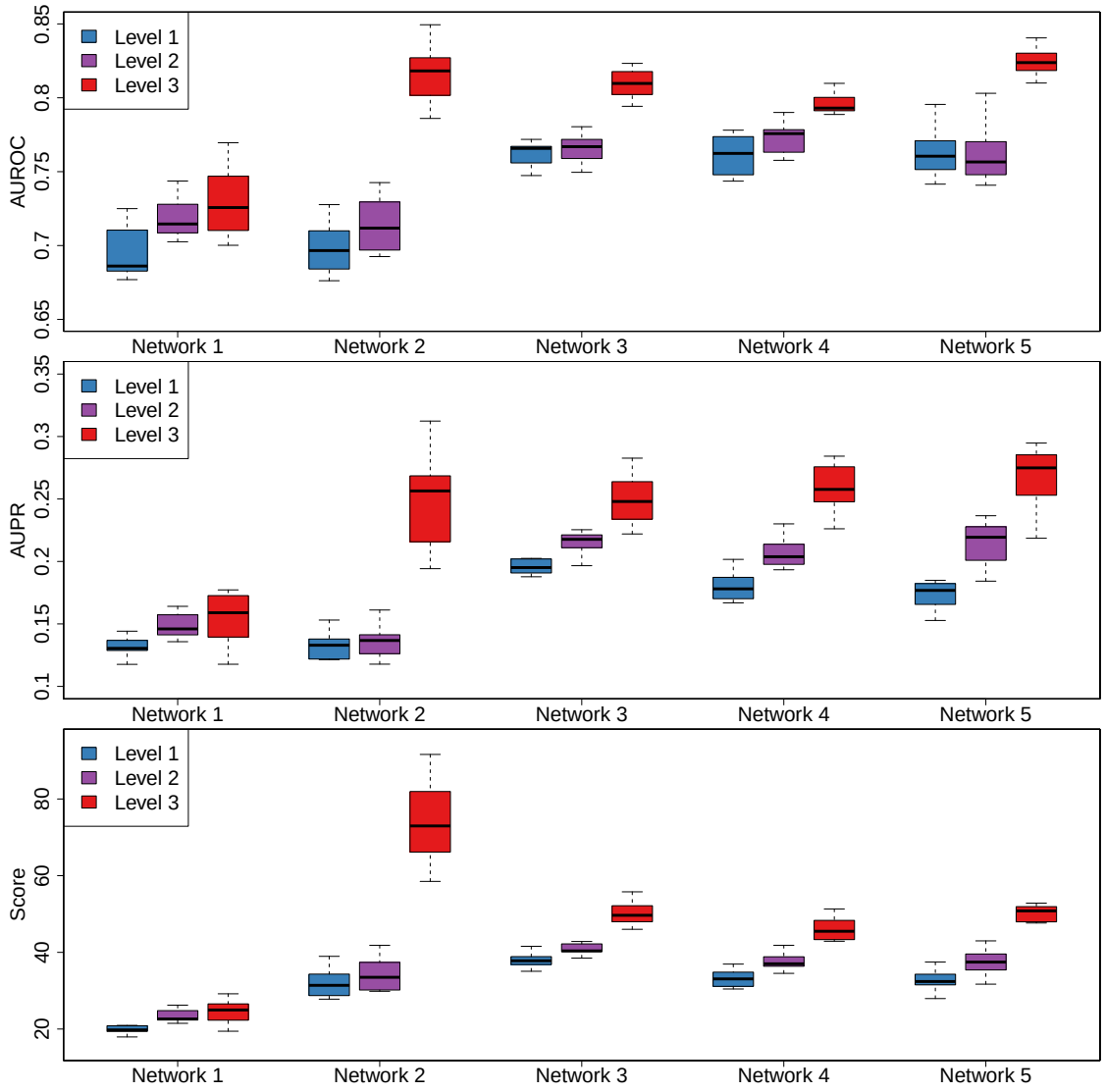


Fig. 4 Evaluation of the performance improvement with the multi-level strategy. The top plot is the comparison of AUROC values. The middle plot is the comparison of AUPR values and the bottom is the prediction score comparison.

22 P. E. Meyer, F. Lafitte and G. Bontempi, *BMC bioinformatics*, 2008, **9**, 461.
23 G. Sanguinetti, M. Rattray and N. D. Lawrence, *Bioinformatics*, 2006, **22**, 1753–1759.
24 P. E. Meyer, K. Kontos, F. Lafitte and G. Bontempi, *EURASIP journal on bioinformatics and systems biology*, 2007, **2007**, 8–8.
25 A.-C. Haury, F. Mordelet, P. Vera-Licona and J.-P. Vert, *BMC systems biology*, 2012, **6**, 145.
26 V. A. Huynh-Thu, A. Irrthum, L. Wehenkel and P. Geurts, *PloS one*, 2010, **5**, e12776.
27 J. Xiong and T. Zhou, *PloS one*, 2012, **7**, e43819.
28 N. Noman, L. Palafox and H. Iba, *Natural Computing and Beyond*, Springer, 2013, pp. 93–103.
29 H. Deng and G. Runger, *Pattern Recognition*, 2013, **46**, 3483–3489.
30 T. Schaffter, D. Marbach and D. Floreano, *Bioinformatics*, 2011, **27**, 2263–2270.