TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 08/11 pp446–455
DOI: 10.26599/TST.2018.9010097
Volume 24, Number 4, August 2019

A Novel Method of Gene Regulatory Network Structure Inference from Gene Knock-Out Expression Data

Xiang Chen, Min Li*, Ruiqing Zheng, Siyu Zhao, Fang-Xiang Wu, Yaohang Li, and Jianxin Wang

Abstract: Inferring Gene Regulatory Networks (GRNs) structure from gene expression data has been a challenging problem in systems biology. It is critical to identify complicated regulatory relationships among genes for understanding regulatory mechanisms in cells. Various methods based on information theory have been developed to infer GRNs. However, these methods introduce many redundant regulatory relationships in the network inference process due to external noise in the original data, topology sparseness in the network structure, and non-linear dependency among genes. Especially as the network size increases, the performance of these methods decreases dramatically. In this paper, a novel network structure inference method named Loc-PCA-CMI is proposed that first identifies local overlapped gene clusters, and then infers the local network structure for each cluster by a Path Consistency Algorithm based on Conditional Mutual Information (PCA-CMI). The final structure of the GRN is denoted as dependence among genes by an ensemble of the obtained local network structures. Loc-PCA-CMI was evaluated on DREAM3 knock-out datasets, and its performance was compared to other information theory-based network inference methods including ARACNE, MRNET, PCA-CMI, and PCA-PMI. Experimental results demonstrate our novel method Loc-PCA-CMI outperforms the other four methods in DREAM3 datasets especially in size 50 and 100 networks.

Key words: gene regulatory networks; network inference; path consistency algorithm

1 Introduction

Inferring and understanding Gene Regulatory Networks (GRNs) is a critical problem in systems biology, which can help biomedical scientists to explicitly identify complicated regulatory relationships among genes and understand regulatory mechanisms in cells[1,2]. In the past, GRNs were inferred from experimental interventions in which regulatory interactions among genes were verified. Obviously, this approach is infeasible[3] and requires substantial time and considerable cost. Owing to the development of microarray technologies, tremendous amounts of gene expression data have been generated[4], which makes it feasible for GRNs to be inferred from these expression data based on computational methods[5]. In recent years, the inference of networks based on computational methods has become one of the most crucial goals[1,6]. Various methods have been proposed for GRNs inference, such as regression-based methods[7–12], ordinary differential equation-based methods[13–15], Bayesian and dynamic Bayesian networks[16–21], and state-space based methods[22,23]. Unfortunately, gene expression data are typically of high dimensions and relatively small sample size that

• Xiang Chen, Min Li, Ruiqing Zheng, Siyu Zhao, and Jianxin Wang are with the School of Computer Science and Engineering, Central South University, Changsha 410083, China. E-mail: limin@mail.csu.edu.cn.

• Fang-Xiang Wu is with the Department of Mechanical Engineering and Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK S7N 5A9, Canada.

• Yaohang Li is with the Department of Computer Science, Old Dominion University, Norfolk, VA 23529, USA.

∗ To whom correspondence should be addressed.

Manuscript received: 2018-04-05; accepted: 2018-05-01

suffer from “dimensionality curse”[24]. Furthermore, gene expression data usually involve large amounts of external noise and non-linear relationships. All of these issues make it more complex and challenging to accurately infer regulatory interactions among genes, especially when dealing with large-scale gene expression data in the post-genome era.

GRNs can be viewed as undirected acyclic graphs if both up-streaming or down-streaming regulatory relationships among genes are not taken into account and the self-regulatory mechanism is ignored[25], in which each node corresponds to a gene and each edge represents a regulatory relationship between genes. Various computational methods to construct accurate structures of GRNs from expression data have been proposed based on a variety of different assumptions and different conditions[26,27]. Current approaches can be broadly divided into model-based and model-free approaches. Model-based methods usually formulate a computational model of the system and further learn the parameters of such a model. Typical computational models include Boolean network[28–31], Bayesian network[17,32–35], and differential equation models[15,36–40]. The Boolean network model is the simplest network model, which is implemented through Boolean variables and Boolean logic. Because the state of gene expression is considered to be only active or inactive, Boolean network models cannot entirely capture complex system behavior[41]. The Bayesian network model is a popular probabilistic graphical model in which the dependency relationships among genes are described via a directed acyclic graph. The Bayesian network model outperforms other models in dealing with noise and incorporating prior knowledge, but structure learning in the model is computationally intensive and has been proven as an NP-hard problem[42]. The differential equation model characterizes the expression level of a gene at a certain time by a function, which involves regulatory interactions with other genes. Differential equation models quantify the change rate (derivative) of the expression of one gene in the system as a function of expression levels of other related genes. A major challenge to using differential equation models for reconstructing GRNs is how to identify the model structure and estimate parameters efficiently in high-dimensional models. Excellent reviews on diverse data-driven modeling schemes and related topics can be found in Refs. [43–47].

Other than model-based methods, model-free approaches identify regulatory interactions mainly by measuring dependences among genes. Typical algorithms include correlation-based and information theory-based methods. In the correlation-based method, a regulatory interaction is determined by the degree of co-expression between two genes such as Pearson correlation, rank correlation, Euclidean distance, and the angle between a pair of observed expression vectors[48]. However, the correlation-based methods cannot identify complex dependencies between genes, such as non-linear dependencies[49]. Furthermore, quite a few functionally related genes might not be co-expressed, which makes it difficult to accurately infer regulatory interactions. The information theory-based method is also a representative model-free method, in which Mutual Information (MI) is favored to measure potential dependency among genes as it can capture non-linear dependencies effectively[50,51]. In recent years, various network inference methods based on information theory have been proposed, which focus on distinguishing direct regulatory interactions from indirect associations[52]. To eliminate indirect interactions, Margolin et al.[53] proposed the ARACNE method based on Data Processing Inequality with interaction triangles considered. The Minimum-Redundancy NETwork (MRNET) by Meyer et al.[54] uses a minimum redundancy feature selection method[55], wherein for each candidate gene in a network, it selects a subset of its highly relevant genes while minimizing the MI-based criteria between the selected genes. Zhang et al.[51] introduced a Path Consistency Algorithm based on Conditional MI (PCA-CMI); Zhao et al.[56] introduced a PCA based on Part MI (PCA-PMI). The PCA is an exhaustive algorithm that is widely used in inferring GRNs[51]. A trade-off is usually made between running time and accuracy in both PCA-CMI and PCA-PMI. As the network size increases, more uncontrollable external noise within the instinctive complex network structure makes prediction accuracy of GRNs decrease dramatically. To improve this situation, motivated by the divide-and-conquer strategy, we first used top-ranked, highly co-expressed genes as centroids of local clusters; each cluster’s accurate structure was then refined with PCA-CMI. The final structure of the GRN was then inferred with an ensemble of all the local network structures together.

We have named this novel approach as Loc-PCA-CMI. Hereafter and intuitively, the Loc-PCA-CMI method can deal with relatively larger datasets and benefit from the relatively accurate structure inference for small size gene subnetworks with PCA-CMI.

2 Methods

In this section, we introduce related work in information theory including entropy, MI, and CMI, as well as the proposed algorithm, Loc-PCA-CMI, for inferring GRNs.

2.1 Related work

With the advantages of measuring non-linear dependence association between two variables and relatively high efficiency, information theory is increasingly used to measure the regulatory strength of genes. The MI is defined as follows:

$$\mathrm{MI}(X,Y)=\iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y \qquad (1)$$

where $p(x,y)$ denotes the joint probability density function of two variables X and Y. X is a vector in gene expression data, in which the elements denote the corresponding gene's expression values in different conditions (samples). $p(x)$ (resp. $p(y)$) represents the marginal probability density distribution of X (resp. Y).

The CMI can be expressed in terms of entropies as

$$\mathrm{CMI}(X,Y|Z)=H(X,Z)+H(Y,Z)-H(Z)-H(X,Y,Z) \qquad (2)$$

where $H(X,Z)$, $H(Y,Z)$, and $H(X,Y,Z)$ are joint entropies. High CMI indicates that there may be a close relationship between the variables X and Y given variable(s) Z.

The entropy is estimated with a Gaussian kernel probability density estimator[2] and we can get the entropy of variable X as follows, where $|C|$ is the determinant of covariance matrix of variable X[51]:

$$H(X)=\log(2\pi\mathrm{e})^{\frac{n}{2}}|C|^{-\frac{1}{2}} \qquad (3)$$

Furthermore, we can obtain the following equation:

$$\mathrm{MI}(X,Y)=\frac{1}{2}\log\frac{|C(X)|\cdot|C(Y)|}{|C(X,Y)|} \qquad (4)$$

2.2 Loc-PCA-CMI

It is well-known that biological systems are seldom fully connected and most nodes are only directly connected to a small number of other nodes[57], rendering the GRN as a sparse network. A key step in identifying the sparse structure of the network is to identify the significant edges that may have a comparatively high co-expression value. Specifically our proposed method Loc-PCA-CMI first selects the top *n* highly co-expressed edges by Pearson correlation analysis with a False Discovery Rate (FDR) correction in the *p*-value, and, secondly, in the reduced-edges space, computes local overlapped clusters with genes connected by edges. Then for each local cluster, we apply the PCA-CMI algorithm, which can construct a high-confidence undirected network[58] by removing the most likely uncorrelated edges repeatedly from low to high-order dependence correlation until no edges can be removed, to obtain each local network structure. Final edge weight of the complete regulatory network is obtained by averaging edge weight with each inferred network structure. The entire framework is provided in Fig. 1 and the implementation details are shown below in Algorithm 1. As PCA-CMI is extremely competent for relatively small GRN structure inference, we ran a

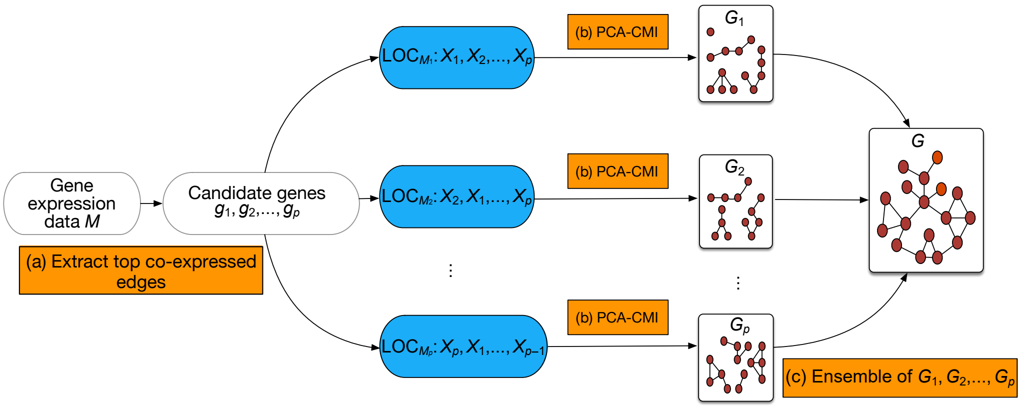

Gene expression data M
(a) Extract top co-expressed edges
Candidate genes $g_1, g_2,..., g_p$
LOC_{M_1}: $X_1, X_2,..., X_p$
LOC_{M_2}: $X_2, X_1,..., X_p$
LOC_{M_p}: $X_p, X_1,..., X_{p-1}$
(b) PCA-CMI
(b) PCA-CMI
(b) PCA-CMI
G_1
G_2
G_p
G
(c) Ensemble of $G_1, G_2,..., G_p$

Fig. 1 Loc-PCA-CMI framework. (a) Top *n* co-expressed edges are extracted from gene expression data matrix *M*, and obtain the candidate genes $g_1, g_2, \ldots, g_p$. The candidate genes are then grouped into different local clusters $\mathrm{LOC}_{M_1}, \mathrm{LOC}_{M_2}, \ldots, \mathrm{LOC}_{M_p}$ with each gene $g_1, g_2, \ldots, g_p$ as the centroid. (b) For each local overlapped cluster, PCA-CMI is applied to get its accurate structure. (c) Ensemble of diverse cluster structure $G_1, G_2, \ldots, G_p$ obtains the final structure of the GRN as *G*.

Algorithm 1 Loc-PCA-CMI

```
Input: $M$ (the gene expression data matrix), $m$ (the number of genes), $n$ (the number of top ranked edges), $c$ (constant number); $k$ (CMI order number), and $\beta$ (order threshold) in subroutine PCA-CMI.
Output: Graph weight matrix $G$
if $m \leqslant c$ then
    $G \leftarrow$ PCA-CMI($M, k, \beta$);
    return $G$
else
    Construct pair-wise gene-gene Pearson correlation matrix $\Omega = \rho(M_i, M_j)$;
    Select top $n$ edges as $E$ with highest Pearson correlation value in $\Omega$ with FDR correction in $p$-value, and according to which to get $p$ candidate genes as $g_1, g_2, \ldots, g_p$;
    for each gene in $g_1, g_2, \ldots, g_p$ do
        Retrieve its directly connected genes that in edges list $E$ as local cluster LOC$_{M_i}$;
    for each cluster LOC$_{M_i}$ in LOC do
        $G_i \leftarrow$ PCA-CMI(LOC$_{M_i}, k, \beta$);
    $G \leftarrow$ mean($G_1, G_2, \ldots, G_p$);
    return $G$
```

preprocessing to check the number of input genes; if it was less than or equal to a constant c, then PCA-CMI was directly applied for the GRN structure inference.

The computational complexity of Algorithm 1 is generally determined by two factors: the first is the number of local overlapped clusters p, which is usually lower than the number of genes m; the second is the complexity of the PCA-CMI subroutine. All of the PCA-CMI subroutines can be executed in parallel to make the proposed algorithm Loc-PCA-CMI perform more efficiently.

3 Materials

We benchmarked the performance of our approach, Loc-PCA-CMI, using six simulation data from well-known DREAM3 challenges[59]. DREAM3 features in-silico networks and expression data simulated using GeneNetWeaver software. Benchmark networks were derived as subnetworks of a system of regulatory interactions from known model organisms: E.coli and S.cerevisiae. Six gene knock-out expression networks in DREAM3 were evaluated in our experiments, which included three different sizes, varying as 10, 50, and 100, with two types of model organisms, E.coli and S.cerevisiae. Table 1 gives detailed descriptions of the datasets.

Table 1 Descriptions of the datasets in our experiments.

Dataset	Number of samples	Average (Max) degree	Number of edges	Network density
DREAM3-10 Ecoli	11	2.2 (5)	11	0.244
DREAM3-50 Ecoli	51	2.48 (14)	62	0.051
DREAM3-100 Ecoli	101	2.5 (14)	125	0.025
DREAM3-10 Yeast	11	2 (4)	10	0.222
DREAM3-50 Yeast	51	3.08 (13)	77	0.063
DREAM3-100 Yeast	101	3.32 (10)	166	0.034

Rows in each input datafile stand for samples (observations) and columns stand for genes (variables). The first line is the wild-type expression data; every gene in this sample stays at a steady state. The n ($n > 1$) line stands for that how other gene expression data changes after the $n - 1$ gene is knock-out.

4 Results and Discussion

As described in Algorithm 1, three intrinsic parameters affected the performance of Loc-PCA-CMI in GRN structure inference. The first parameter was the number of top-selected edges, n. If n increased, more edges were considered and the local cluster size would increase subsequently. The second parameter was β, which acted as the threshold value of MI and CMI to decide independence. The third parameter was CMI order number k. Theoretically, by increasing k, the structure was more accurate if CMI did not reach the threshold β in $k - 1$ order. The latter two parameters were with PCA-CMI and PCA-PMI. The best value of n could be obtained by cross-validation and generally, the larger value of n could contribute to a larger size cluster and more genes were covered in the network; in our experiments, we set its value to be $n = \binom{m}{2}/5$ uniformly. Besides the above three intrinsic parameters, we set constant $c = 10$ in Algorithm 1, i.e., if the number of genes was less than or equal to 10, Loc-PCA-CMI called PCA-CMI directly, and in this case, performances of Loc-PCA-CMI and PCA-CMI were the same.

We assessed the performance of Loc-PCA-CMI by evaluating the Area Under the Receiver Operating Characteristic curve (AUROC) and the Area Under the Precision-Recall curve (AUPR). As in sparse biological networks, the number of non-existing edges (negatives) outweighed the number of existing edges (positives) significantly; AUPR was therefore, in fact, more informative to AUROC[60]. We tended to use AUPR for evaluation, but for a conservative comparison with other methods that adopt AUROC as an evaluation metric, we also took AUROC as a supplementary metric. Higher

AUROC and AUPR values indicated more accurate GRN predictions. For this purpose, we computed the numbers of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) edges by comparing the regulatory edges in the gold standard network with the top q edges from the ranked list output of Loc-PCA-CMI. The ROC curve was constructed by plotting the True Positive Rates (TPR $=$ TP$/$(TP $+$ FN)) versus the False Positive Rates (FPR $=$ FP$/$(FP$+$ TN)) for increasing $q\,(q = 1, 2, \ldots, m^2)$. Similarly, the precision (TP$/$(TP $+$ FP)) and recall (TP$/$(TP $+$ FN)) curve was plotted for increasing q.

It should be noted that in Algorithm 1, after each local cluster was obtained, both PCA-CMI and PCA-PMI were alternatives for the subsequent structure refinement. If PCA-CMI was replaced with PCA-PMI, a novel method was generated, which we named Loc-PCA-PMI, analogously. Four PCA based methods were then derived including PCA-PMI, PCA-CMI, Loc-PCA-PMI, and Loc-PCA-CMI, all of which, at present, belong to model-free methods. As shown in Table 1, among the six benchmark datasets, DREAM3-10 Ecoli and DREAM3-10 Yeast datasets contained only 10 genes, hence Loc-PCA-CMI and PCA-CMI were identical in performance, as was the case with Loc-PCA-PMI and PCA-PMI according to the principle of Algorithm 1. For a meaningful comparison of these PCA based methods, we selected four other datasets whose gene number was greater than 10. Order number was not explicitly discussed in Refs. [51, 56], wherein $\beta = 0.03$ and $k = 2$ were set directly. We were curious as to how order number k affects the performance of these methods as well. By varying the order number k from 1 to 10 in these four methods with fixed threshold $\beta = 0.03$, AUROC and AUPR could be calculated, respectively. Figure 2 illustrates the result, in summary, of the benchmark datasets, and from which we could draw the following two conclusions:

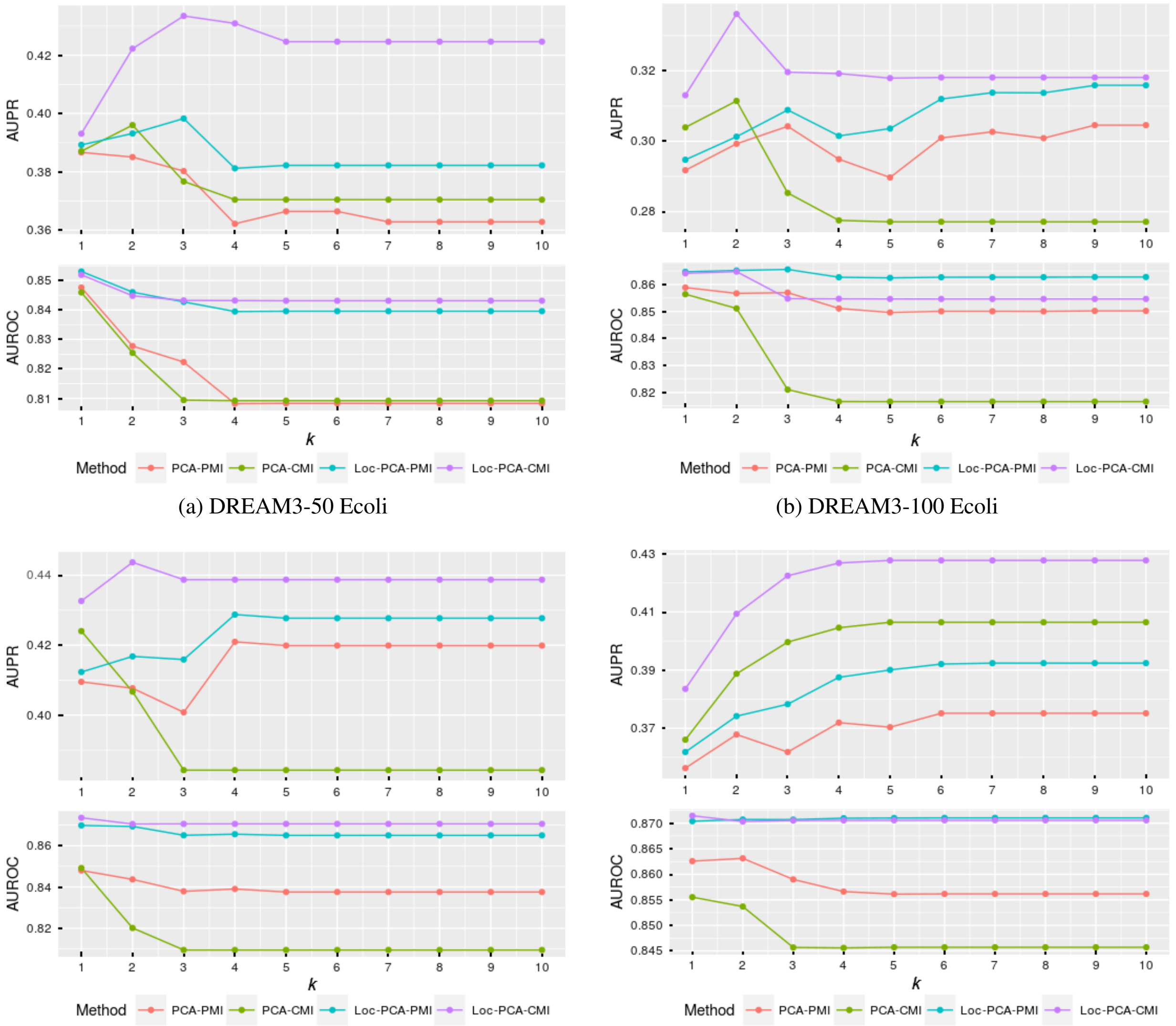
- Order number k affects the results of these four PCA based methods, generally when k reaches 4, AUPR and AUROC become stable, except those in the DREAM3-100 Ecoli dataset.
- Loc-PCA-CMI and Loc-PCA-PMI yield higher AUPR and AUROC than PCA-CMI and PCA-PMI, respectively, hence the local cluster strategy adopted in the algorithm helps to improve the performance of PCA-CMI and PCA-PMI.

We also conducted a comparison experiment using Loc-PCA-CMI with four previously proposed methods on all the six of our benchmark datasets, which included ARACNE, MRNET, PCA-PMI, and PCA-CMI. We used the R package “minet” with default parameters for evaluation of ARACNE and MRNET[61]. The MI matrices of the methods were approximated using Pearson correlation directly from continuous gene knock-out expression data[62,63]. For the implementation of PCA-PMI and PCA-CMI, we downloaded the MATLAB codes according to the URL provided in Refs. [51, 56]. We preferred the default value of parameters in PCA-PMI and PCA-CMI, where $\beta = 0.03$ and $k = 2$. For Loc-PCA-CMI, we also adopted the same values for these two parameters for comparison. Table 2 gives the AUROC and AUPR of this experiment. From the table, we can see that AUPR of all these contending methods decreased dramatically when the network size increased. Loc-PCA-CMI was only after PCA-PMI (or Loc-PCA-PMI in this case) in the DREAM3-10 Yeast dataset, while in the other five datasets, it outperformed the other four methods, ARACNE, MRNET, PCA-PMI, and PCA-CMI, in terms of both AUROC and AUPR. Additionally, for a more complete demonstration, we also include the performance of Loc-PCA-PMI in the table, with $\beta = 0.03$ and $k = 2$. Loc-PCA-CMI and Loc-PCA-PMI were almost identical in terms of AUROC. However, in most of the datasets, Loc-PCA-CMI had a superior AUPR than Loc-PCA-PMI. We provide the source codes including all the methods, benchmark datasets, and evaluation scripts at https://github.com/chenxofhit/Loc-PCA-CMI.git.

5 Conclusion

We have proposed a novel model-free GRN structure inference method named Loc-PCA-CMI, which was motivated by the divide-and-conquer strategy. At present, all the experiments are conducted in the DREAM3 challenge silico datasets. Experiments on DREAM3 knock-out datasets show that Loc-PCA-CMI benefits from the local overlapped cluster strategy. Furthermore, Loc-PCA-CMI outperforms other comparing methods including ARACNE, MRNET, PCA-PMI, and PCA-CMI, especially for networks of sizes of 50 and 100.

Loc-PCA-CMI was an extended version of PCA-CMI and its limitations in computational efficiency can be inherited, especially when dealing with large-size datasets. The number of local clusters in the case of a



(a) DREAM3-50 Ecoli

(b) DREAM3-100 Ecoli

(c) DREAM3-50 Yeast

(d) DREAM3-100 Yeast

Fig. 2 AUPR and AUROC by varying *k* from 1 to 10 of four PCA based methods on four different datasets: (a) DREAM3-50 Ecoli; (b) DREAM3-100 Ecoli; (c) DREAM3-50 Yeast; (d) DREAM3-100 Yeast.

Table 2 AUROC and AUPR for the six datasets using different methods.

Dataset	ARACNE		MRNET		PCA-PMI		Loc-PCA-PMI		PCA-CMI		Loc-PCA-CMI	
	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR
DREAM3-10 Ecoli	0.523	0.255	0.518	0.258	0.816	0.483	0.816	0.483	0.825	0.499	**0.825**	**0.499**
DREAM3-50 Ecoli	0.474	0.050	0.529	0.061	0.828	0.385	**0.846**	0.393	0.825	0.396	0.845	**0.422**
DREAM3-100 Ecoli	0.505	0.027	0.488	0.025	0.857	0.299	0.865	0.301	0.851	0.311	**0.865**	**0.336**
DREAM3-10 Yeast	0.628	0.321	0.644	0.322	0.995	0.933	**0.995**	**0.933**	0.993	0.918	0.993	0.918
DREAM3-50 Yeast	0.507	0.074	0.524	0.080	0.844	0.408	0.869	0.417	0.820	0.406	**0.871**	**0.444**
DREAM3-100 Yeast	0.547	0.040	0.556	0.042	0.863	0.368	**0.871**	0.374	0.854	0.389	0.870	**0.409**

large network can be extremely large. However, if we can control the size of each local cluster, our method should be applicable to large size datasets. One of our future works was to improve the cluster strategy, such as by integrating protein complexes[64, 65], to be more efficient and effective in dealing with large-size data. We mainly focus on inferring GRNs structure and have not considered the stability of networks in this study. As a result, our future studies will attempt to infer stable networks.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (Nos. 61622213 and 61732009), the 111 Project (No. B18059), and the Hunan Provincial Science and Technology Program (No. 2018WK4001).

References

[1] G. Altay and F. Emmert-Streib, Inferring the conservative causal core of gene regulatory networks, *BMC Systems Biology*, vol. 4, no. 1, p. 132, 2010.

[2] K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano, Reverse engineering of regulatory networks in human b cells, *Nature Genetics*, vol. 37, no. 4, p. 382, 2005.

[3] L. Elnitski, V. X. Jin, P. J. Farnham, and S. J. Jones, Locating mammalian transcription factor binding sites: A survey of computational and experimental techniques, *Genome Research*, vol. 16, no. 12, pp. 1455–1464, 2006.

[4] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, et al., Functional discovery via a compendium of expression profiles, *Cell*, vol. 102, no. 1, pp. 109–126, 2000.

[5] S. R. Maetschke, P. B. Madhamshettiwar, M. J. Davis, and M. A. Ragan, Supervised, semi-supervised and unsupervised inference of gene regulatory networks, *Briefings in Bioinformatics*, vol. 15, no. 2, pp. 195–211, 2013.

[6] A. A. Margolin, K. Wang, W. K. Lim, M. Kustagi, I. Nemenman, and A. Califano, Reverse engineering cellular networks, *Nature Protocols*, vol. 1, no. 2, p. 662, 2006.

[7] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, Inferring regulatory networks from expression data using tree-based methods, *PLoS One*, vol. 5, no. 9, pp. 1–10, 2010.

[8] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert, TIGRESS: Trustful Inference of Gene REgulation using Stability Selection, *BMC Syst. Biol.*, vol. 6, no. 1, p. 145, 2012.

[9] V. A. Huynh-Thu, G. Sanguinetti, A. Huynh-thu, and T. Jump, Combining tree-based and dynamical systems for the inference of gene regulatory networks, *Bioinformatics*, vol. 31, no. 10, pp. 1614–1622, 2014.

[10] L.-Z. Liu, F.-X. Wu, and W.-J. Zhang, A group lasso-based method for robustly inferring gene regulatory networks from multiple time-course datasets, *BMC Systems Biology*, vol. 8, no. S3, p. S1, 2014.

[11] M. Li, R. Zheng, Y. Li, F.-X. Wu, and J. Wang, Mgt-sm: A method for constructing cellular signal transduction networks, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, doi: 10.1109/TCBB.2017.2705143.

[12] R. Zheng, M. Li, X. Chen, F.-X. Wu, Y. Pan, and J. Wang, Bixgboost: A scalable, flexible boosting-based method for reconstructing gene regulatory networks, *Bioinformatics*, doi: 10.1093/bioinformatics/bty908.

[13] E. Sakamoto and H. Iba, Inferring a system of differential equations for a gene regulatory network by using genetic programming, in *Proceedings of the 2001 Congress on Evolutionary Computation*, 2001, vol. 1, pp. 720–726.

[14] A. R. Chowdhury, M. Chetty, and R. Evans, Stochastic s-system modeling of gene regulatory network, *Cognitive Neurodynamics*, vol. 9, no. 5, pp. 535–547, 2015.

[15] Z. Li, P. Li, A. Krishnan, and J. Liu, Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic bayesian network analysis, *Bioinformatics*, vol. 27, no. 19, pp. 2686–2691, 2011.

[16] K. Murphy and S. Mian, Modelling gene expression data using dynamic Bayesian networks, Technical report, Computer Science Division, University of California, Berkeley, CA, USA, 1999.

[17] M. Zou and S. D. Conzen, A new Dynamic Bayesian Network (DBN) approach for identifying gene regulatory networks from time course microarray data, *Bioinformatics*, vol. 21, no. 1, pp. 71–79, 2004.

[18] N. X. Vinh, M. Chetty, R. Coppel, and P. P. Wangikar, Globalmit: Learning globally optimal dynamic Bayesian network with the mutual information test criterion, *Bioinformatics*, vol. 27, no. 19, pp. 2765–2766, 2011.

[19] W. C. Young, A. E. Raftery, and K. Y. Yeung, Fast Bayesian inference for gene regulatory networks using scanbma, *BMC Systems Biology*, vol. 8, no. 1, p. 47, 2014.

[20] F. Liu, S.-W. Zhang, W.-F. Guo, Z.-G. Wei, and L. Chen, Inference of gene regulatory network based on local Bayesian networks, *PLOS Comput. Biol.*, vol. 12, no. 8, p. e1005024, 2016.

[21] N. Omranian, J. M. Eloundou-Mbebi, B. Mueller-Roeber, and Z. Nikoloski, Gene regulatory network inference using fused lasso on multiple data sets, *Scientific Reports*, vol. 6, p. 20533, 2016.

[22] F.-X. Wu, W.-J. Zhang, and A. J. Kusalik, Modeling gene expression from microarray expression data with state-space equations, in *Biocomputing 2004*. World Scientific, 2003, pp. 581–592.

[23] M. Quach, N. Brunel, and F. d'Alché Buc, Estimating parameters and hidden variables in non-linear state-space models based on odes for biological networks inference, *Bioinformatics*, vol. 23, no. 23, pp. 3209–3216, 2007.

[24] Y. Wang, T. Joshi, X.-S. Zhang, D. Xu, and L. Chen, Inferring gene regulatory networks from multiple microarray datasets, *Bioinformatics*, vol. 22, no. 19, pp. 2413–2420, 2006.

[25] V. A. Huynh-Thu, A. Irrthum, L. Wehenkel, and P. Geurts, Inferring regulatory networks from expression data using tree-based methods, *PloS One*, vol. 5, no. 9, p. e12776, 2010.

[26] W. J. Longabaugh, E. H. Davidson, and H. Bolouri, Computational representation of developmental genetic regulatory networks, *Developmental Biology*, vol. 283, no. 1, pp. 1–16, 2005.

[27] G. Karlebach and R. Shamir, Modelling and analysis of

gene regulatory networks, *Nature Reviews — Molecular Cell Biology*, vol. 9, no. 10, p. 770, 2008.

[28] I. Shmulevich, E. R. Dougherty, S. Kim, and W. Zhang, Probabilistic boolean networks: A rule-based uncertainty model for gene regulatory networks, *Bioinformatics*, vol. 18, no. 2, pp. 261–274, 2002.

[29] H. Kim, J. K. Lee, and T. Park, Boolean networks using the chi-square test for inferring large-scale gene regulatory networks, *BMC Bioinformatics*, vol. 8, no. 1, p. 37, 2007.

[30] S. Bornholdt, Boolean network models of cellular regulation: Prospects and limitations, *Journal of the Royal Society Interface*, vol. 5, no. Suppl 1, pp. S85–S94, 2008.

[31] J. X. Zhou, A. Samal, A. F. d'Hérouël, N. D. Price, and S. Huang, Relative stability of network states in boolean network models of gene regulation in development, *Biosystems*, vol. 142, pp. 15–24, 2016.

[32] S. Y. Kim, S. Imoto, and S. Miyano, Inferring gene networks from time series microarray data using dynamic bayesian networks, *Briefings in Bioinformatics*, vol. 4, no. 3, pp. 228–235, 2003.

[33] X.-W. Chen, G. Anantha, and X. Wang, An effective structure learning method for constructing gene networks, *Bioinformatics*, vol. 22, no. 11, pp. 1367–1374, 2006.

[34] C. J. Needham, J. R. Bradford, A. J. Bulpitt, and D. R. Westhead, A primer on learning in Bayesian networks for computational biology, *PLoS Computational Biology*, vol. 3, no. 8, p. e129, 2007.

[35] L.-Y. Lo, M.-L. Wong, K.-H. Lee, and K.-S. Leung, High-order dynamic Bayesian network learning with hidden common causes for causal gene regulatory network, *BMC Bioinformatics*, vol. 16, no. 1, p. 395, 2015.

[36] T. S. Gardner, D. Di Bernardo, D. Lorenz, and J. J. Collins, Inferring genetic networks and identifying compound mode of action via expression profiling, *Science*, vol. 301, no. 5629, pp. 102–105, 2003.

[37] D. di Bernardo, M. J. Thompson, T. S. Gardner, S. E. Chobot, E. L. Eastwood, A. P. Wojtovich, S. J. Elliott, S. E. Schaus, and J. J. Collins, Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks, *Nature Biotechnology*, vol. 23, no. 3, pp. 377–383, 2005.

[38] M. Bansal, G. D. Gatta, and D. Di Bernardo, Inference of gene regulatory networks and compound mode of action from time course gene expression profiles, *Bioinformatics*, vol. 22, no. 7, pp. 815–822, 2006.

[39] A. Honkela, C. Girardot, E. H. Gustafson, Y.-H. Liu, E. E. Furlong, N. D. Lawrence, and M. Rattray, Model-based method for transcription factor target identification with limited data, *Proceedings of the National Academy of Sciences*, vol. 107, no. 17, pp. 7793–7798, 2010.

[40] T. Lu, H. Liang, H. Li, and H. Wu, High-dimensional odes coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification, *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1242–1258, 2011.

[41] W.-P. Lee and W.-S. Tzou, Computational methods for discovering gene networks from expression data, *Briefings in Bioinformatics*, vol. 10, no. 4, pp. 408–423, 2009.

[42] D. M. Chickering, D. Heckerman, and C. Meek, Large-sample learning of Bayesian networks is np-hard, *Journal of Machine Learning Research*, vol. 5, pp. 1287–1330, 2004.

[43] M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke, Gene regulatory network inference: Data integration in dynamic models—A review, *Biosystems*, vol. 96, no. 1, pp. 86–103, 2009.

[44] 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., Wisdom of crowds for robust gene network inference, *Nature Methods*, vol. 9, no. 8, pp. 796–804, 2012.

[45] F.-X. Wu, Inference of gene regulatory networks and its validation, *Current Bioinformatics*, vol. 2, no. 2, pp. 139–144, 2007.

[46] L.-Z. Liu, F.-X. Wu, and W.-J. Zhang, Reverse engineering of gene regulatory networks from biological data, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 5, pp. 365–385, 2012.

[47] M. Li, H. Gao, J. Wang, and F.-X. Wu, Control principles for complex biological networksli control principles for biological networks, *Briefings in Bioinformatics*, doi: 10.1093/bib/bby088.

[48] Y. R. Wang and H. Huang, Review on statistical methods for gene network reconstruction using expression data, *Journal of Theoretical Biology*, vol. 362, pp. 53–61, 2014.

[49] J. Ruyssinck, P. Geurts, T. Dhaene, P. Demeester, and Y. Saeys, Nimefi: Gene regulatory network inference using multiple ensemble feature importance algorithms, *PLoS One*, vol. 9, no. 3, p. e92709, 2014.

[50] H. Brunel, J.-J. Gallardo-Chacón, A. Buil, M. Vallverdú, J. M. Soria, P. Caminal, and A. Perera, Miss: A non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis, *Bioinformatics*, vol. 26, no. 15, pp. 1811–1818, 2010.

[51] X. Zhang, X.-M. Zhao, K. He, L. Lu, Y. Cao, J. Liu, J.-K. Hao, Z.-P. Liu, and L. Chen, Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information, *Bioinformatics*, vol. 28, no. 1, pp. 98–104, 2011.

[52] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky, Revealing strengths and weaknesses of methods for gene network inference, *Proceedings of the National Academy of Sciences*, vol. 107, no. 14, pp. 6286–6291, 2010.

[53] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano, Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *BMC Bioinformatics*, vol. 7, no. 1, p. S7, 2006.

[54] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi, Information-theoretic inference of large transcriptional regulatory networks, *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, no. 1, p. 79879, 2007.

[55] H. Peng, F. Long, and C. Ding, Feature selection based on mutual information criteria of max-dependency, max-

relevance, and min-redundancy, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[56] J. Zhao, Y. Zhou, X. Zhang, and L. Chen, Part mutual information for quantifying direct associations in networks, *Proceedings of the National Academy of Sciences*, vol. 113, no. 18, pp. 5130–5135, 2016.

[57] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, The large-scale organization of metabolic networks, *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.

[58] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, Prediction, and Search*. MIT Press, 2000.

[59] T. Schaffter, D. Marbach, and D. Floreano, GeneNetWeaver: In silico benchmark generation and performance profiling of network inference methods, *Bioinformatics*, vol. 27, no. 16, pp. 2263–2270, 2011.

[60] T. Saito and M. Rehmsmeier, The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets, *PloS One*, vol. 10, no. 3, p. e0118432, 2015.

[61] P. E. Meyer, F. Lafitte, and G. Bontempi, minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information, *BMC Bioinformatics*, vol. 9, no. 1, p. 461, 2008.

[62] C. Olsen, P. E. Meyer, and G. Bontempi, On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information, *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2009, no. 1, p. 308959, 2008.

[63] P. Meyer, D. Marbach, S. Roy, and M. Kellis, Information-theoretic inference of gene networks using backward elimination, in *BioComp*, 2010, pp. 700–705.

[64] M. Li, X. Meng, R. Zheng, F.-X. Wu, Y. Li, Y. Pan, and J. Wang, Identification of protein complexes by using a spatial and temporal active protein interaction network, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, doi: 10.1109/TCBB.2017.2749571.

[65] M. Li, J. Yang, F.-X. Wu, Y. Pan, and J. Wang, Dynetviewer: A cytoscape app for dynamic network construction, analysis and visualization, *Bioinformatics*, vol. 34, no. 9, pp. 1597–1599, 2017.

Xiang Chen received the BS degree from Central South University, China, in 2010 and the MS degree in computer science from Harbin Institute of Technology, China, in 2012. He is currently a PhD candidate in bioinformatics at Central South University. His research interest is machine learning and its application in bioinformatics.

Min Li received the PhD degree in computer science from Central South University, China, in 2008. She is currently the vice dean and a professor at the School of Computer Science and Engineering, Central South University, Changsha, China. Her research interests include computational biology, systems biology, and bioinformatics. She has published more than 80 technical papers in refereed journals such as *Bioinformatics*, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *Proteomics*, and conference proceedings such as *BIBM*, *GIW*, and *ISBRA*.

Ruiqing Zheng received the BS degree from Central South University, China, in 2013, MS degree in 2016. He is currently a PhD candidate in bioinformatics at Central South University. His research interest is gene regulation and analysis of single-cell RNA-seq data.

Yaohang Li received the BS degree from South China University of Technology in 1997, and the MS and PhD degrees in computer science from Florida State University, Tallahassee, FL, USA, in 2000 and 2003, respectively. He is an associate professor in computer science at Old Dominion University, Norfolk, VA, USA. His research interests are in protein structure modeling, computational biology, bioinformatics, Monte Carlo methods, big data algorithms, and parallel and distributive computing. After graduation, he worked at Oak Ridge National Laboratory as a Postdoctoral Researcher for a short period in 2003. He was a Summer Research Fellow at National Center of Supercomputing Applications (NCSA) in 2007 and a Summer Faculty Research Participation member at Oak Ridge National Laboratory in 2006 and 2008. Dr. Li is the author of over 70 papers in international journals and refereed conference proceedings. He is the program committee co-chair of 2015 International Symposium on Bioinformatics Research and Applications (ISBRA2015). He also serves on the editorial boards of *International Journal of Computational Mathematics* and *Computational Biology Journal*. Dr. Li received the best poster awards at ISBRA2015 and the best paper awards at Modeling, Simulation, and Visualization Capstone Conferences in 2014 and 2015. He is the recipient of the Ralph E. Powe Award in 2005 and an NSF CAREER Award in 2009.

Fang-Xiang Wu received the BSc and MSc degrees in applied mathematics, both from Dalian University of Technology, China, in 1990 and 1993, respectively, the first PhD degree in control theory and its applications from Northwestern Polytechnical University in 1998, and the second PhD degree in biomedical engineering from the University of Saskatchewan, Canada, in 2004. Currently, he is working as an associate professor of bioengineering with the Department of Mechanical Engineering and graduate chair of the Division of Biomedical Engineering at University of Saskatchewan, Canada. His current research interests include systems biology, genomic and proteomic data analysis, biological system identification and parameter estimation, and applications of control theory to biological system.

Siyu Zhao received the BS degree from Hunan University of Science and Technology in 2015. He is currently a master student in computer technology at Central South University. His research interest is network analysis in bioinformatics.

Jianxin Wang received the BEng and MEng degrees in computer engineering from Central South University, China, in 1992 and 1996, respectively, and the PhD degree in computer science from Central South University, China, in 2001. He is the dean and a professor in School of Computer Science and Engineering, Central South University, Changsha, China. His current research interests include algorithm analysis and optimization, parameterized algorithm, bioinformatics, and computer network. He has published more than 150 papers in various international journals and refereed conferences. He is a senior member of the IEEE.