Article

A novel method of gene regulatory network structure inference on gene knock-out expression data

Xiang Chen [1], Min Li [1]*, Fang-Xiang Wu[1,2], Yaohang Li[3] and Jianxin Wang [1]

[1] School of Information Science and Engineering, Central South University, Changsha 410083, China; chenxofhit@gmail.com (X.C.); faw341@mail.usask.ca (F.W.); jxwang@mail.csu.edu.cn (J.W.)

[2] Department of Mechanical Engineering and Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, Canada

[3] Department of Computer Science, Old Dominion University, Norfolk, VA, USA; yaohang@cs.odu.edu (Y.L.)

* Correspondence: limin@mail.csu.edu.cn (M.L.); Tel.: +86-0731-888-30212

Academic Editor: Quan Zou; Le Zhang
Version October 19, 2017 submitted to Int. J. Mol. Sci.

Abstract: Inferring gene regulatory networks (GRNs) structure from gene expression data has always been a pretty challenging problem in systems biology. It is critical to identify complicated regulatory relationships among genes and understanding regulatory mechanisms in cells. Various methods based on information theory have been developed to infer networks. However, these methods introduce many redundant regulatory relationships in the network inference process due to external noise in the data, topology sparseness in the network structure and non-linear dependency among genes. As a result they are not suitable for accurate structure inference on gene expression data. In this paper, a novel network structure inference method named Loc-PCA-CMI is proposed as to first identify Local Overlapped gene Clusters, and then in conjunction with PCA-CMI method each local cluster structure is refined. The final dependence score of every edge between genes is denoted as the average score of the corresponding edge in each local cluster structure. The proposed method is evaluated on DREAM3 knock-out dataset, and its performance is compared to other information theory based network inference methods including ARACNE, MRNET, PCA-CMI and PCA-PMI. Experimental results demonstrate that our method Loc-PCA-CMI outperforms significantly than other previous methods. Code is available at http://github.com/chenxofhit/Loc-PCA-CMI.

Keywords: Gene Regulatory Networks ; Network inference; Path consistency algorithm

1. Introduction

A critical problem in systems biology is to recover gene regulatory networks (GRNs), which can help biomedical scientists to identify complicated regulatory relationships among genes and also to look deeply into 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 require substantial time and considerable cost. With the development of high-throughput technologies, expression data for tens of thousands of genes can be produced, which makes it possible for GRNs to be inferred from these expression data based on computational methods [4]. In recent years, the inference of networks based on computational methods has become one of the most important goals [1,5], and which is more complex and challenging especially when to deal with large scale gene expression data in the post-genome era.

A gene regulatory network can be typically described by a graph in which each node corresponds to a gene and each edge represents a regulatory relationship between genes in graph theory. And GRNs can be viewed as undirected acyclic graph if not considering up-streaming or down-streaming regulatory relationship among genes and ignore the self-regulatory mechanism [6]. Thus, network structure can be reconstructed by accurately inferring the underlying regulatory interactions among genes from the gene expression data. Unfortunately, gene expression data are typically in high dimensions and with small sample size which suffer from "dimensionality curse" [7]. Furthermore, gene expression data usually involve large amounts of external noise and non-linear relationships. All of these issues make it difficult and challenging to accurately identify regulatory interactions among genes from gene expression data.

To construct accurate GRN structures from expression data, various computational methods have been proposed based on a variety of different assumptions and different conditions [8,9]. These algorithms can be divided into two main categories: model-based and similarity-based [10,11]. Model-based algorithms usually infer regulatory interactions based on computational model learning. The typical models include Boolean network [12,13], Bayesian network [14,15], and differential equation models [11]. The Boolean network model is the simplest network model, which is implemented through Boolean variables and abstract Boolean logic. Because the state of gene expression is considered to be only active or inactive, Boolean network models can not capture complex system behaviour [16]. The Bayesian network model is a popular probabilistic graphical model in which the dependency relationships among genes are described via a directed acyclic graph (DAG). The Bayesian network model outperforms other models in dealing with noise and incorporating prior knowledge, but structure learning in the model is an NP-hard problem [17]. 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. Therefore, the regulatory interactions among genes can be identified by the parameter set, which is obtained according to the expression data and the equation model.

Other than model-based algorithms, similarity-based algorithms identify regulatory interactions only by measuring dependences between 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. Pearson's correlation, rank correlation and Euclidean distance are typically used [18] to measure gene-gene co-expression. However, the correlation-base method can not identify complex dependencies between genes, such as non-linear dependencies [19]. Furthermore, some functionally related genes might not be co-expressed, which makes it difficult to accurately identify regulatory interactions. The information theory-based method is also a representative similarity-based algorithm, in which mutual information (MI) is used to measure potential dependency among genes. As MI can capture non-linear dependencies effectively [20,21], the information theory-based method is widely applied to identify complicated regulatory interactions in GRNs.

In this paper, we focus on the network inference method based on information theory. In recent years, various network inference methods based on information theory have been developed, which focus on distinguishing direct regulatory interactions from indirect associations [22]. To eliminate indirect interactions, Margolin et al.[23] proposed the ARACNE method based on Data Processing Inequality (DPI) with interaction triangles considered. The minimum-redundancy network (MRNET) by Meyer [24] uses a minimum redundancy feature selection method [25], wherein for each candidate gene in a obtained MI network, it selects a subset of its highly relevant genes while minimizing the MI-based criteria between the selected genes. Zhang et al. [21] introduce a network inference algorithm called the conditional mutual information-based path consistency algorithm (PCA-CMI) and later Zhao et al.[26] introduce a network inference algorithm called the part mutual information-based path consistency algorithm (PCA-PMI). Both CMI and PMI are presented to measure the nonlinearly direct associations between genes.

Path consistency (PC) algorithm is an exhaustive algorithm and a trade-off is usually made between time and accuracy in both PCA-CMI and PCA-PMI, hence it is not suitable for relatively large gene networks. To overcome this difficulty we first use top ranked highly co-expressed genes as candidates for local clusters and then each cluster's accurate structure is refined with PCA-CMI. The final structure of the GRN is then inferred with an ensemble all the local structure together. We named this novel approach or workflow as Loc-PCA-CMI and apparently Loc-PCA-CMI method can deal with relatively large datasets and benefit from the relatively accurate structure inference to small number gene subnetworks with PCA-CMI.

2. Methods

In this section, we will introduce some definitions of information theory including entropy, MI and CMI, as well as the algorithm of Loc-PCA-CMI for inferring GRNs.

2.1. Information Theory

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 between genes. The definitions of mutual information (MI) and conditional mutual information (CMI) are as follows:

$$MI(X,Y) = \int\int p(x,y)log\frac{p(x,y)}{p(x)p(y)}dxdy \tag{1}$$

$$I(X,Y|Z) = \int\int\int p(x,y,z)log\frac{p(x,y|z)}{p(x|z)p(y|z)} \tag{2}$$

where $p(x,y)$ denotes the joint distribution of X and Y. $p(x)$ and $p(y)$ represent the marginal distribution of x and y, respectively.

CMI can also be expressed in terms of entropies as :

$$H(X,Y|Z) = H(X,Z) + H(Y,Z) - H(Z) - H(X,Y,Z) \tag{3}$$

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

Since it is widely accepted that gene expression data follow Gaussian distribution [21], formulation of entropy subject to n-dim Gaussian distribution can be easily calculated by a simple equation, where $|C|$ is the determinant of covariance matrix of variables x_1,x_2,...,x_n [27].

$$H(X) = log(2\pi e)^{\frac{n}{2}}|C|^{-\frac{1}{2}} \tag{4}$$

After mathematical transformation, we can obtain the following equation:

$$MI(X,Y) = \frac{1}{2}log\frac{|C(X)| * |C(X)|}{|C(X,Y)|} \tag{5}$$

2.2. Path consistency algorithms(PCA)

Path consistency (PC) algorithm is widely used in inferring gene regulatory networks [21]. By removing the most likely uncorrelated edges repeatedly from low to high order dependence correlation until no edges can be removed, PC-algorithm can construct a high-confidence undirected network [28].

2.3. Loc-PCA-CMI

Our proposed method Loc-PCA-CMI method, firstly reduces the search space by only picking the top M ranked edges by Pearson correlation analysis with FDR correction in p-value, and secondly obtain local overlapped cluster with each edge linked gene. The overlapped cluster differ from each other with different node set after a filter operation. Then for each local cluster we apply PCA-CMI algorithm to infer its structure. Final edge weight of the complete regulatory network is obtained by averaging edge weight with each inferred cluster structure. The implementation details are shown below in Algorithm 1:

Algorithm 1 Loc-PCA-CMI

Require: X (gene expression matrix), GS (gene names set), m (number of gene), n (number of top ranked edges); k (CMI order number) and β (order threshold) in subroutine PCA-CMI.
Ensure: Weight matrix W

```
if $N \leq 10$ then
    $W \leftarrow PCA-CMI(X, k, \beta)$;
    return $W$
else
    Construct pair-wise Pearson correlation matrix $\Omega = \rho(X_i X_j)$;
    Select top $n$ edges as $E$ with highest Pearson correlation values in $\Omega$ with FDR correction in p-value;
    for each gene $i$ in $GS$ do
        Retrieve its direct connected genes that in edges list $E$ as local cluster $Loc(i)$;
    Filter repeated local cluster and obtain final local cluster as $Loc$;
    for each cluster $C$ in $Loc$ do
        $w_C \leftarrow PCA-CMI(C, k, \beta)$;
    $W_{X_i X_j} \leftarrow mean(w)$
    return $W$
```

3. Materials

We benchmarked the performance of our approach, Loc-PCA-CMI using six simulation data from well known DREAM3 challenge [29]. 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 are evaluated in our experiments, which include three different size varying in 10, 50, 100 with two types E.coli and S.cerevisiae respectively. Table 1 gives detailed descriptions of the datasets.

Table 1. Descriptions of the datasets in our experiments

Datasets	#Samples	#Average degree	#Max degree	#Network density
DREAM3-10 Ecoli	11	2.2	5	0.244
DREAM3-50 Ecoli	51	2.48	14	0.051
DREAM3-100 Ecoli	101	2.5	14	0.025
DREAM3-10 Yeast	11	2	4	0.222
DREAM3-50 Yeast	51	3.08	13	0.063
DREAM3-100 Yeast	101	3.32	10	0.034

4. Results and Discussion

As described in Algorithm 1 three parameters affect the performance of the method Loc-PCA-CMI in network structure inference. The first parameter is the top selected edges number n, if n increase more edges will be taken into consideration and the local cluster size will increase. The second parameter is β which acts as the threshold value of MI and CMI to decide independence. The third parameter is CMI order number k, theoretically by increasing k the structure will be more accurate if CMI not reach the threshold β in $k-1$ order. Latter two parameters exist in PCA-CMI, and for comparison we denote β=0.03 and k=2 as described also in [21]. Best n can be obtained by cross validation and in our experiments $n = 20\% * \binom{m}{2}$.

Another option should be noticed in Algorithm 1 is that after local cluster structure obtained in Algorithm 1 , both PCA-CMI and PCA-PMI are alternatives for the subsequent structure refinement process in fact. We also validate the PCA-PMI alogrithm for experiment integrity on the benchmarked datasets and named it as Loc-PCA-PMI for further presentation as a convenience.

We assessed the performance of Loc-PCA-CMI by evaluating the areas under the Receiver Operating Characteristic (AUROC) and the Precision-Recall curve (AUPR). As in sparse biological networks the number of non-existing edges (negatives) outweighs the number of existing edges (positives) significantly, which AUPR is more informative to AUROC in fact [30]. We tend to use AUPR for evaluation, but for a further comparison with other methods that adopt AUROC as evaluation metric we also take AUROC as supplementary metric. Higher AUROC and AUPR values indicate 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 ($q = 1, 2, ..., m^2$). Similarly, the precision (TP/(TP+FP)) and recall (TP/(TP+FN)) curve was plotted for increasing q.

Table 2 gives the AUROC and AUPR value of six methods including ARACNE, MRNET, PCA-PMI, PCA-CMI, Loc-PCA-PMI, Loc-PCA-CMI on the benchmarked datasets. As a whole we can observe that Loc-PCA-PMI and Loc-PCA-CMI significantly outperformed than the other four methods ARACNE, MRNET, PCA-PMI, PCA-CMI both in AUROC and AUPR value. Furthermore, going deep into the table and we can get two conclusions below:

- Comparing Loc-PCA-PMI to PCA-PMI or Loc-PCA-CMI to PCA-CMI it can be inferred that local cluster strategy adopted in the algorithm can greatly improve the performance of PCA-PMI or PCA-CMI both in AUROC and AUPR value.
- Comparing Loc-PCA-PMI to Loc-PCA-CMI , AUPR value of Loc-PCA-CMI is significantly higher than that of Loc-PCA-PMI especially in a larger size 50 and size 100 dataset. As stated previously AUPR is more meaningful when to tackle with sparse network structure prediction issues, it can be inferred that Loc-PCA-CMI is more suitable for large gene regulatory network structure inference.

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

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

5. Conclusion

We have proposed a novel method of gene regulatory network structure inference named Loc-PCA-CMI in this paper. Experiments on DREAM3 knock-out dataset show that out proposed method performs better compared with other methods. The algorithm has a tremendous potential to be applied for gene regulatory network structure inference in large scale gene expression data.

Acknowledgments: This work was supported in part by the National Natural Science Foundation of China underGrants (No. 61622213, No. 61370024 and No. 61420106009).

Author Contributions: Xiang Chen designed the experiments and wrote the manuscript. Fang-Xiang Wu, Yaohang Li and Jianxin Wang revised the manuscript; Min Li systemically revised the manuscript before submission.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Altay, G.; Emmert-Streib, F. Inferring the conservative causal core of gene regulatory networks. *BMC Systems Biology* **2010**, *4*, 132.
2. Basso, K.; Margolin, A.A.; Stolovitzky, G.; Klein, U.; Dalla-Favera, R.; Califano, A. Reverse engineering of regulatory networks in human B cells. *Nature genetics* **2005**, *37*, 382.
3. Elnitski, L.; Jin, V.X.; Farnham, P.J.; Jones, S.J. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome research* **2006**, *16*, 1455–1464.
4. Maetschke, S.R.; Madhamshettiwar, P.B.; Davis, M.J.; Ragan, M.A. Supervised, semi-supervised and unsupervised inference of gene regulatory networks. *Briefings in bioinformatics* **2013**, *15*, 195–211.
5. Margolin, A.A.; Wang, K.; Lim, W.K.; Kustagi, M.; Nemenman, I.; Califano, A. Reverse engineering cellular networks. *Nature protocols* **2006**, *1*, 662.
6. Irrthum, A.; Wehenkel, L.; Geurts, P.; others. Inferring regulatory networks from expression data using tree-based methods. *PloS one* **2010**, *5*, e12776.
7. Wang, Y.; Joshi, T.; Zhang, X.S.; Xu, D.; Chen, L. Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics* **2006**, *22*, 2413–2420.
8. Longabaugh, W.J.; Davidson, E.H.; Bolouri, H. Computational representation of developmental genetic regulatory networks. *Developmental biology* **2005**, *283*, 1–16.
9. Karlebach, G.; Shamir, R. Modelling and analysis of gene regulatory networks. *Nature reviews. Molecular cell biology* **2008**, *9*, 770.
10. Bansal, M.; Belcastro, V.; Ambesi-Impiombato, A.; Di Bernardo, D. How to infer gene networks from expression profiles. *Molecular systems biology* **2007**, *3*, 78.
11. Li, Z.; Li, P.; Krishnan, A.; Liu, J. Large-scale dynamic gene regulatory network inference combining differential equation models with local dynamic Bayesian network analysis. *Bioinformatics* **2011**, *27*, 2686–2691.
12. Kim, H.; Lee, J.K.; Park, T. Boolean networks using the chi-square test for inferring large-scale gene regulatory networks. *BMC bioinformatics* **2007**, *8*, 37.
13. Zhou, J.X.; Samal, A.; d'Hérouël, A.F.; Price, N.D.; Huang, S. Relative stability of network states in Boolean network models of gene regulation in development. *Biosystems* **2016**, *142*, 15–24.
14. Chen, X.w.; Anantha, G.; Wang, X. An effective structure learning method for constructing gene networks. *Bioinformatics* **2006**, *22*, 1367–1374.
15. Lo, L.Y.; Wong, M.L.; Lee, K.H.; Leung, K.S. High-order dynamic Bayesian Network learning with hidden common causes for causal gene regulatory network. *BMC bioinformatics* **2015**, *16*, 395.
16. Lee, W.P.; Tzou, W.S. Computational methods for discovering gene networks from expression data. *Briefings in bioinformatics* **2009**, *10*, 408–423.
17. Chickering, D.M.; Heckerman, D.; Meek, C. Large-sample learning of Bayesian networks is NP-hard. *Journal of Machine Learning Research* **2004**, *5*, 1287–1330.
18. Wang, Y.R.; Huang, H. Review on statistical methods for gene network reconstruction using expression data. *Journal of theoretical biology* **2014**, *362*, 53–61.
19. Ruyssinck, J.; Geurts, P.; Dhaene, T.; Demeester, P.; Saeys, Y.; others. Nimefi: gene regulatory network inference using multiple ensemble feature importance algorithms. *PLoS One* **2014**, *9*, e92709.
20. Brunel, H.; Gallardo-Chacón, J.J.; Buil, A.; Vallverdú, M.; Soria, J.M.; Caminal, P.; Perera, A. MISS: a non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis. *Bioinformatics* **2010**, *26*, 1811–1818.
21. Zhang, X.; Zhao, X.M.; He, K.; Lu, L.; Cao, Y.; Liu, J.; Hao, J.K.; Liu, Z.P.; Chen, L. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics* **2011**, *28*, 98–104.

22. Marbach, D.; Prill, R.J.; Schaffter, T.; Mattiussi, C.; Floreano, D.; Stolovitzky, G. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the national academy of sciences* **2010**, *107*, 6286–6291.

23. Margolin, A.A.; Nemenman, I.; Basso, K.; Wiggins, C.; Stolovitzky, G.; Dalla Favera, R.; Califano, A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics* **2006**, *7*, S7.

24. Meyer, P.E.; Kontos, K.; Lafitte, F.; Bontempi, G. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP journal on bioinformatics and systems biology* **2007**, *2007*, 79879.

25. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* **2005**, *27*, 1226–1238.

26. Zhao, J.; Zhou, Y.; Zhang, X.; Chen, L. Part mutual information for quantifying direct associations in networks. *Proceedings of the National Academy of Sciences* **2016**, *113*, 5130–5135.

27. Shannon, C.E.; Weaver, W. *The mathematical theory of communication*; University of Illinois press, 1998.

28. Spirtes, P.; Glymour, C.N.; Scheines, R. *Causation, prediction, and search*; MIT press, 2000.

29. Schaffter, T.; Marbach, D.; Floreano, D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics* **2011**, *27*, 2263–2270.

30. Saito, T.; Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one* **2015**, *10*, e0118432.