Article

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

Xiang Chen [1], Min Li [1]*, Ruiqing Zheng [1], Siyu Zhao [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.); rqzheng@csu.edu.cn (R.Z.); syzhao@csu.edu.cn (S.Z.); 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 December 2, 2017 submitted to Int. J. Mol. Sci.

Abstract: Inferring gene regulatory networks (GRNs) structure from gene expression data has 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 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. 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 the accurate local cluster structure is inferred by PCA-CMI method. The final structure of the GRN is denoted as dependence among genes by ensemble of the obtained local cluster structures. Loc-PCA-CMI 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 this novel method Loc-PCA-CMI outperforms other four methods significantly. Source code which includes algorithm and evaluation is freely available for download at http://github.com/chenxofhit/Loc-PCA-CMI.

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

1. Introduction

To identify and understand gene regulatory networks (GRNs) is a critical problem in systems biology, which can help biomedical scientists explicitly to 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 require substantial time and considerable cost. Owing to the development of micro-array technologies tremendous amounts of gene expression data has 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], and which is more complex and challenging especially when to deal with large scale gene expression data in the post-genome era.

GRNs can be viewed as undirected acyclic graph if not considering both up-streaming or down-streaming regulatory relationship among genes and ignore the self-regulatory mechanism [7], in which each node corresponds to a gene and each edge represents a regulatory relationship between genes. Unfortunately, gene expression data are typically in high dimensions and relative small sample size which suffer from "dimensionality curse" [8]. Furthermore, gene expression data usually involve large amounts of external noise and non-linear relationships. All of these issues make it challenging to accurately identify regulatory interactions among genes from gene expression data.

Various computational methods to construct accurate GRN structures from expression data have been proposed based on a variety of different assumptions and different conditions [9,10]. 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 [11–14], Bayesian network [15–19], and differential equation models [20–25]. 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 entirely capture complex system behavior [26]. 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 computationally intensive and has been proved as an NP-hard problem [27]. 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 all other related genes. A major challenge to use 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 [28,29].

Other than model-based methods, model-free approaches identify regulatory interactions mainly 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 include Pearson correlation, rank correlation, Euclidean distance, and the angle between a pair of observed expression vectors [30]. However, the correlation-based method can not identify complex dependencies between genes, such as non-linear dependencies [31]. Furthermore, quite a few 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 model-free method, in which mutual information (MI) is favored to measure potential dependency among genes as it can capture non-linear dependencies effectively [32,33]. In recent years, various network inference methods based on information theory have been proposed, which focus on distinguishing direct regulatory interactions from indirect associations [34]. To eliminate indirect interactions, Margolin et al. [35] proposed the ARACNE method based on Data Processing Inequality (DPI) with interaction triangles considered. The minimum-redundancy network (MRNET) by Meyer [36] uses a minimum redundancy feature selection method [37], 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. [33] introduce a network inference algorithm called the conditional mutual information-based path consistency algorithm (PCA-CMI) and later Zhao et al. [38] introduce a network inference algorithm called the part mutual information-based path consistency algorithm (PCA-PMI). Path consistency algorithm (PCA) is an exhaustive algorithm which is widely used in inferring GRNs [33]. A trade-off is usually made between time and accuracy in both PCA-CMI and PCA-PMI, and as the network size expands more uncontrollable external noise to the instinct complex network structure makes prediction accuracy of GRNs decreases dramatically. To improve this situation motivated by the divide and conquer strategy, we first use top ranked highly co-expressed genes as centroids of 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 of all the local structure together. We name this novel approach as Loc-PCA-CMI hereafter and intuitively Loc-PCA-CMI method can deal with relatively larger 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 related work of information theory including entropy, MI and CMI, as well as the algorithm of 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 between genes. The definitions of mutual information (MI) and conditional mutual information (CMI) are as follows:

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

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

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 [33], 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 [39].

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

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. 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 [40], consequently the GRN is a sparse network. A key step of identifying the sparse structure of the network is to identify the significant edges that may have a comparatively high co-expressed expression value. As a result, our proposed method Loc-PCA-CMI firstly select the top n highly co-expressed edges by Pearson correlation analysis with FDR correction in p-value, and secondly in the reduced edges space we can obtain local overlapped cluster with genes connected by edges. Then for each local cluster we apply PCA-CMI algorithm, which can construct a high-confidence undirected network [41] 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 cluster's accurate structure. Final edge weight of the complete regulatory network is obtained by averaging edge weight simply with each inferred cluster structure. The entire pipeline is provided in Figure 1 and the implementation details are shown below in Algorithm 1. As PCA-CMI is extremely competent for relative small GRN structure inference, we make a preprocessing to check the number of gene m in the algorithm, if m is less or equal than a constant c then PCA-CMI is directly applied for the GRN structure inference.

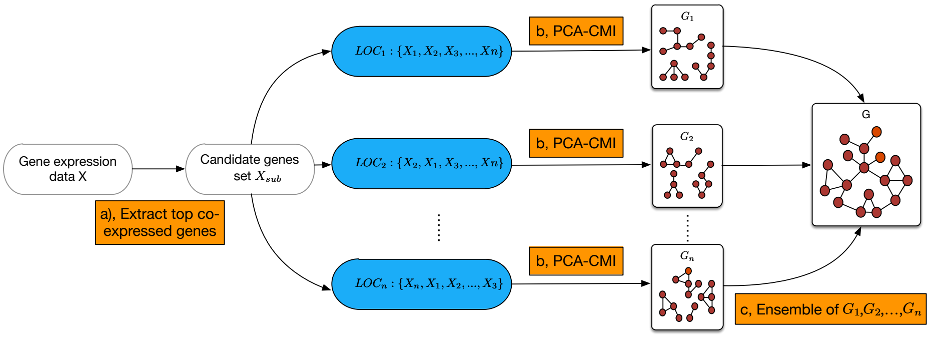

Gene expression data X
Candidate genes set X_{sub}
a), Extract top co-expressed genes
$LOC_1 : \{X_1, X_2, X_3, ..., Xn\}$
$LOC_2 : \{X_2, X_1, X_3, ..., Xn\}$
$LOC_n : \{X_n, X_1, X_2, ..., X_3\}$
b, PCA-CMI
b, PCA-CMI
b, PCA-CMI
G_1
G_2
G_n
G
c, Ensemble of G_1,G_2,...,G_n

Figure 1. The Loc-PCA-CMI framework. a), We first extract top co-expressed genes from gene expression data matrix X as candidate genes X_{sub}. The candidate genes X_{sub} are then grouped into different local clusters with each gene in X_{sub} as the centroid. b), For each local overlapped cluster PCA-CMI is applied to get accurate structure. c), Ensemble of diverse cluster structure G_1, G_2, ..., G_n to obtain the final structure of the GRN as G.

Algorithm 1 Loc-PCA-CMI

Require: X (gene expression data matrix), m (number of gene), n (number of top ranked edges), c (constant number); k (CMI order number) and β (order threshold) in subroutine PCA-CMI.
Ensure: Graph weight matrix G

```
if $m \leq c$ then
    $G \leftarrow$ PCA-CMI$(X, k, \beta)$;
    return $G$
else
    Construct pair-wise Pearson correlation matrix $\Omega = \rho(X_i, X_j)$;
    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 the candidate genes as $X_{sub}$ ;
    for each gene $i$ in $X_{sub}$ do
        Retrieve its direct connected genes that in edges list $E$ as local cluster $Loc(i)$;
    for each cluster $C$ in $Loc$ do
        $g_C \leftarrow$ PCA-CMI$(C, k, \beta)$;
    $G \leftarrow$ mean$(g_1, g_2, ..., g_t)$, where $t$ stands for total cluster number of $Loc$;
    return $G$
```

The runtime complexity of Algorithm 1 is generally determined by total cluster number t and the complexity of each invoked subroutine PCA-CMI to the special cluster. t is in the same order with the number of genes m. The runtime complexity of PCA-CMI comes mainly from the parameter CMI order number k and cluster size $|C|$ of C, which can be roughly estimated as $O(|C|^k)$. As a result the final runtime complexity of PCA-CMI can be calculated as $O(m * |C|^k)$. At worst, if cluster size $|C|$ equals to m i.e. every cluster contains all the genes in it and the runtime complexity is thus $m * m^k = m^{k+1}$. However, this worst case scenario never happens in practice; $|C|$ is usually much lower than m.

3. Materials

We benchmarked the performance of our approach, Loc-PCA-CMI using six simulation data from well known DREAM3 challenge [42]. 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(Max) degree	#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

4. Results and Discussion

As described in Algorithm 1 three intrinsic parameters affect the performance of Loc-PCA-CMI in GRN structure inference. The first parameter is the top selected edges number n, if n increase more edges will be taken as significant edges into consideration and the local cluster size will increase subsequently. 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 also the following described method PCA-PMI. Best n can be obtained by cross validation and generally larger n can contribute to a larger size cluster and more genes will be covered in the network, in our experiments we attribute it as a constant with value $n = 20\% * \binom{m}{2}$ uniformly. Besides the above three intrinsic parameters we set constant $c = 10$ in Algorithm 1, i.e. if number of gene is less or equal to 10 Loc-PCA-CMI invokes PCA-CMI directly and in this case performance of Loc-PCA-CMI and PCA-CMI are the same.

We assess 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 [43]. We tend to use AUPR for evaluation, but for a conservative comparison with other methods that adopt AUROC as evaluation metric we also take AUROC as supplementary metric. Higher AUROC and AUPR value indicate more accurate GRN predictions. For this purpose, we compute 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 is 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 is plotted for increasing q.

It is should be noticed that in Algorithm 1 after each local cluster is obtained both PCA-CMI and PCA-PMI are alternatives for the subsequent structure refinement. If PCA-CMI is replaced with PCA-PMI a novel method is generated and we name it as Loc-PCA-PMI analogously. Then four PCA based methods are derived including PCA-PMI, PCA-CMI, Loc-PCA-PMI, Loc-PCA-CMI at present, all of which are belong to model-free methods. As showed in Table 1 among the six benchmark datasets DREAM10-Ecoli and DREAM10-Yeast datasets contain only 10 genes, hence Loc-PCA-CMI and PCA-CMI are identical in performance and the same with Loc-PCA-PMI and PCA-PMI according to the principle of Algorithm 1. For meaningful comparison of these PCA based methods we select other four datasets whose gene number is greater than 10. Order number is not explicitly discussed in [33,38] wherein $\beta = 0.03$ and $k = 2$ are configured directly, we are curious about how order number k affect 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 value can be calculated respectively. Figure 2 illustrates the result in summary on the benchmark datasets, and from which we can induce three conclusions:

- Order number k affect the results of these four PCA based methods, generally when k reaches 4 AUPR and AUROC become stable except those in DREAM3-100 Ecoli dataset.
- Loc-PCA-CMI and Loc-PCA-PMI yield higher AUPR and AUROC value than PCA-CMI and PCA-PMI respectively, hence that local cluster strategy adopted in the algorithm helps to improve the performance of PCA-CMI and PCA-PMI significantly.
- Loc-PCA-CMI outperforms significantly than the other three methods in the metric of AUPR, which is more meaningful when to tackle with sparse network structure prediction issues.

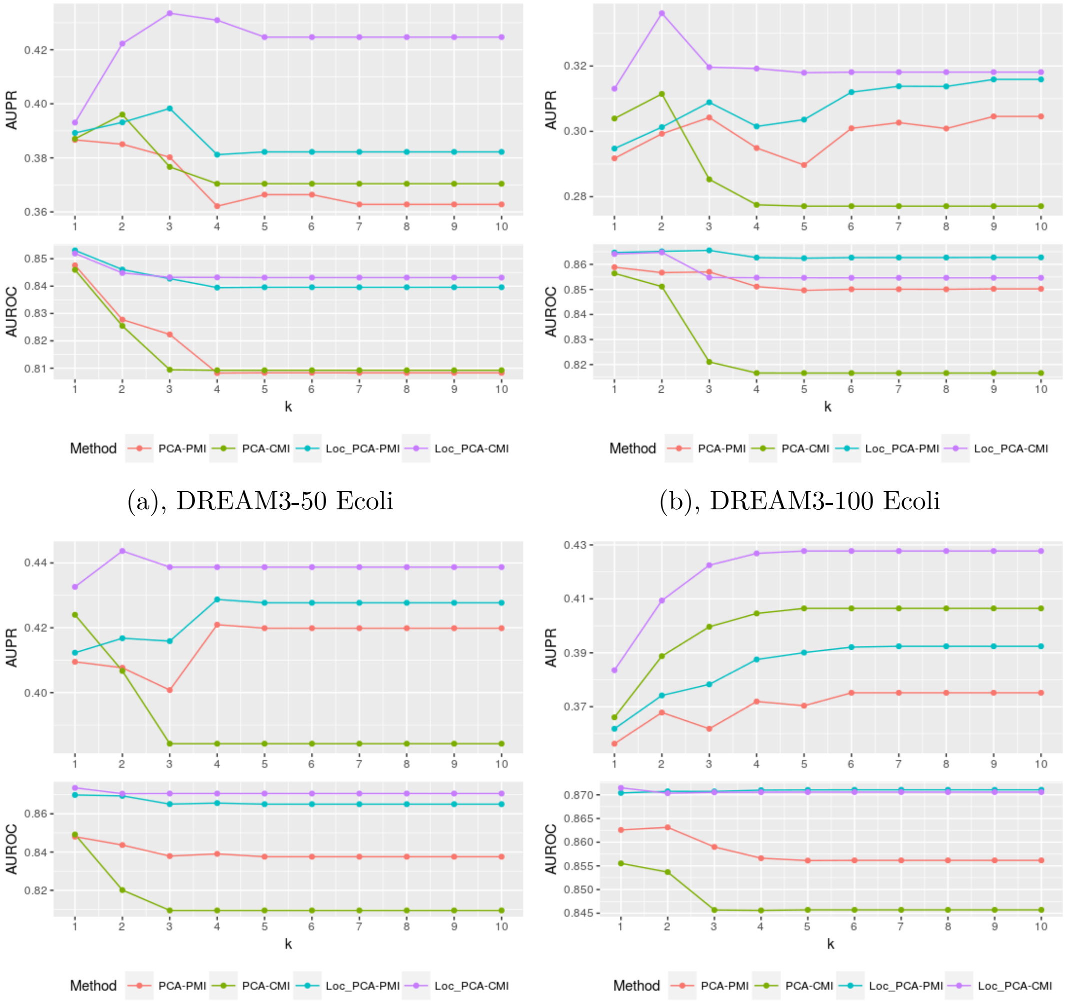
We also conduct a comparison experiment using Loc-PCA-CMI with four previously proposed methods on all the six benchmark datasets, which include ARACNE, MRNET, PCA-PMI, PCA-CMI. we use the R package "minet" with default parameters for evaluation of ARACNE and MRNET [44]. The MI matrices of the methods are approximated using Pearson correlation directly from continuous gene knock-out expression data [45,46]. For implementation of PCA-PMI and PCA-CMI we have downloaded the MATLAB codes according to URL provided in [33,38]. We prefer the default value of parameters in PCA-PMI and PCA-CMI, where $\beta = 0.03$ and $k = 2$. For Loc-PCA-CMI we also attribute the same value to these two parameters for comparison. Table 2 gives the AUROC and AUPR value of this experiment. From the table we can see that AUPR value of both PCA-CMI and PCA-PMI decrease dramatically when the network size expands. And Loc-PCA-CMI rank only second to PCA-PMI in DREAM3-10 Yeast dataset, while in other five datasets it outperforms significantly the other four methods ARACNE, MRNET, PCA-PMI, PCA-CMI both in AUROC and AUPR value. We provide the source codes including all the methods, benchmark datasets and evaluation scripts at http://github.com/chenxofhit/Loc-PCA-CMI.

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

Dataset	ARACNE		MRNET		PCA-PMI		PCA-CMI		Loc-PCA-CMI	
	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	**0.825**	**0.499**
DREAM3-50 Ecoli	0.474	0.050	0.529	0.061	0.828	0.385	0.825	0.396	**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**	**0.336**
DREAM3-10 Yeast	0.628	0.321	0.644	0.322	**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	0.820	0.406	**0.871**	**0.444**
DREAM3-100 Yeast	0.547	0.040	0.556	0.042	0.863	0.368	0.854	0.389	**0.870**	**0.409**

5. Conclusion

We have proposed a novel gene regulatory network structure inference method named Loc-PCA-CMI, which can be divided into model-free method in this manuscript. Experiments on DREAM3 knock-out datasets show that Loc-PCA-CMI benefits from the local cluster strategy. Loc-PCA-CMI outperforms other comparing methods including ARACNE, MRNET, PCA-PMI, PCA-CMI especially in size 50 and 100 networks. All the experiments are conducted in the DREAM 3 challenge silico datasets, and a major limitation for many studies in computational biology is the



(a), DREAM3-50 Ecoli

(b), DREAM3-100 Ecoli

(c), DREAM3-50 Yeast

(d), DREAM3-100 Yeast

Figure 2. AUPR and AUROC value 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.

lack of systematic, large-scale gold standards on which to evaluate the models. Accordingly, lack of real large GRNs we can not deduce directly that Loc-PCA-CMI can outperform other methods but we are believing that Loc-PCA-CMI could be one of the applicable approaches for large GRNs structure inference in the future.

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 raw manuscript. Min Li, Ruiqing Zheng and Siyu Zhao proposed constructive suggestion about the manuscript and also detail of the experiments. Min Li, Fang-Xiang Wu, Yaohang Li and Jianxin Wang 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. Hughes, T.R.; Marton, M.J.; Jones, A.R.; Roberts, C.J.; Stoughton, R.; Armour, C.D.; Bennett, H.A.; Coffey, E.; Dai, H.; He, Y.D.; others. Functional discovery via a compendium of expression profiles. *Cell* **2000**, *102*, 109–126.
5. 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.
6. Margolin, A.A.; Wang, K.; Lim, W.K.; Kustagi, M.; Nemenman, I.; Califano, A. Reverse engineering cellular networks. *Nature protocols* **2006**, *1*, 662.
7. Irrthum, A.; Wehenkel, L.; Geurts, P.; others. Inferring regulatory networks from expression data using tree-based methods. *PloS one* **2010**, *5*, e12776.
8. Wang, Y.; Joshi, T.; Zhang, X.S.; Xu, D.; Chen, L. Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics* **2006**, *22*, 2413–2420.
9. Longabaugh, W.J.; Davidson, E.H.; Bolouri, H. Computational representation of developmental genetic regulatory networks. *Developmental biology* **2005**, *283*, 1–16.
10. Karlebach, G.; Shamir, R. Modelling and analysis of gene regulatory networks. *Nature reviews. Molecular cell biology* **2008**, *9*, 770.
11. Shmulevich, I.; Dougherty, E.R.; Kim, S.; Zhang, W. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics* **2002**, *18*, 261–274.
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. Bornholdt, S. Boolean network models of cellular regulation: prospects and limitations. *Journal of the Royal Society Interface* **2008**, *5*, S85–S94.
14. 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.
15. Kim, S.Y.; Imoto, S.; Miyano, S. Inferring gene networks from time series microarray data using dynamic Bayesian networks. *Briefings in bioinformatics* **2003**, *4*, 228–235.
16. Zou, M.; Conzen, S.D. A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics* **2004**, *21*, 71–79.
17. Chen, X.w.; Anantha, G.; Wang, X. An effective structure learning method for constructing gene networks. *Bioinformatics* **2006**, *22*, 1367–1374.
18. Needham, C.J.; Bradford, J.R.; Bulpitt, A.J.; Westhead, D.R. A primer on learning in Bayesian networks for computational biology. *PLoS computational biology* **2007**, *3*, e129.
19. 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.

20. Gardner, T.S.; Di Bernardo, D.; Lorenz, D.; Collins, J.J. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **2003**, *301*, 102–105.
21. di Bernardo, D.; Thompson, M.J.; Gardner, T.S.; Chobot, S.E.; Eastwood, E.L.; Wojtovich, A.P.; Elliott, S.J.; Schaus, S.E.; Collins, J.J. Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks. *Nature biotechnology* **2005**, *23*, 377–383.
22. Bansal, M.; Gatta, G.D.; Di Bernardo, D. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics* **2006**, *22*, 815–822.
23. Honkela, A.; Girardot, C.; Gustafson, E.H.; Liu, Y.H.; Furlong, E.E.; Lawrence, N.D.; Rattray, M. Model-based method for transcription factor target identification with limited data. *Proceedings of the National Academy of Sciences* **2010**, *107*, 7793–7798.
24. Lu, T.; Liang, H.; Li, H.; Wu, H. High-dimensional ODEs coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. *Journal of the American Statistical Association* **2011**, *106*, 1242–1258.
25. 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.
26. Lee, W.P.; Tzou, W.S. Computational methods for discovering gene networks from expression data. *Briefings in bioinformatics* **2009**, *10*, 408–423.
27. 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.
28. Hecker, M.; Lambeck, S.; Toepfer, S.; Van Someren, E.; Guthke, R. Gene regulatory network inference: data integration in dynamic models—a review. *Biosystems* **2009**, *96*, 86–103.
29. Marbach, D.; Costello, J.C.; Küffner, R.; Vega, N.M.; Prill, R.J.; Camacho, D.M.; Allison, K.R.; Kellis, M.; Collins, J.J.; Stolovitzky, G.; others. Wisdom of crowds for robust gene network inference. *Nature methods* **2012**, *9*, 796–804.
30. Wang, Y.R.; Huang, H. Review on statistical methods for gene network reconstruction using expression data. *Journal of theoretical biology* **2014**, *362*, 53–61.
31. 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.
32. 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.
33. 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.
34. 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.
35. 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.
36. 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.
37. 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.
38. 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.
39. Shannon, C.E.; Weaver, W. *The mathematical theory of communication*; University of Illinois press, 1998.
40. Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z.N.; Barabási, A.L. The large-scale organization of metabolic networks. *Nature* **2000**, *407*, 651–654.
41. Spirtes, P.; Glymour, C.N.; Scheines, R. *Causation, prediction, and search*; MIT press, 2000.

42. Schaffter, T.; Marbach, D.; Floreano, D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics* **2011**, *27*, 2263–2270.
43. 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.
44. Meyer, P.E.; Lafitte, F.; Bontempi, G. minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC bioinformatics* **2008**, *9*, 461.
45. Olsen, C.; Meyer, P.E.; Bontempi, G. On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information. *EURASIP Journal on Bioinformatics and Systems Biology* **2008**, *2009*, 308959.
46. Meyer, P.; Marbach, D.; Roy, S.; Kellis, M. Information-Theoretic Inference of Gene Networks Using Backward Elimination. BioComp, 2010, pp. 700–705.