Methods 67 (2014) 294–303

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods

Jing Qin [a,d,1], Yaohua Hu [b,c,1], Feng Xu [a,d], Hari Krishna Yalamanchili [a,d], Junwen Wang [a,d,e,*]

[a] Department of Biochemistry, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong Special Administrative Region, China
[b] Department of Mathematics, Zhejiang University, Hangzhou, China
[c] Department of Applied Mathematics, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong Special Administrative Region, China
[d] Shenzhen Institute of Research & Innovation, The University of Hong Kong, Shenzhen, China
[e] Centre for Genomic Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong Special Administrative Region, China

ARTICLE INFO

Article history:
Received 21 November 2013
Revised 4 March 2014
Accepted 5 March 2014
Available online 17 March 2014

Keywords:
Gene regulatory networks
LASSO-type regularization methods
Integrative omics data
ChIP-seq/chip
Transcriptome

ABSTRACT

Inferring gene regulatory networks from gene expression data at whole genome level is still an arduous challenge, especially in higher organisms where the number of genes is large but the number of experimental samples is small. It is reported that the accuracy of current methods at genome scale significantly drops from *Escherichia coli* to *Saccharomyces cerevisiae* due to the increase in number of genes. This limits the applicability of current methods to more complex genomes, like human and mouse. Least absolute shrinkage and selection operator (LASSO) is widely used for gene regulatory network inference from gene expression profiles. However, the accuracy of LASSO on large genomes is not satisfactory. In this study, we apply two extended models of LASSO, L_0 and $L_{1/2}$ regularization models to infer gene regulatory network from both high-throughput gene expression data and transcription factor binding data in mouse embryonic stem cells (mESCs). We find that both the L_0 and $L_{1/2}$ regularization models significantly outperform LASSO in network inference. Incorporating interactions between transcription factors and their targets remarkably improved the prediction accuracy. Current study demonstrates the efficiency and applicability of these two models for gene regulatory network inference from integrative omics data in large genomes. The applications of the two models will facilitate biologists to study the gene regulation of higher model organisms in a genome-wide scale.

1. Introduction

Inferring gene regulatory networks from high-throughput genome-wide data is still a major challenge in systems biology. Transcriptome that is measured by microarray or RNA-seq describes the expression of all the genes of a genome. Various methods have been developed to infer gene regulatory networks from such transcriptome data (reviewed in [1]). Even though most of the methods perform very well on smaller genomes such as *Escherichia coli*, very few of them can accurately handle larger genomes, such as human and mouse. In large genomes, the complexity of gene regulatory system dramatically increases. Thousands of regulators, such as transcription factors (TFs), communicate in different ways to regulate tens of thousands of target genes in various tissues or biological processes. However, for a specific gene, only a few key TFs collaborate and control its expression change in a specific cell type or developmental stage. Thus, the gene regulatory network inference for such large genomes becomes a sparse optimization problem, which is to search a small number of key TFs from a pool of thousands of TFs for tens of thousands of targets based on the dependencies between the expression of TFs and the targets. The sparsity levels of the gene regulatory networks in large genomes are much higher than those in small genomes.

One of the most popular approaches is to estimate the pair-wise correlation between genes using metrics like Pearson's correlation coefficient, Spearman's correlation coefficient, mutual information, partial correlation coefficient or expression alignment, and then filter for causal relationships to infer TF-target pairs [2–6]. For example, ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) is proved to achieve low error rate and scaled up to mammalian system [2,7,8]. Another popular approach is to use the regression-based models to select TFs with target gene-specific sparse linear-regression [1,9,10]. In regression-based models, least

* Corresponding author at: Department of Biochemistry, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong Special Administrative Region, China. Fax: +852 2855 1254.
E-mail address: junwen@hku.hk (J. Wang).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.ymeth.2014.03.006
1046-2023/

absolute shrinkage and selection operator (LASSO) is most commonly used for gene regulatory network inference. In the field of optimization, LASSO is also called the L_1 regularization model [11]. Many efficient algorithms, such as ISTA (Iterative Soft Thresholding Algorithm), LAR (Least Angle Regression) and YALL1 (Your ALgorithms for L_1), have been developed for this model, and some of them are applied to large datasets [11–20]. Recently developed methods, such as NARROMI, take advantages of both correlation- and regression-based approaches and achieve improved accuracy [10].

However, both approaches suffer from several limitations when dealing with large genomes. Marbach et al. have shown that all of the 35 methods assessed, including both approaches, have much less precision for gene regulatory network inference in *Saccharomyces cerevisiae* than those in *E. coli*. Because the genome of *S. cerevisiae* is larger than that of *E. coli*, and its gene regulatory network is much more complex, only about 2.5% area under the precision-recall curve (AUPR) is achieved by these methods in *S. cerevisiae*, which is close to random. The correlation-based approaches need to calculate the correlation of all gene pairs, thus the computation cost increases exponentially with the number of genes. The sparsity level of the gene regulatory networks in large genomes is much higher than those in small genomes, hence false positives also increase remarkably when a similar correlation cutoff is used to predict the gene regulatory networks [2,7,21]. The accuracy of LASSO in large-scale problems with high sparsity is also reported to be not satisfactory [22–24].

To improve the performance, current methods require a large number of transcriptome profiles, usually at least one fold of the number of regulators [1]. However, in most biological studies, sample size is much smaller than the number of regulators due to high experimental cost. The limitation in sample size impedes performance of both approaches. In correlation-based methods, limited sample size makes the correlations between genes sensitive to noise, and thus high correlated gene pair needs not imply a true regulatory relationship. In large genomes, it is even more difficult to infer true regulatory links from a larger pool of highly correlated gene pairs with smaller sample size. In LASSO-based methods, when sample size is smaller than the number of regulators, multiple solutions are available, which makes it difficult to determine which solution is more biologically meaningful. To encounter the small sample size problem, heterogeneous datasets from different tissues, biological processes or experimental conditions are usually pooled together before the modeling, which increases prediction accuracy. However, the inferred network will lose its cell-type or condition specificity [1,25]. Further, heterogeneous datasets may weaken dependencies between the expression of TFs and their targets, since gene regulatory network topology varies among different biological processes, and one TF may regulate different gene sets in different cell states.

Chromatin immunoprecipitation (ChIP) coupled with high-throughput techniques, such as sequencing or microarray (ChIP-seq/chip, hereafter refer to as ChIP-X) data, which is also called cistrome, are also widely used to construct gene regulatory network in recent years [26–28]. However, TF binding sites detected by ChIP-X show only the genomic positions of the TF binding, but could not tell which gene is its target and whether and how the TF binding affects the transcription of its targets. Recently developed web server ChIP-Array that integrates both ChIP-X and transcriptome data to construct gene regulatory networks takes the advantages of both technologies and provides high accuracy, but it can be used for single TF-centered network only and requires the transcriptome data to be generated under the perturbation of the same TF as that of ChIP-X data. However, genome-wide gene regulatory networks contain multiple TFs, but only a limited number of TFs have both omics data.

In summary, although several methods have been proposed to infer genome-wide gene regulation networks from transcriptome profiles either alone [1,29], or in combination with predicted TF binding data [30], they are all limited by high computation cost and low accuracy in large genomes. To tackle these limitations, here we propose to integrate ChIP-X data with transcriptome profiles for gene regulatory network inference and use the L_p ($p < 1$) regularization model to improve the accuracy of LASSO, hereafter referred to as the L_1 regularization model. It is reported that it is able to achieve more sparse and accurate solutions by virtue of the L_p ($p < 1$) regularization model, even from small amount of samples [22,23,31]. However, the L_p ($p < 1$) regularization model suffers from its non-convexity and it is very difficult in general to design efficient algorithm for its solutions. Fortunately, the iterative hard thresholding algorithm [32] and iterative half thresholding algorithm [23] have been developed to solve the L_0 and $L_{1/2}$ regularization models respectively, but they have not been applied to gene regulatory network inference. Due to their low computation cost and fast convergence rate, we found that they are suitable for the gene regulatory network inference problem. Thus in this study, we apply the L_0 and $L_{1/2}$ regularization models to infer gene regulatory networks from ChIP-X and transcriptome data in mouse embryonic stem cells (mESCs). We compare their performance with the L_1 regularization model and find that ChIP-X data dramatically improved the accuracy of all three models, and the L_0 and $L_{1/2}$ regularization models significantly outperform the L_1 regularization model in the presence of ChIP-X data. The proposed models biologists to infer gene regulatory networks in higher model organisms using integrative omics data, efficiently. All the datasets and codes are available at: http://jjwanglab.org/LpRGNI.

2. Materials and Methods

2.1. L_p regularization models

Regulatory relationship between TFs and targets can be represented approximately by a linear system (Fig. 1A)

$$AX = B + \varepsilon$$

where $A \in R^{m\times r}$ denotes the expression matrix of candidate TFs, $B \in R^{m\times n}$ denotes the expression matrix of all target genes, $\varepsilon \in R^{m\times n}$ denotes an error matrix and $X \in R^{r\times n}$ denotes the regulation matrix that describes the regulatory relationship between all TFs and the targets, m denotes the number of samples, r denotes the number of factors and n denotes the number of target genes. In gene regulatory network inference, for each target gene j, we want to minimize the difference between $AX_{\cdot j}$ and $B_{\cdot j}$ with a small number of selected TFs, which is a sparse optimization problem described as

$$\min \|AX_{\cdot j} - B_{\cdot j}\|_2$$

$$\text{s.t.} \|X_{\cdot j}\|_0 \leqslant K,$$

where $\|\cdot\|_2$ denotes the Euclidean norm as $\|X_{\cdot j}\|_2 = \sqrt{\sum_{i=1}^{r} X_{ij}^2}$ and $\|X_{\cdot j}\|_0$ denotes the number of non-zero elements in $X_{\cdot j}$. The less $\|X_{\cdot j}\|_0$ means higher sparsity of $X_{\cdot j}$. It indicates how many TFs are found to regulate the target gene j.

For this problem, a popular and practical technique is to transform the sparse optimization problem into an unconstrained optimization problem, called a regularization problem (Fig. 1B). For example, given a gene j and its expression profile $B_{\cdot j}$, the L_0 regularization model is to minimize the difference between $AX_{\cdot j}$ and $B_{\cdot j}$, and maximize the sparsity of $X_{\cdot j}$:

$$\min_{X_{\cdot j}\in R^r} \|AX_{\cdot j} - B_{\cdot j}\|_2^2 + \lambda \|X_{\cdot j}\|_0,$$

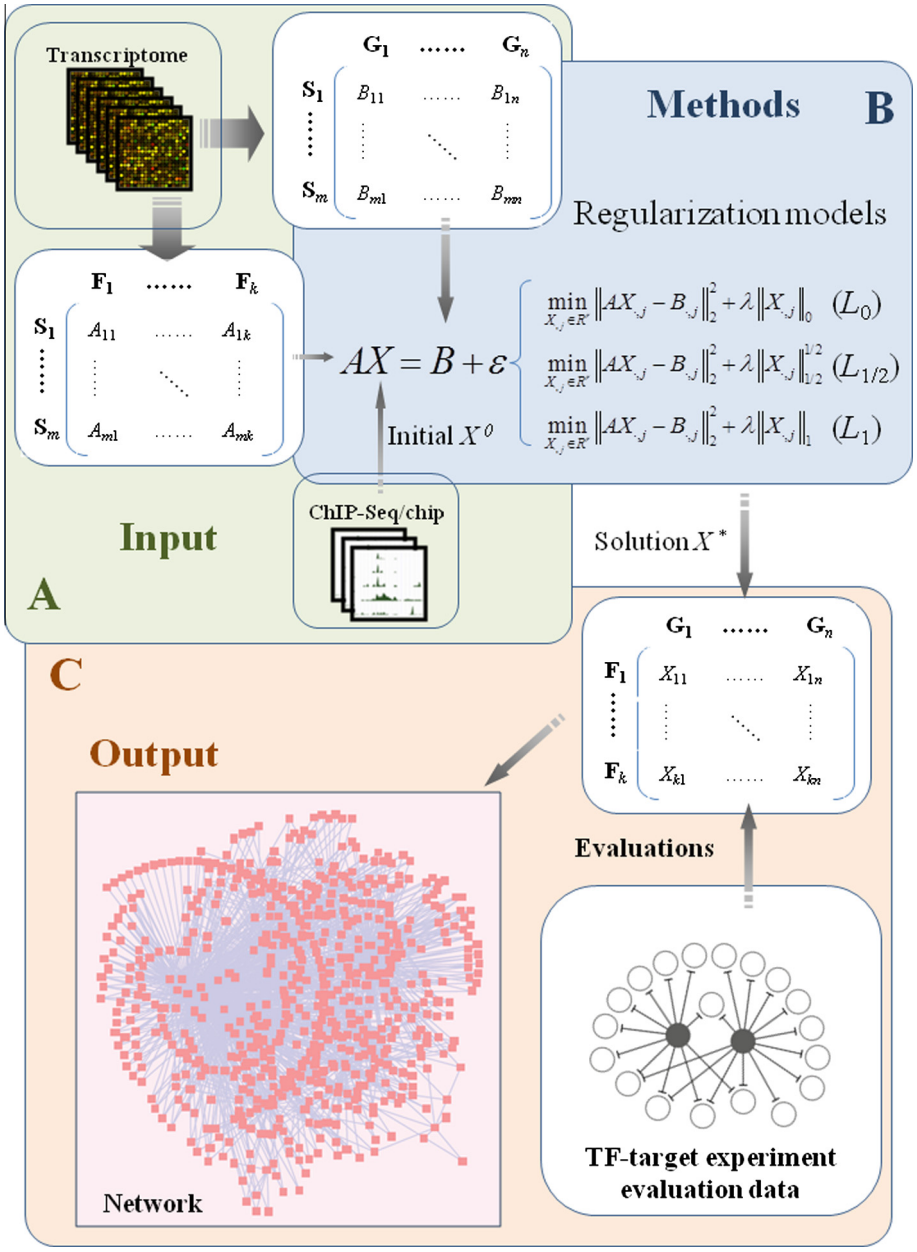

Fig. 1. Workflow of gene regulatory network inference with three regularization models. (A) matrix A and B containing the expression profiles of TFs and targets respectively are generated from transcriptome data, while ChIP-X identified TF-target interaction are converted into an initial X^0; (B) three regularization models are applied to solve the L_p (p = 1,1/2,0) regularization models; (C) output is a sparse matrix X^* that describes the TF-target relationships, which is evaluated by two sets of golden standards.

where $\lambda > 0$ is the regularization parameter, providing a tradeoff between accuracy and sparsity. Even though the L_0 regularization model is close to the original problem we want to solve, it is NP-hard to achieve a global optimal solution [33]. Thus, the L_1 regularization model (LASSO), a popular relaxation of the L_0 regularization model, is introduced to solve the following problem:

$$\min_{X_{.j}\in R^r} \|AX_{.j} - B_{.j}\|_2^2 + \lambda \|X_{.j}\|_1,$$

where $\|X_{.j}\|_1 = \sum_{i=1}^{r} |X_{ij}|$. However, in many practical applications, the solutions yielded from the L_1 regularization model are less sparse than those of the L_0 regularization model [22–24].

Recently, the $L_{1/2}$ regularization model is proposed and proved to perform better than the L_1 regularization model [23]. This model is described as

$$\min_{X_{.j}\in R^r} \|AX_{.j} - B_{.j}\|_2^2 + \lambda \|X_{.j}\|_{1/2}^{1/2},$$

where $\|X_{.j}\|_{1/2} = \left(\sum_{i=1}^{r} \sqrt{|X_{ij}|}\right)^2$. Neither L_0 nor $L_{1/2}$ regularization model has been used in gene regulatory network inference.

2.2. Algorithms

In this study, we apply the iterative thresholding algorithms to solve the L_p (p = 1,1/2,0) regularization models for gene regulatory network inference from omics data. The iterative thresholding algorithm is the most widely studied class of the first-order methods for the sparse optimization problem. It is convergent and of very low computational complexity [11,23,32]. Benefitting from its simple formulation and low storage requirement, it is very efficient and applicable even for the large-scale sparse optimization problem. In particular, the iterative soft thresholding algorithm is

introduced and developed to solve the L_1 regularization problem; the iterative hard thresholding algorithm is proposed to solve the L_0 regularization problem; and the iterative half thresholding algorithm is designed for the $L_{1/2}$ regularization problem. Briefly, in each iteration, these three algorithms firstly have a same gradient step

$$Z^k_{\cdot j} = X^k_{\cdot j} - 2\upsilon A^T(AX^k_{\cdot j} - B_{\cdot j}),$$

and then perform the thresholding operator respectively

$$(L_1)\quad X^{k+1}_{ij} = \begin{cases} Z^k_{ij} - sign(Z^k_{ij})\upsilon\lambda, & |Z^k_{ij}| > \upsilon\lambda \\ 0, & |Z^k_{ij}| \leqslant \upsilon\lambda \end{cases}$$

$$(L_0)\quad X^{k+1}_{ij} = \begin{cases} Z^k_{ij}, & |Z^k_{ij}| > \sqrt{2\upsilon\lambda} \\ 0, & |Z^k_{ij}| \leqslant \sqrt{2\upsilon\lambda} \end{cases},$$

$$(L_{1/2})\quad X^{k+1}_{ij} = \begin{cases} \frac{2}{3}Z^k_{ij}\left(1+\cos\left(\frac{2}{3}\pi - \frac{2}{3}\psi(Z^k_{ij})\right)\right), & |Z^k_{ij}| > \frac{3}{2}(\upsilon\lambda)^{2/3} \\ 0, & |Z^k_{ij}| \leqslant \frac{3}{2}(\upsilon\lambda)^{2/3} \end{cases}$$

$$\text{with } \psi(Z^k_{ij}) = \arccos\left(\frac{\upsilon\lambda}{4}\left(\frac{3}{|Z^k_{ij}|}\right)^{3/2}\right),$$

where the upper indexes of X and Z denote the number of iterations, υ denotes the stepsize, which we always choose as 1/2, and $sign(\cdot)$ denotes the sign function. The solutions of three algorithms achieved after 200 iterations, except those indicated, are used for further evaluation and comparison. For all the three algorithms, the regularization parameter λ is updated iteratively so as to keep the sparsity of $X^k_{\cdot j}$. For the details, one can refer to [23,32]. Since the number of TFs (the sparsity of $X_{\cdot j}$) that regulate a particular gene is usually unknown and biologists need to select a small number of TFs for the experimental verification, we make this parameter adjustable to the user. For further evaluation and comparison of three models, we test a series of factor numbers ($\|X_{\cdot j}\|_0$) from 1 to 100 (sparsity level ~0.1%-10%). In each test, we fix the same sparsity ($\|X_{\cdot j}\|_0$) for all three models. We assume that the TFs which are detected in a higher sparsity will be more important than those detected in a lower sparsity. Thus we score each factor according to the highest sparsity where it gets a non-zero index in the final solution (Section 2.5).

2.3. Data collections

Transcriptome data were downloaded from Gene Expression Omnibus (GEO). 245 experiments under perturbations in mESC were collected from three papers (Table 1) [34–36]. Each experiment produced transcriptome data with or without overexpression or knockdown of a gene, two replicates for control and two replicates for treatment. Gene expression fold changes between control and TF perturbation samples of 19978 genes in all experiments were log2 transformed and formed matrix B (Fig. 1A). Candidate regulators, including TFs, mediators, co-factors, chromatin modifiers and repressors, were collected from four TF databases, TRANSFAC, JASPAR, UniPROBE and TFCat, as well as literatures. Matrix A was made up of the expression profiles of 939 regulators (Fig. 1A). A literature-based golden standard TF-target pair set from biological studies (Fig. 1C), including 97 TF-target interactions between 23 TFs and 48 target genes (low-throughput golden standard), was downloaded from iScMiD (Integrated Stem Cell Molecular Interactions Database). Another golden standard mESC network was constructed from high-throughput ChIP-X and transcriptome data under TF perturbation (high-throughput golden standard). 28 TFs with evidences from both high-throughput ChIP-X and transcriptome data under perturbation were collected from literatures (Tables 2 and 3). TF binding sites were called with MACS [37] for ChIP-seq data and Cisgenome [38] for ChIP-chip data. Distance cutoff between a TF binding site and a potential target gene was set as 10 kbp. Differentially expressed genes under TF perturbation were defined as top 5% up-regulated and top 5% down-regulated genes, whose expression changes were significant with p-value <0.05. Single TF-centered network was constructed for each TF by ChIP-Array [39] with both high-throughput data. Direct target of all TFs were combined as a golden standard mESC network, which contains 40006 links between 13092 notes (Fig. 1C). Basically, each target in the network is evidenced by the cell-type specific binding sites on its promoter and the expression change in the perturbation experiment of the TF, which is generally accepted as a true target.

Table 1
Transcriptome data for gene regulatory network inference.

#Experiment	Perturbation	GEO accession	Pubmed ID
53	Overexpression	GSE16375	19796622
84	Overexpression	GSE31381	22355682
107	Knockdown	GSE26520	23462645

2.4. Integration of ChIP-X and transcriptome data

ChIP-X identifies *in vivo* active and cell-specific TF binding sites of a particular TF. A gene with an active TF binding site around its promoter is considered to be a potential target of the TF. Thus, ChIP-X data provides possible direct TF-target connections and may help regularization models to approximate the biologically meaningful solutions for the whole genome. Since matrix X describes the connections between TFs and targets, the TF-target connections defined by ChIP-X data were converted into an initial matrix X^0 (Fig. 1A and Table 2). Without ChIP-X data as a prior, the initial X^0 was artificially set as 0. When integrating ChIP-X data, if TF i has binding site around the gene j promoter within 10 kbp, except those indicated, the Pearson's correlation coefficient (PCC) between the expression profiles of TF i and gene j was calculated and assigned on X^0_{ij}. PCC can be positive or negative, representing the TF can activate or repress the target gene expression.

2.5. Evaluations

The area under the curve (AUC) of a receiver operating characteristic (ROC) curve is widely applied as an important index of the overall classification performance of an algorithm. We applied AUC to evaluate the performance of these three regularization models. For each pair of TF i and target j, if the X_{ij} is non-zero in the final solution matrix X, this TF is regarded as a potential regulator of the target. A series of factor numbers ($\|X_{\cdot j}\|_0$) from 1 to 100 were tested for each target. We assume that the TFs which are detected in a higher sparsity (smaller $\|X_{\cdot j}\|_0$) will be more important than those detected in a lower sparsity (larger $\|X_{\cdot j}\|_0$). Thus in the process of calculating the AUC, a score S_{ij} was applied as the predictor for TF i on target j:

$$S_{ij} = \begin{cases} \max(1/\|X_{\cdot j}\|_0), & X_{ij} \neq 0 \\ 0, & X_{ij} = 0 \end{cases}$$

And either high-throughput or low-throughput evaluation dataset was used as the golden standard. Furthermore, 1000 times' bootstrap was used to test the stability of AUCs. After the 1000 times' bootstrap, 1000 AUCs of each model was obtained. We then used Wilcoxon test to compare the performance of three models.

3. Results and discussions

Since, current methods are mostly designed for transcriptome data, we firstly inferred the gene regulatory network using current

Table 2
ChIP-X data for gene regulatory network inference and evaluation.

Factor	GEO accession	Pubmed ID	Factor	GEO accession	Pubmed ID
Cdx2	GSE16375	19796622	Rest	GSE26680	–
Ctr9	GSE14654	19345177	Rest	GSE27844	22297846
Ctr9	GSE20530	20434984	Sall4	GSE20551	20946988
Esrrb	GSE11431	18555785	Sfpi1	GSE11329	18358816
Jarid2	GSE19365	20075857	Smad1	GSE11431	18555785
Jarid2	GSE18776	20064375	Sox2	GSE11431	18555785
Kdm1a	GSE27844	22297846	Sox2	GSE11329	18358816
Klf4	GSE11431	18555785	Sox2	GSE11724	18692474
Myc	GSE11431	18555785	Stat3	GSE11431	18555785
Myc	GSE11329	18358816	Suz12	GSE11431	18555785
Mycn	GSE11431	18555785	Suz12	GSE11724	18692474
Nacc1	GSE11329	18358816	Suz12	GSE19365	20075857
Nanog	GSE11431	18555785	Suz12	GSE13084	18974828
Nanog	GSE11329	18358816	Suz12	GSE18776	20064375
Nanog	GSE11724	18692474	Tbx3	GSE19219	20139965
Nelfa	GSE20530	20434984	Tcf3	GSE11724	18692474
Nr0b1	GSE11329	18358816	Tcfcp2l1	GSE11431	18555785
Nr5a2	GSE19019	20096661	Trim28	GSE12283	19339689
Pou5f1	GSE11431	18555785	Wdr5	GSE22934	21477851
Pou5f1	GSE22934	21477851	Zfp281	GSE11329	18358816
Pou5f1	GSE11329	18358816	Zfp42	GSE11329	18358816
Pou5f1	GSE11724	18692474			

Table 3
Transcriptome data of TF perturbation for the construction of high-throughput golden standard network.

Factor	GEO accession	Pubmed ID	Factor	GEO accession	Pubmed ID
Cdx2	GSE12986	20081188	Pou5f1	GSE19588	21477851
Cdx2	GSE16375	19796622	Pou5f1	GSE26520	23462645
Ctr9	GSE12078	19345177	Pou5f1	GSE4189	16518401
Esrrb	GSE26520	23462645	Rest	GSE26520	23462645
Jarid2	GSE26520	23462645	Rest	GSE31381	22355682
Jarid2	GSE31381	22355682	Sall4	GSE16375	19796622
Kdm1a	GSE27844	22297846	Sall4	GSE26520	23462645
Klf4	GSE16375	19796622	Sfpi1	GSE16375	19796622
Klf4	GSE26520	23462645	Smad1	GSE16375	19796622
Myc	GSE16375	19796622	Sox2	GSE16375	19796622
Myc	GSE26520	23462645	Sox2	GSE26520	23462645
Mycn	GSE16375	19796622	Stat3	GSE16375	19796622
Mycn	GSE26520	23462645	Stat3	GSE26520	23462645
Nacc1	GSE26520	23462645	Suz12	GSE16375	19796622
Nanog	GSE16375	19796622	Suz12	GSE26520	23462645
Nanog	GSE26520	23462645	Tbx3	GSE26520	23462645
Nanog	GSE4189	16518401	Tbx3	GSE31381	22355682
Nelfa	GSE16375	19796622	Tcf3	GSE16375	19796622
Nelfa	GSE26520	23462645	Tcfcp2l1	GSE26520	23462645
Nr0b1	GSE16375	19796622	Tcfcp2l1	GSE31381	22355682
Nr0b1	GSE26520	23462645	Trim28	GSE26520	23462645
Nr5a2	GSE16375	19796622	Wdr5	GSE19588	21477851
Nr5a2	GSE26520	23462645	Zfp281	GSE26520	23462645
Pou5f1	GSE16375	19796622	Zfp42	GSE26520	23462645

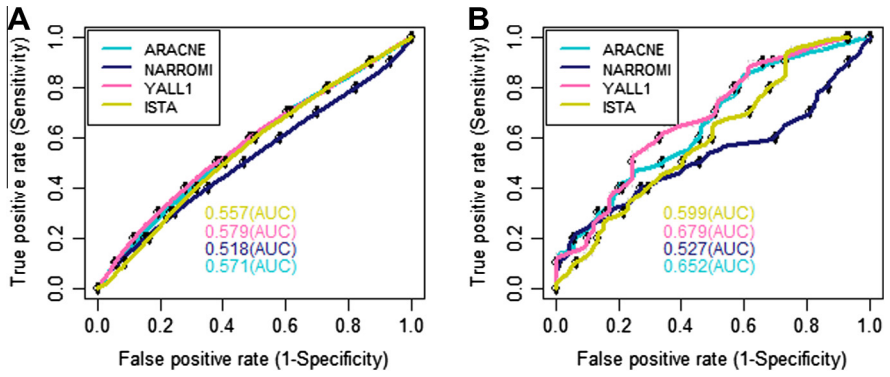

Fig. 2. ROC curves and AUCs of current methods on transcriptome data alone. Networks were inferred from transcriptome data alone. (A) evaluation with high-throughput golden standard; (B) evaluation with literature-based low-throughput golden standard.

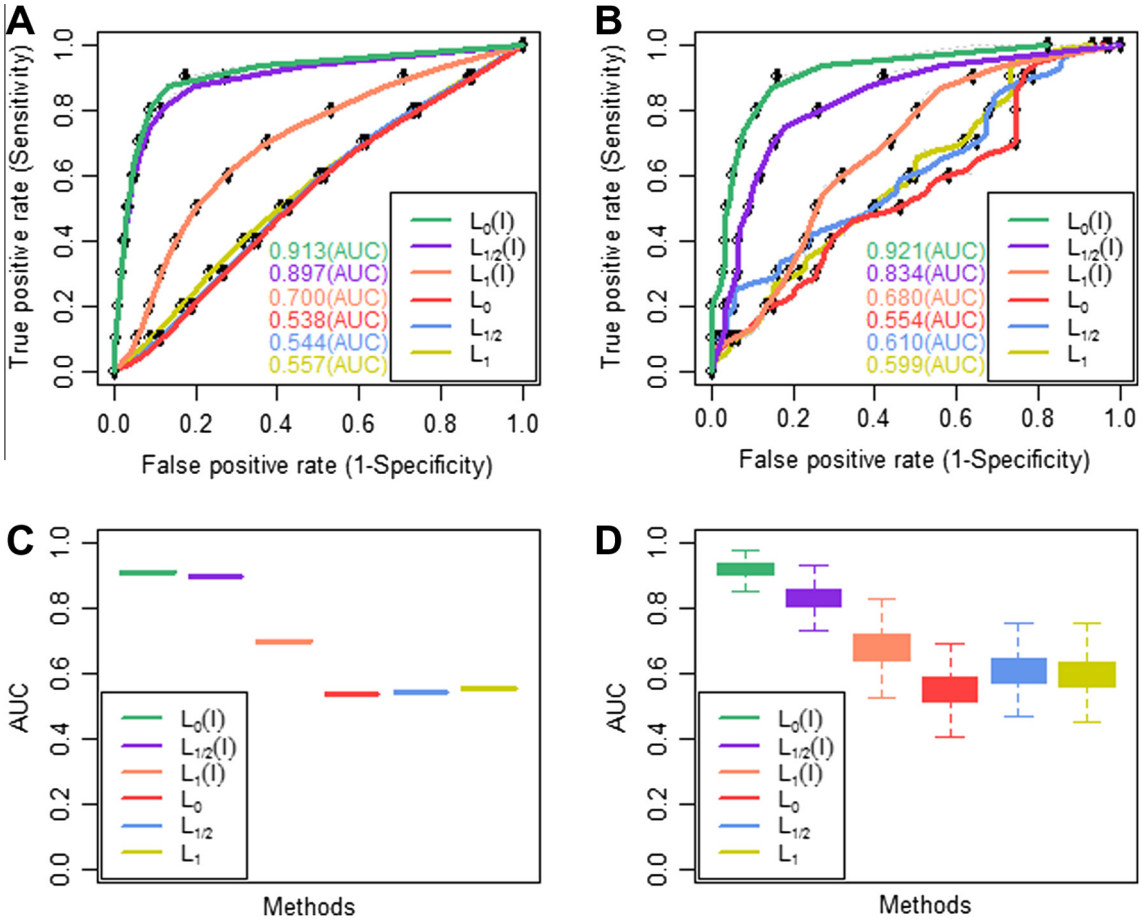

Fig. 3. ROC curves and AUCs of different methods on mESC gene regulatory network inference. (A) evaluation with high-throughput golden standard; (B) evaluation with literature-based low-throughput golden standard. L_0 (I), $L_{1/2}$ (I), L_1 (I): integrate ChIP-X and transcriptome data; L_0, $L_{1/2}$, L_1: transcriptome data alone. (C and D) 1000 times' bootstrap was used to test the stability of AUCs. Box plot of the 1000 AUCs shows the variability of the AUCs.

methods from transcriptome data alone. To retain the cell-type specificity of the inferred gene regulatory networks, we incorporated data from only mESC. Two TF-target datasets from high-throughput and low-throughput studies respectively are used as golden standards to evaluate the accuracy. As expected, their AUCs are all close to random on both evaluation data because the number of samples is much less than the number of regulators in our mESC data set (Fig. 2). Thus we incorporated ChIP-X data to improve the accuracy. We applied three L_p (p = 1,1/2,0) regularization models on the integration of ChIP-X and transcriptome data, and compared their performance. The L_1 regularization model used the same algorithm as ISTA, but using a different initial X^0 derived from ChIP-X data. Without the integration of ChIP-X data, the performance of the three models is similar to other methods and very poor when evaluated with either high-throughput (HGS) or low-throughput (LGS) golden standard (Fig. 3A and B). When ChIP-X data are integrated for network inference, the performance of all three regularization models dramatically improved, and the L_0 and $L_{1/2}$ regularization models significantly outperformed the L_1 regularization model (Fig. 3A and B, Table 4). The stabilities of the AUCs for all models are high when evaluated on high-throughput golden standard (Fig. 3C), while, due to the small number of

Table 4
Comparison of different LASSO-type regularization methods on mESC gene regulatory network inference. Pair-wise comparisons of AUCs were performed with Wilcoxon test on high-throughput golden standard (upper triangle, p-values in orange) or low-throughput golden standard (lower triangle, p-value in blue).

p-value	L_0 (I)	$L_{1/2}$ (I)	L_1 (I)	L_0	$L_{1/2}$	L_1
AUC (H)*	0.913	0.897	0.7	0.538	0.544	0.557
L_0 (I)		<1.000E-284	<1.000E-284	<1.000E-284	<1.000E-284	<1.000E-284
$L_{1/2}$ (I)	<1.000E-284		<1.000E-284	<1.000E-284	<1.000E-284	<1.000E-284
L_1 (I)	<1.000E-284	<1.000E-284		<1.000E-284	<1.000E-284	<1.000E-284
L_0	<1.000E-284	<1.000E-284	4.890E-257		7.44E-261	<1.000E-284
$L_{1/2}$	<1.000E-284	<1.000E-284	2.999E-127	9.304E-98		<1.000E-284
L_1	<1.000E-284	<1.000E-284	3.134E-154	3.306E-67	1.850E-05	
AUC (L)*	0.921	0.834	0.68	0.554	0.61	0.599

* AUC (HGS): AUC for high-throughput evaluation data (red); AUC (LGS): AUC for low-throughput evaluation data (purple).

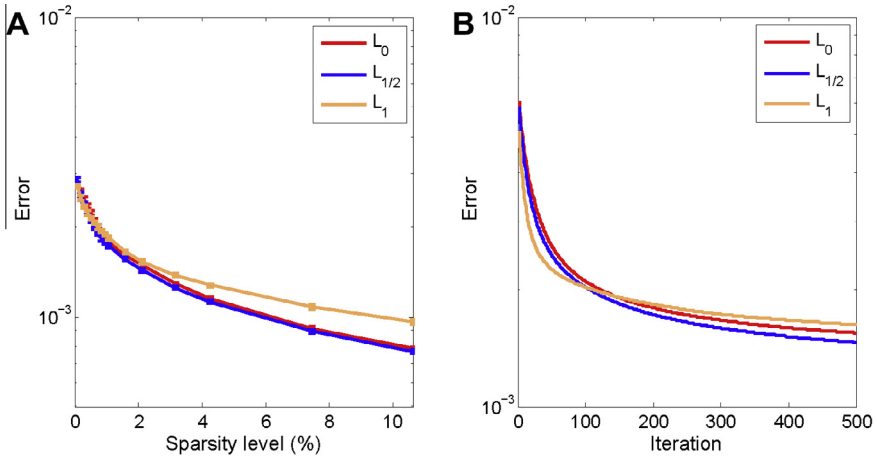

Fig. 4. Errors of three regularization models. A: errors against the sparsity level ($\|X_{x,j}\|_0$), which represents the number of TFs that are inferred as regulators of the target (solutions are achieved after 200 iterations); B: errors against the number of iterations (sparsity level is ~1% ($\|X_{x,j}\|_0 = 10$)). Each data point is the average of all genes. Error bars are the 95% confidence intervals. Errors were recorded when gene regulatory networks were inferred from integrative omics data.

known TF-target pairs, AUCs of different models calculated on low-throughput golden standard are less stable (Fig. 3D). But the L_0 and $L_{1/2}$ regularization models for integrative data still showed significantly better performance when evaluated on this data set (Table 4 and Fig. 3B). ROCs describe the information of false positive rate (FPR) and true positive rate (TPR) of all models. In biological studies, 0.05 is commonly used as the cutoff of FPR. At the FPR of 0.05, when evaluated with HGS, integration of ChIP-X data achieved TPRs of 0.637, 0.594 and 0.079 for the L_0, $L_{1/2}$ and L_1 regularization models respectively, and calculation with transcriptome data alone had TPRs of 0.031, 0.034 and 0.044 for the L_0, $L_{1/2}$ and L_1 regularization models respectively (Fig. 3A). The L_0, and $L_{1/2}$ regularization models achieved much higher sensitivity when integrating ChIP-X data, which meets biological researches' demand much better. Fig. 7 shows an example networks known to be active in mESC. A strict and identical cutoff (score $S_{i,j} \geqslant 0.1$) is used for all three models. The L_0 and $L_{1/2}$ regularization models reported much more true targets than the L_1 regularization model.

The advantages of the L_0 and $L_{1/2}$ regularization models are also demonstrated by the smaller error between $AX_{\cdot j}$ and $B_{\cdot j}$ when compared with the L_1 regularization model (Fig. 4). The errors are calculated along different sparsity levels (Fig. 4A) or iterations (Fig. 4B). When more TFs are selected, smaller error is achieved. To obtain solutions with same sparsity level, the L_1 regularization model showed a larger error than the L_0 and $L_{1/2}$ regularization models, which means that it gets less sparse if we fix the error allowance. This observation is consistent with several previous numerical experiments [22–24]. The error reduction along the iteration of the L_0 and $L_{1/2}$ regularization models were also faster than the L_1 regularization model after 100 iterations (Fig. 4B). Heatmap in Fig. 5B illustrates an example gene regulatory network inferred by the L_0 regularization model in mESC, in which only a small portion of TFs regulates most of the targets. The inferred TFs with more targets significantly overlapped with the TFs that were intensively reported to be associated with ESCs (Fig. 5**A**, p-value is 9.95E-10 in hypergeometric test).

The iterative thresholding algorithms we applied here only require low storage and computation cost [11,23,32]. The computational complexity of iterative hard and soft thresholding algorithms for the L_0 and L_1 regularization models, respectively, has been reported to be $O(kr\log m)$, where k is the number of iterations, r is the number of TFs and m is the number of targets [11,32,40]. The $L_{1/2}$ regularization model costs the similar computation time, since its computational complexity is similar to that of

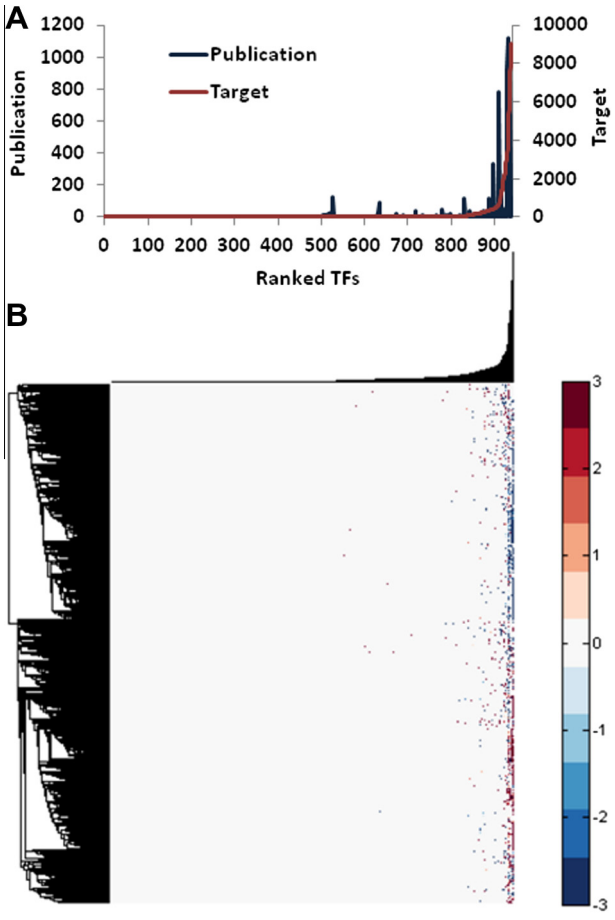

Fig. 5. Inferred TFs and their regulations on the targets. (A) TFs are ranked according to the number of targets in the inferred network (red line); and the number of publications, in which the TF and embryonic stem cell are co-occurred, is plotted for each ranked TF (blue line). (B) heatmap shows an example network with the regulatory connections between TFs and targets. Each row is a target gene, whereas each column is a TF. 10 TFs are detected for each target (sparsity level is ~1%, $\|X_{x,j}\|_0 = 10$). Red points are positive regulations, blue points are negative regulations, and white points indicate no regulation relationship.

the L_0 and L_1 regularization models (Fig. 6). Here, these three regularization models inferred the gene regulatory network with 939 TF, 19978 targets and 245 samples of mouse genome within one

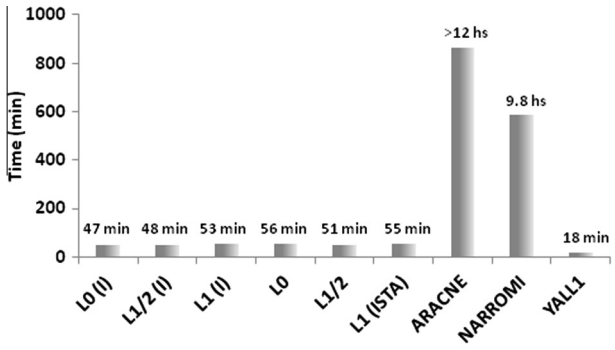

Fig. 6. Runtime of different methods. L_0 (I), $L_{1/2}$ (I), L_1 (I): integrate ChIP-X and transcriptome data; L_0, $L_{1/2}$, L_1: transcriptome data alone.

hour with one Intel Core i7 in personal laptop (2.00 GHz, 8.00 GB of RAM), slower than YALL1, but much faster than ARACNE and NARROMI (Fig. 6).

The ChIP-Array web server we developed previously can also integrate ChIP-X and transcriptome data to construct gene regulatory network for a single TF [39]. Even though it provides more confident network, it could be used only if both ChIP-X and transcriptome data of perturbation are available for the same TF. However, only a limited number of TFs have both omics data. Moreover, ChIP-Array constructs network for only one TF, although in most cases, target genes are regulated by multiple TFs. Unlike ChIP-Array, transcriptome data used by the L_p (p = 1,1/2,0) regularization models are not necessary to be the paired transcriptome data obtained from the perturbation of the same TF of ChIP-X experiment. Moreover, the L_p (p = 1,1/2,0) regularization models consider multiple TFs at the same time to infer more comprehensive gene regulatory network in a genome-wide scale. To assess how three models rely on the ChIP-X data, we tested their performance with different initial X^0 s. A series of initial X^0 s are made up of ChIP-X defined TF-target relationships with different distance cutoffs between a TF binding site and a potential target gene from 200 bp to 50 kbp, which are commonly used in biological studies (Table 5). After the TF binding sites are detected from ChIP-X data, a gene that locates closely to a TF binding site is usually considered as a potential target of the TF. However, proximity may not always indicate a true target, because the TF may bind on a distal enhancer, it may have other unknown function rather than transcription regulation, or sometimes multiple genes are close to a single TF binding site. Thus a large portion of potential targets defined by only ChIP-X data are false positives. Table 5 has shown that only 12.468-16.277% of potential targets defined by ChIP-X data alone are true targets that are verified by the TF knockdown/overexpression experiments (HGS). When a shorter distance cutoff is used, fewer genes will be defined as potential targets in the initial X^0, less false targets, but some true targets in a longer distance may be lost. When a longer distance cutoff is chosen, more true targets will be included, but false targets will be increased also. With different initial X^0 s, the L_0 and $L_{1/2}$ regularization models consistently outperformed the L_1 regularization model (Table 5). Even though high-throughput golden standard shares the ChIP-X data with the initial X^0 s, the large proportions of false targets in the initial X^0 s show that ChIP-X data alone could not infer the true targets accurately. PCC values between TF and target expression profiles in the initial X^0 s were used to indicate the possible regulatory effects of the TFs on the targets (activated or repressed), however, using PCC value in

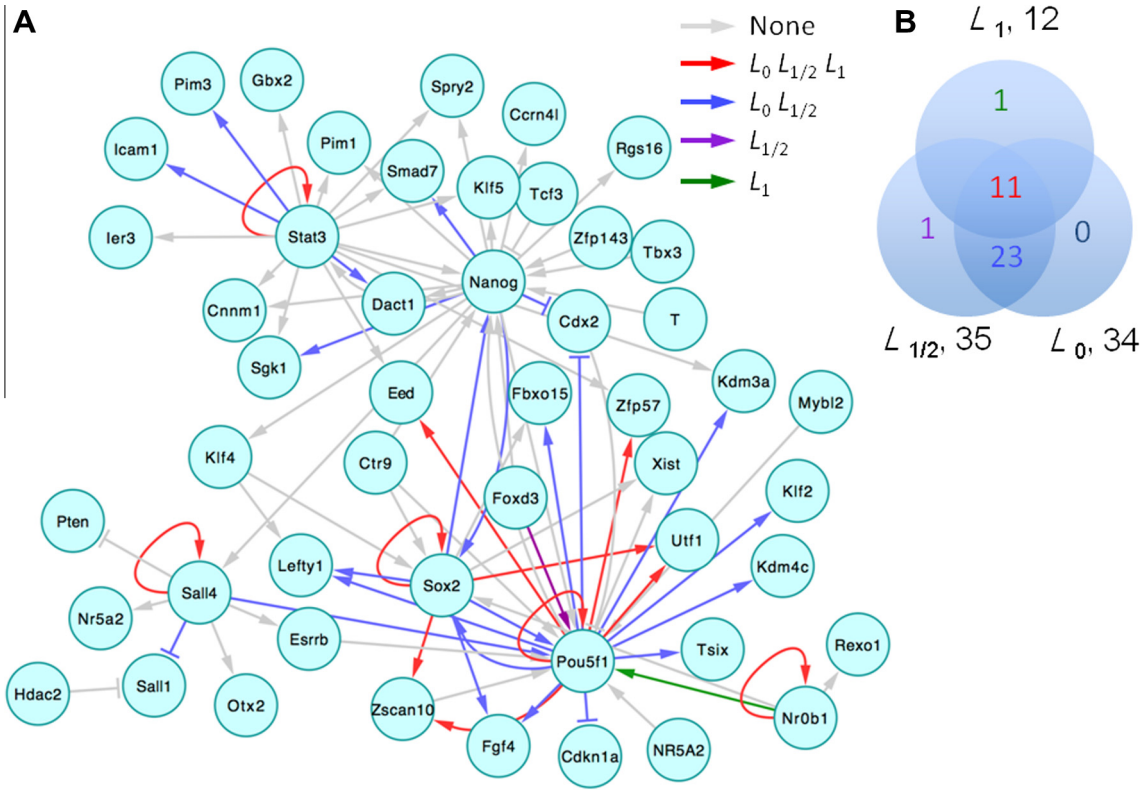

Fig. 7. Example networks known to be active in mESC and inferred by the L_p (p = 1,1/2,0) regularization models from integrative omics data. (A) Example networks in mESC. A strict and identical cutoff (score $S_{i,j} \geqslant 0.1$) is used for all three models. Grey arrows are regulations that are not identified by any of three models; red arrows are regulations that are identified by all three models; blue ones are identified by the L_0 and $L_{1/2}$ regularization models but not the L_1 regularization model; purple and green ones are only identified by the $L_{1/2}$ regularization model and the L_1 regularization model, respectively. (B) Venn diagram shows numbers of true regulations that identified by three models in the example network.

Table 5
Sensitivities of Lp (p = 1,1/2,0) regularization models to different initial X^0s. Initial X^0s were produced with different distance cutoffs for the potential targets defined by ChIP-X data.

Distance cutoff (bp)	200	500	1000	2000	5000	10000	20000	50000
AUC (HGS, L_0 (I))*	0.855	0.865	0.875	0.887	0.900	0.914	0.918	0.914
AUC (HGS, $L_{1/2}$ (I))	0.843	0.855	0.861	0.870	0.886	0.9	0.902	0.897
AUC (HGS, L_1 (I))	0.705	0.709	0.705	0.709	0.703	0.703	0.696	0.677
AUC (LGS, L_0 (I))*	0.737	0.860	0.810	0.933	0.876	0.907	0.908	0.872
AUC (LGS, $L_{1/2}$ (I))	0.812	0.797	0.800	0.868	0.865	0.895	0.883	0.864
AUC (LGS, L_1 (I))	0.716	0.757	0.697	0.794	0.679	0.742	0.679	0.670
AUC (HGS, PCC)	0.531	0.531	0.530	0.531	0.537	0.534	0.538	0.533
No. of true targets in initial X^0 (HGS)	15050	17049	18853	20816	24660	29271	32953	37183
No. of all potential targets in initial X^0	96083	107644	117281	128442	151500	180456	222186	298217
% of true targets in initial X^0 (HGS)	15.664	15.838	16.075	16.207	16.277	16.221	14.831	12.468

* AUC (HGS): AUC for high-throughput golden standard; AUC (LGS): AUC for low-throughput golden standard.

the initial X^0 s as the predictor to classify true targets in those potential targets defined by ChIP-X data resulted in very low AUCs (0.530–0.538, Table 5). Besides, least norm minimization can provide the solution having the smallest L_2 norm in all the possible solutions, which is commonly used as the initial point for the algorithms for solving the regularization problems of sparse optimization [22]. We created initial X^0 s via least norm minimization from transcriptome data, then performed the applied iterative thresholding algorithms for the L_p (p = 0,1/2,1) regularization models and evaluated results with the same methods. Three regularization models, L_0, $L_{1/2}$ and L_1, obtained AUCs of 0.529, 0.531 and 0.498 with high-throughput golden standard, and AUCs of 0.681, 0.672 and 0.629 with low-throughput golden standard, respectively. Consistent with [22], the L_0, $L_{1/2}$ regularization models are slightly better than L_1 regularization model. However, the results starting from L_2 norm solution are still much worse than those starting from the ChIP-X data. Thus, either ChIP-X or transcriptome data alone cannot achieve satisfactory accuracy, and the L_0 and $L_{1/2}$ regularization models did improve the performance of gene regulatory network inference from integrating ChIP-X and transcriptome data.

Recursive optimization of these three models iteratively moves the initial $X^0_{\cdot j}$ to the solution with minimum value of error between $AX_{\cdot j}$ and $B_{\cdot j}$. When sample size is smaller than the number of factors, the solution is not unique. Without prior knowledge, $X_{\cdot j}$ reaches one of the solutions that are close to the artificially assigned initial $X^0_{\cdot j}$, like 0. Thus, even though the models achieve a small error, which is mathematically profound; the non-uniqueness of the solution makes it biologically contradictory, i.e. it may not be the true biological solution we want. Integrating ChIP-X data provides partial knowledge of the gene regulatory network. The solutions that are close to the initial $X^0_{\cdot j}$ defined by ChIP-X are biologically more meaningful. Thus the performances of these three models improve after ChIP-X data are incorporated. Since the L_1 regularization model is convex, its local minimizer is also the global minimizer. Thus different initial X^0 s have less influence on the solution of L_1 regularization model (Table 5). However, the L_0 and $L_{1/2}$ regularization models are non-convex, the corresponding algorithms only converge to some local minimizers [41]. Here, we have shown that these local minimizers of the L_0 and $L_{1/2}$ regularization models obtained from integrative data are much closer to the biological solutions we expect than those of the L_1 regularization model (Table 5).

4. Conclusion

In this study, we apply the L_0 and $L_{1/2}$ regularization models to gene regulatory network inference from integrative omics data in large genome with a small number of samples. Integrating ChIP-X data with transcriptome profiles significantly improves the performance of network inference. Compared with the commonly used L_1 regularization model, the L_0 and $L_{1/2}$ regularization models have much higher accuracy for integrative omics data. We evaluated the inferred networks with both high-throughput and low-throughput golden standards. The L_0 and $L_{1/2}$ regularization models consistently outperformed the L_1 regularization model for integrative omics data. Besides, the algorithms we applied here are computationally efficient and can be executed by a personal computer within one hour. In summary, we have demonstrated that the L_0 and $L_{1/2}$ regularization models are applicable to gene regulatory network inference in biological researches that study higher organisms but generate only a small number of omics data, and facilitate biologists to analyze gene regulation at whole system level.

Funding

This work was supported by funding from the Research Grants Council, Hong Kong SAR, China (Grant number 781511M), National Natural Science Foundation of China, China (Grant numbers 91229105 and 11101186).

Reference

[1] D. Marbach, J.C. Costello, R. Kuffner, N.M. Vega, R.J. Prill, D.M. Camacho, K.R. Allison, M. Kellis, J.J. Collins, G. Stolovitzky, Nat. Methods 9 (2012) 796–804.
[2] A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, A. Califano, BMC Bioinf. 7 (Suppl 1) (2006) S7.
[3] A.J. Butte, I.S. Kohane, Pac. Symp. Biocomput. (2000) 418–429.
[4] A. de la Fuente, N. Bing, I. Hoeschele, P. Mendes, Bioinformatics 20 (2004) 3565–3574.
[5] X. Zhang, X.M. Zhao, K. He, L. Lu, Y. Cao, J. Liu, J.K. Hao, Z.P. Liu, L. Chen, Bioinformatics 28 (2012) 98–104.
[6] H.K. Yalamanchili, B. Yan, M.J. Li, J. Qin, Z. Zhao, F.Y. Chin, J. Wang, Bioinformatics 30 (2014) 377–383.
[7] K. Basso, A.A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, A. Califano, Nat. Genet. 37 (2005) 382–390.
[8] A.A. Margolin, K. Wang, W.K. Lim, M. Kustagi, I. Nemenman, A. Califano, Nat. Protoc. 1 (2006) 662–671.
[9] E.P. van Someren, B.L. Vaes, W.T. Steegenga, A.M. Sijbers, K.J. Dechering, M.J. Reinders, Bioinformatics 22 (2006) 477–484.
[10] X. Zhang, K. Liu, Z.P. Liu, B. Duval, J.M. Richer, X.M. Zhao, J.K. Hao, L. Chen, Bioinformatics 29 (2013) 106–113.
[11] I. Daubechies, M. Defrise, C. De Mol, Commun. Pur. Appl. Math. 57 (2004) 1413–1457.
[12] M.A.T. Figueiredo, R.D. Nowak, S.J. Wright, IEEE J.-Stsp 1 (2007) 586–597.
[13] J.F. Yang, Y. Zhang, Siam J. Sci. Comput. 33 (2011) 250–278.
[14] A.C. Haury, F. Mordelet, P. Vera-Licona, J.P. Vert, BMC Syst. Biol. 6 (2012) 145.
[15] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Ann. Stat. 32 (2004) 407–451.
[16] R. Tibshirani, J. R. Stat. Soc. B: Met. 58 (1996) 267–288.
[17] R. Tibshirani, J. R. Stat. Soc. B 73 (2011) 273–282.
[18] M.K.S. Yeung, J. Tegner, J.J. Collins, Proc. Natl. Acad. Sci. USA 99 (2002) 6163–6168.
[19] Y. Wang, T. Joshi, X.S. Zhang, D. Xu, L.N. Chen, Bioinformatics 22 (2006) 2413–2420.

[20] R. Bonneau, D.J. Reiss, P. Shannon, M. Facciotti, L. Hood, N.S. Baliga, V. Thorsson, Genome Biol. 7 (2006).
[21] V. Belcastro, V. Siciliano, F. Gregoretti, P. Mithbaokar, G. Dharmalingam, S. Berlingieri, F. Iorio, G. Oliva, R. Polishchuck, N. Brunetti-Pierri, D. di Bernardo, Nucleic Acids Res. 39 (2011) 8677–8688.
[22] R. Chartrand, V. Staneva, Inverse Prob. 24 (2008).
[23] Z.B. Xu, X.Y. Chang, F.M. Xu, H. Zhang, IEEE Trans. Neural Net. Lear 23 (2012) 1013–1027.
[24] T. Zhang, J. Mach. Learn Res. 11 (2010) 1081–1107.
[25] D. Marbach, S. Roy, F. Ay, P.E. Meyer, R. Candeias, T. Kahveci, C.A. Bristow, M. Kellis, Genome Res. 22 (2012) 1334–1349.
[26] C.G. de Boer, T.R. Hughes, Nucleic Acids Res. 40 (2012) D169–D179.
[27] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V.B. Vega, E. Wong, Y.L. Orlov, W. Zhang, J. Jiang, Y.H. Loh, H.C. Yeo, Z.X. Yeo, V. Narang, K.R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W.K. Sung, N.D. Clarke, C.L. Wei, H.H. Ng, Cell 133 (2008) 1106–1117.
[28] A. Marson, S.S. Levine, M.F. Cole, G.M. Frampton, T. Brambrink, S. Johnstone, M.G. Guenther, W.K. Johnston, M. Wernig, J. Newman, J.M. Calabrese, L.M. Dennis, T.L. Volkert, S. Gupta, J. Love, N. Hannett, P.A. Sharp, D.P. Bartel, R. Jaenisch, R.A. Young, Cell 134 (2008) 521–533.
[29] L. Zhang, X. Ju, Y. Cheng, X. Guo, T. Wen, BMC Syst. Biol. 5 (2011) 152.
[30] N. Novershtern, A. Regev, N. Friedman, Bioinformatics 27 (2011) i177–i185.
[31] R. Chartrand, IEEE Signal Process Lett. 14 (2007) 707–710.
[32] T. Blumensath, M.E. Davies, J. Fourier Anal. Appl. 14 (2008) 629–654.
[33] B.K. Natarajan, SIAM J. Comput. 24 (1995) 227–234.
[34] A. Nishiyama, L. Xin, A.A. Sharov, M. Thomas, G. Mowrer, E. Meyers, Y. Piao, S. Mehta, S. Yee, Y. Nakatake, C. Stagg, L. Sharova, L.S. Correa-Cerro, U. Bassey, H. Hoang, E. Kim, R. Tapnio, Y. Qian, D. Dudekula, M. Zalzman, M. Li, G. Falco, H.T. Yang, S.L. Lee, M. Monti, I. Stanghellini, M.N. Islam, R. Nagaraja, I. Goldberg, W. Wang, D.L. Longo, D. Schlessinger, M.S. Ko, Cell Stem Cell 5 (2009) 420–433.
[35] L.S. Correa-Cerro, Y. Piao, A.A. Sharov, A. Nishiyama, J.S. Cadet, H. Yu, L.V. Sharova, L. Xin, H.G. Hoang, M. Thomas, Y. Qian, D.B. Dudekula, E. Meyers, B.Y. Binder, G. Mowrer, U. Bassey, D.L. Longo, D. Schlessinger, M.S. Ko, Sci. Rep. 1 (2011) 167.
[36] A. Nishiyama, A.A. Sharov, Y. Piao, M. Amano, T. Amano, H.G. Hoang, B.Y. Binder, R. Tapnio, U. Bassey, J.N. Malinou, L.S. Correa-Cerro, H. Yu, L. Xin, E. Meyers, M. Zalzman, Y. Nakatake, C. Stagg, L. Sharova, Y. Qian, D. Dudekula, S. Sheer, J.S. Cadet, T. Hirata, H.T. Yang, I. Goldberg, M.K. Evans, D.L. Longo, D. Schlessinger, M.S. Ko, Sci. Rep. 3 (2013) 1390.
[37] J. Feng, T. Liu, B. Qin, Y. Zhang, X.S. Liu, Nat. Protoc. 7 (2012) 1728–1740.
[38] H. Ji, H. Jiang, W. Ma, W.H. Wong, Current protocols in bioinformatics/editoral board, Andreas D. Baxevanis, et al., Chapter 2 (2011) Unit2 13.
[39] J. Qin, M.J. Li, P. Wang, M.Q. Zhang, J. Wang, Nucleic Acids Res. 39 (2011) W430–W436.
[40] T. Blumensath, M.E. Davies, Appl. Comput. Harmon A 27 (2009) 265–274.
[41] Y.H. Hu, C. Li, X.Q. Yang, J. Mach. Learn. Res. (submitted for publication), http://www.acad.polyu.edu.hk/~mayangxq/GPA-SO.pdf.