BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 8 2014, pages 1190–1192
doi:10.1093/bioinformatics/btt761

Systems biology

Advance Access publication January 2, 2014

CMGRN: a web server for constructing multilevel gene regulatory networks using ChIP-seq and gene expression data

Daogang Guan[1], Jiaofang Shao[1], Youping Deng[2], Panwen Wang[3], Zhongying Zhao[1], Yan Liang[1], Junwen Wang[3,4,*] and Bin Yan[1,5,*]

[1]Department of Biology, Hong Kong Baptist University, Hong Kong, China, [2]Department of Internal Medicine and Biochemistry, Rush University Medical Center, Chicago, IL, USA, [3]Department of Biochemistry and HKU-SIRI, [4]Centre for Genomic Sciences, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China and [5]Key Laboratory of Acupuncture and Medicine Research of Minister of Education , Nanjing University of Chinese Medicine, Nanjing, Jiangsu 210046, China

Associate Editor: Gunnar Ratsch

ABSTRACT

Summary: ChIP-seq technology provides an accurate characterization of transcription or epigenetic factors binding on genomic sequences. With integration of such ChIP-based and other high-throughput information, it would be dedicated to dissecting cross-interactions among multilevel regulators, genes and biological functions. Here, we devised an integrative web server CMGRN (constructing multilevel gene regulatory networks), to unravel hierarchical interactive networks at different regulatory levels. The newly developed method used the Bayesian network modeling to infer causal interrelationships among transcription factors or epigenetic modifications by using ChIP-seq data. Moreover, it used Bayesian hierarchical model with Gibbs sampling to incorporate binding signals of these regulators and gene expression profile together for reconstructing gene regulatory networks. The example applications indicate that CMGRN provides an effective web-based framework that is able to integrate heterogeneous high-throughput data and to reveal hierarchical 'regulome' and the associated gene expression programs.

Availability: http://bioinfo.icts.hkbu.edu.hk/cmgrn; http://www.byan bioinfo.org/cmgrn.

Contact: yanbinai6017@gmail.com or junwen@hku.hk

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on July 3, 2013; revised on November 12, 2013; accepted on December 25, 2013

1 INTRODUCTION

The combinational performance of transcriptional and epigenetic factors is central regulatory mechanisms to control gene expression networks. A fundamental challenge in post-genome era is to dissect such complex regulation associations underlying biological functions. ChIP-seq technology allows high-fidelity mapping of different regulators, such as transcription factors (TFs) or epigenetic modifications to genomic locations, thus providing a basis for profiling transcriptional or epigenetic regulatory relationships. Accurate binding features of these factors on the genomic sequences can also be used to identify regulatory modules and reconstruct gene regulatory networks (GRNs). Various algorithms and computational methods have been proposed to investigate gene regulation promoted by cooperation of TFs and epigenetic alterations. Some online tools or servers were designed to build GRNs derived from ChIP-seq data, for instance, Cscan (Zambelli *et al.*, 2012) and ChIP-Array (Qin *et al.*, 2011). Although the existing web-based software can predict potential target genes of the multiple regulators, they are not able to discover hierarchical organizations formed by cross-interactions between the regulators and genes simultaneously. Currently, there is a lack of effective web resources to generate the topology of networks controlled by the interacting factors at transcriptional, post-transcriptional and epigenetic layers.

To provide an easy-to-use bioinformatics tool to interpret these ChIP-seq high-throughput data, we devised an integrative web server, constructing multilevel gene regulatory networks (CMGRN) to unveil interrelationships between TFs or epigenetic modifications and to robustly construct hierarchical regulatory network structures in biological complex systems. It is system wide and enables biologists to analyze the mixture information of ChIP-seq and gene expression levels, using standard data formatted with minimal need of bioinformatics skills.

2 METHODS

2.1 Overview

The schematic procedure and layout of CMGRN are illustrated in Figure 1. Our server provides a web-based framework with two main tasks: (i) inference of causal relationships between TFs or epigenetic modifications by using Bayesian network-based methodology. One input is required: gene counts of TFs or epigenetic modifications based on ChIP-seq BED data mapped to genomic regions. (ii) Construction of GRNs coordinated by the multilevel factors. Two types of input are required: regulatory signal of TFs, epigenetic modifications or microRNAs and gene expression data. The regulatory signal represents binding/modification activities of TFs or epigenetic factors identified by calling peaks of ChIP-seq or binding targets of microRNAs. Through the integrated analysis, CMGRN would establish hierarchical regulatory network organizations at multiple levels, which quantify the interactions among regulators and genes by using Bayesian hierarchical model and Gibbs sampling. Thus, CMGRN can generate two outputs following the two procedures and also provides an option for network reconstruction

*To whom correspondence should be addressed.

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

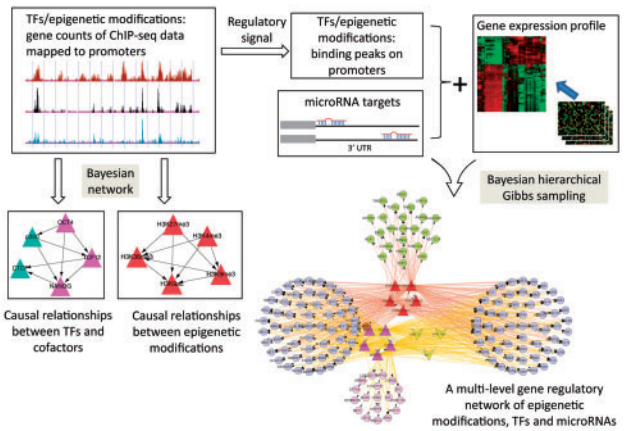

Fig. 1. An overview of data-integrated analysis and regulatory network construction. The triangle nodes represent TFs (pink), epigenetic regulators (red) and cofactors (blue), whereas the V-type nodes represent microRNAs (yellow). Target genes were indicated by using the circle nodes

depending on what are the types of input data (Fig. 1). The first one will obtain a causal network linking TFs or epigenetic factors if only ChIP-seq gene counts are provided. The second one will construct a final network exploring interactions of both regulator–regulator and regulator–gene if all three types of input are satisfied. The output tables and figures will be displayed via a user-friendly web interface and be easily downloaded. Usage of our service is simple and does not need user's knowledge on specific algorithm. The detailed information for the web tool implementation is explained in supplementary information and help page of the Web site.

2.2 Algorithms and statistical methods

2.2.1 Inference of causal relationships between regulators We used a Bayesian network-based model to infer causal interactions among regulators (TFs or epigenetic modifications) and examine how these regulators influence or associate with each other. The basic algorithm of Bayesian network structure searching for possibility of dependence between two nodes was described by Heckerman *et al.* (1995). There are two challenges to overcome, high posterior probability and the compelled edges in the network. In particular, compelled edge identification is critical because these edges indicate causal relationships when certain assumptions hold. A series of optimization is needed for such 'cause-effect' network structures with high posterior probability. In this study, we adopted *BD metric* (Bayesian metric with Dirichlet prior) strategy proposed by Heckerman *et al.* (1995) to find the high likelihood of posterior probability. First, we set a Bayesian network of regulators as follows: $B = \{D, \xi\}$, $D = \{(x_i, \ldots, x_n), E_k\}$, D is an acyclic-directed graph (DAG); x_i represents nodes (regulators) of the DAG, E_k is set of edges of DAG and ξ contains the set of conditional probability distributions that correspond to D. To find likelihood structures, the probability can be calculated by the following formula:

$$\log p(D, B_s^h|\xi) = \sum_{i=1}^{n} \log s(x_i|\pi_i) = \sum_{i=1}^{n} w(x_i, \pi_i) + \sum_{i=1}^{n} \log s(x_i|\theta)$$

where π_i is the (possibly) null parent of x_i. For edge $x_j \rightarrow x_i$, the strategy is to associate with a weight $w(x_i, x_j) = \log s(x_i|x_j) - \log s(x_i|\theta)$, θ is a node set without edges. A weight is defined for edge x_i to x_j as $\omega(x_i, x_j) = logG(x_i|x_j) - logG(x_i|\varphi)$. To search for the best network structure, we used maximal value for $\sum_{i=1}^{n} w(x_i, \pi_i)$. The aforementioned algorithm uses greedy strategy, so the next step is to identify compelled edges as potential network structures. Chickering developed mathematical algorithm for identifying compelled edges between two nodes to determine their causal relationship (Chickering, 1995). Yu *et al.* (2008) applied a similar methodology to predict causal relationships between histone modifications and gene expression in human T cells.

The ChIP-seq reads in BED format were mapped to genomic regions and were then transferred to gene counts or signals of TFs or epigenetic modifications that could be detected. CMGRN used the gene count of ChIP-seq as input for inferring the causal relationships between regulators. First, the input gene counts will be discretized by using k-means++ algorithm. The k-means method is a widely used clustering technique that seeks to minimize the average and squared distance between points in the same cluster. Although its simplicity and speed are appealing in practice, it offers no accuracy guarantees for discretization. One reason is that k-means usually chooses initial k centers in an arbitrary way. If the centers are not properly taken, the resulting clusters will be also arbitrary and thus leading to generation of unstable 'cause-effect' relationships. In this study, we adopted a k-means++ method, which proposes a specific approach to choose centers like 'D^2 *weighting*' (Arthur and Vassilvitskii, 2007). The main step of k-means++ is as follows: (i) take one center k_1, chosen uniformly at random from χ; (ii) take a new center k_i, choosing $x \in \chi$ with probability $\frac{D(x)^2}{\sum_{x \in \chi} D(x)^2}$; (iii) repeat the second step until we take k centers altogether; and (iv) do the k-means algorithm. The strategy used by the k-means++ is able to improve both speed and accuracy of k-means. We have also demonstrated that the k-means++ method can increase the stability of the inferred causal relationships (see Supplementary Fig. S1).

2.2.2 GRNs construction We used Bayesian hierarchical model with Gibbs sampling implementation to identify correct sets of regulator–gene interactions that explore underlying network structures. This method will combine two types of data, regulatory signals of multifactors and gene expression profile. The ChIP-seq binding peaks of TFs and epigenetic modifications on the promoters were used as their regulatory signals. The regulatory signals of microRNAs were expressed by their binding targets based on both known and predicted sites on 3′-untranslated regions. The main aim of this method is to incorporate the heterogeneous data, to capture essential regulatory features behind high-throughput data and to provide a basis for reconstructing GRNs. An integrated model COGRIM using the similar algorithm showed its capacity of identifying potential regulatory interactions between TFs and genes (Chen *et al.*, 2007). Previously, we successfully used this statistical model to identify target regulons of oncogenic TF family of NF-κB in different cell types of head and neck cancer (Yan *et al.*, 2008). Here, we assign the gene expression profile and binding data of regulators to Bayesian hierarchical with an extensible linear model as following:

$$e_{it} = \alpha_i + \sum_{j=1}^{j} \beta_j Pr_{ij} f_{jt} + \delta_{it}$$

$$p(Pr_{ij}|\emptyset, e_{it}, b_{ij}) \propto \prod_{t=1}^{t} p(e_{it}|Pr_{ij}, \emptyset) \cdot \prod_{j=1}^{j} p(Pr_{ij}|b_{ij})$$

Where $\delta_{it} \sim Normal(0, \mu^2)$, e_{it} is the log expression of gene i ($i = 1 \ldots N$) in experiment t ($t = 1 \ldots T$). The α_i is baseline expression for gene i in absence of known regulators. The f_{jt} is the regulation activity estimated by log expression of TF gene j ($j = 1 \ldots J$) in experiment t. Pr_{ij} refers to the probability of gene i regulated by regulator j. Prior $Pr_{ij} \sim Bernoulli(h(b_{ij}))$ where b_{ij} means regulator j physically binds in proximity to gene i. In this study, this parameter can be considered as the regulatory signal from ChIP-seq binding peaks, or targets contain binding sites of microRNAs. $\emptyset$ represents parameter set $(\alpha_i, \beta_j, \mu^2)$. Both Pr_{ij} and $\emptyset$ can be estimated one set given the other by Gibbs

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

sampling, one of Markov chain Monte Carlo algorithm that is able to iteratively sample new values for each set of unknown parameters conditional on the current values of all other parameters. This process can be described as follows: estimating $\emptyset$ given Pr_{ij}, e_{it} and b_{ij}, estimating Pr_{ij} given $\emptyset$, e_{it} and b_{ij}. To test the performance of CMGRN, we compared it with other three methods by using four different types of cancer gene expression data. The result showed that our method could improve the reliability of TF target identification from the cancer datasets (see Supplementary Fig. S2).

2.3 Implementation

CMGRN was implemented with PHP and pascal.

3 SAMPLES TEST

We applied CMGRN to address the complex structures of regulatory networks underlying mouse embryonic stem cell pluripotency and differentiation. Our analysis provides computational proof that three pluripotent TFs (Oct4, Sox2 and Nanog), active histone methylations and microRNAs form a hierarchical 'regulome' that downregulates gene expression related to transcriptional processes and upregulates developmental genes facilitating differentiation. The hierarchical regulatory feature also works in an estrogen-stimulated breast cancer cell line and reveals that estrogen receptor α is a driving factor that links to other TFs and cofactors that control estrogen-associated gene expression programs (see Supplementary Figs S3 and S4).

Funding: Grant of Science Faculty of Hong Kong Baptist University (Grant no: FRG2/12-13/066), National Nature Science Foundation of China (Grant no: 91229105), Hong Kong Research Grants Council (Grant no: 781511M) and Collaborative Research Fund of Hong Kong Research Grants Council (Grant no: HKBU5/CRF/11G).

Conflict of Interest: none declared.

REFERENCES

Arthur,D. and Vassilvitskii,S. (2007) kmeans++: the advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, pp. 1027–1035.

Chen,G. *et al.* (2007) Clustering of genes into regulons using integrated modeling-COGRIM. *Genome Biol.*, **8**, R4.

Chickering,D.M. (1995) A transformational characterization of equivalent bayesian network structures. In: Besnard,P. and Hanks,S. (eds) *Proceedings of 11th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 87–98.

Heckerman,D. *et al.* (1995) Learning Bayesian networks: the combination of knowledge and statistical data. *Mach. Learn.*, **20**, 197–243.

Qin,J. *et al.* (2011) ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor. *Nucleic Acids Res.*, **39**, W430–W436.

Yan,B. *et al.* (2008) Systems biology-defined NF-kappaB regulons, interacting signal pathways and networks are implicated in the malignant phenotype of head and neck cancer cell lines differing in p53 status. *Genome Biol.*, **9**, R53.

Yu,H. *et al.* (2008) Inferring causal relationships among different histone modifications and gene expression. *Genome Res.*, **18**, 1314–1324.

Zambelli,F. *et al.* (2012) Cscan: finding common regulators of a set of genes by using a collection of genome-wide ChIP-seq datasets. *Nucleic Acids Res.*, **40**, W510–W515.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016