This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2866836, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Integration of Multi-omics Data for Gene Regulatory Network Inference and Application to Breast Cancer

Lin Yuan, Le-Hang Guo, Chang-An Yuan, You-Hua Zhang, Kyungsook Han, Asoke K. Nandi, Barry Honig, and De-Shuang Huang

Abstract—Underlying a cancer phenotype is a specific gene regulatory network that represents the complex regulatory relationships between genes. However, it remains a challenge to find cancer-related gene regulatory network because of insufficient sample sizes and complex regulatory mechanisms in which gene is influenced by not only other genes but also other biological factors. With the development of high-throughput technologies and the unprecedented wealth of multi-omics data give us a new opportunity to design machine learning method to investigate underlying gene regulatory network. In this paper, we propose an approach, which use biweight midcorrelation to measure the correlation between factors and make use of nonconvex penalty based sparse regression for gene regulatory network inference (BMNPGRN). BMNCGRN incorporates multi-omics data (including DNA methylation and copy number variation) and their interactions in gene regulatory network model. The experimental results on synthetic datasets show that BMNPGRN outperforms popular and state-of-the-art methods (including DCGRN, ARACNE and CLR) under false positive control. Furthermore, we applied BMNPGRN on breast cancer (BRCA) data from The Cancer Genome Atlas database and provided gene regulatory network.

Index Terms—biweight midcorrelation, differential correlation, nonconvex penalty, gene regulatory network, stability selection

1 INTRODUCTION

Gene regulatory network (GRN) is a biological process that represents the complex regulatory relationships between genes. Research on the structures and dynamics of GRNs provides an important insights into the mechanisms of complex diseases (e.g. breast cancer or brain tumors). It remains, however, a challenge to understand gene regulatory network because of complex regulatory mechanisms in which gene is influenced by not only other genes but also other biological factors such as DNA methylation (DM) and copy number variation (CNV) of genes. It is essential to research GRN using multi-omics data. Meanwhile, the rapid development in high-throughput sequencing (HTS) has led to detailed clinical records and heterogeneous data for more than 1000 samples of breast cancer.

DNA methylation, one of the best-known epigenetic marker, plays an important role in modifying gene expression level, and it is one of heritable changes in gene expression, which are not due to any alteration in the DNA sequence . DNA methylation controls gene expression by means of changes of DNA stability, chromatin structure and DNA-protein interactions. In addition, DNA methylation generally inhibits gene expression [1, 2]. However, DNA methylation-gene regulatory mechanism is still an unsolved problem. CNV is a gene of more than 1 kilo-base in size. Specifically, CNV is a kind of structural variation and a type of duplication or deletion event that affects a considerable number of base pairs. Many researches have shown that CNV is an important biological factor that affects gene expression. Recently, biological experiments have shown that GRN research that integrate DNA methylation data and CNV data have a better performance than methods which use gene expression data only [3-6].

In the beginning, experiments show that the expression levels of function-related genes are correlated [7]. Gene correlation measure methods were applied to gene regulatory network analysis [8]. One of the well-known correlation measures is the Pearson correlation measure [9, 10]. Meanwhile, gene co-expression network analysis is one of the most important gene regulatory network analyses. Weighted gene co-expression network analysis (WGCNA) is a representative method of the correlation-based methods [11]. For different characteristics of data, methods which based on different correlation coefficients were proposed. For example, mutual information-based gene network analysis, maximum information

- *L. Yuan and D.S. Huang are with the Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Caoan Road 4800, Shanghai 201804, China. E-mail: yuanlindc @126.com, dshuang @tongji.edu.cn.*
- *L.H. Guo is with the Department of Medical Ultrasound, Shanghai Tenth People's Hospital, Ultrasound Research and Education Institute, Tongji University School of Medicine, Yanchang Middle Road 301, Shanghai 200072, China. E-mail: gopp1314@hotmail.com.*
- *C.A. Yuan is Science Computing and Intelligent Information Processing of GuangXi Higher Education Key Laboratory, Guangxi Teachers Education University, Nanning, Guangxi, 530001, China, E-mail: yca@gxtc.edu.cn*
- *K. Han, School of Computer Science and Engineering Inha University Incheon South Korea. E-mail: khan@inha.ac.kr*
- *Y.H. Zhang is with School of Information and Computer, Anhui Agricultural University, Changjiang West Road 130, Hefei, Anhui, China. E-mail: zhangyh@ahau.edu.cn*
- *A.S. Nandi is with Department of Electronic and Computer Engineering, Brunel University London, Uxbridge, UB8 3PH, United Kingdom. Email: Asoke.Nandi@brunel.ac.uk*
- *B. Honig is with Center for Computational Biology and Bioinformatics, 1130 St. Nicholas Avenue, Room 815, New York, NY, 10032, USA. Email: bh6@columbia.edu.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2866836, IEEE/ACM Transactions on Computational Biology and Bioinformatics

correlation-based gene network analysis, cosine similarity-based gene network analysis, spearman rank correlation-based gene network analysis and conditional mutual information-based gene network analysis. Compared with the gene co-expression network analysis, gene differential co-expression analysis based on correlation measure can discover tiny but important biomolecule changes through identifying subtle changes in gene expression levels from case-control group [12]. For example, gene regulatory network analysis based on GO term and identifying differentially co-expression genes and links from gene expression microarray data (DCGL) outperforms than WGCNA in terms of sensitivity and specificity and real data sets result . However, the relationship between genes is still unclear due to gene interactions are mediated by other biological factors. Meanwhile, correlation measurement-based method use an undirected graph, which cannot explain the regulatory relationship between genes [7, 13, 14].

Bayesian-based methods make it possible to infer the regulatory relationship of genes based on the directed acyclic graph (DAG) [15]. Many studies reported directed gene regulation graphs [16]. With the development of Bayesian network, dynamic Bayesian network was used to model time varying gene regulatory network. Moreover, Bayesian-based methods can deal with data missing problem and noise [17]. Bayesian-based methods can also efficiently incorporate prior biological knowledge such as structural information in the model [16]. However, those Bayesian-based methods are not suitable for large-scale biological data sets because its runtime will increase exponentially as the number of genes increases [18-20].

Regression-based methods decomposes the complex gene regulatory network problem into P (number of genes) regression problems. In practice, regression-based methods need to solve small sample size and high-dimensional problems, thus regression-based gene regulatory network inference models contain regularization and are sparse models. Regularization regression models have often adapted LASSO for selecting significant network-related biological factors (e.g. genes, DNA methylation sites, CNVs) [21]. Regularization regression models can ensure efficiency even with large-scale biological data sets. Meanwhile, Regularization regression models are high-ranking methods in the HPN-DREAM network inference challenge. Recently, several regression-based researches focus on finding gene regulatory network by combining heterogeneous data [22]. Experimental results on synthetic data and real data show that methods can improve the accuracy of gene regulatory network inference by combining a variety of biological factor information [23, 24].

DCGRN (DNA methylation and CNV Gene Regulatory Network) is a representative method of the regularization regression models [22]. DCGRN proposed an integrative gene regulatory network inference method by using gene expression, DNA methylation and CNV [22]. Such method assumes the number of gene-related DNA methylation is one [25]. However, biological experiments have shown that multiple methylation sites affect one gene [26]. A function module formed by the interaction of multiple methylation sites affects gene expression. [22] have applied LASSO to select biological factors of gene regulatory network. LASSO may not be applied directly to multi-omics data since LASSO ignores prior information present in wealth of multi-omics data. In addition, although the LASSO achieved good results [21, 22], L1-regularization is a convex relaxation formulation of L0-regularization. L1-regularization often leads to suboptimal solution because it is not a good approximation to L0-regularization. Nonconvex penalty functions such SCAD (Smoothly Clipped Absolute Deviation) [27] and MCP (Minimax Concave Penalty) [28] were applied to sparse problems and achieved better performance than traditional methods.

In this paper, we propose an approach, which use Biweight Midcorrelation to measure the correlation between factors and make use of Nonconvex Penalty based sparse regression for Gene Regulatory Network inference (BMNPGRN). BMNPGRN integrates heterogeneous multi-omics data to infer gene regulatory network with gene expression data, DNA methylation data and copy number variation data. In order to infer gene regulatory network. Firstly, we combine biweight midcorrelation coefficient algorithm, which is an efficient algorithm for computing correlation coefficient, with 'differential correlation strategy' to learn associations among DNA methylation sites. Then, nonconvex penalty based sparse regression is used to find gene-related biological factors, and the parameter of method is determined by cross-validation. Meanwhile, nonconvex penalty based sparse regression is used under stability selection which can control false positives effectively [29]. Finally, BMNPGRN identifies gene regulatory network based on the probabilities of biological factors (i.e. gene, DNA methylation sites, and CNV).

Our proposed approach BMNPGRN has advantages over existing gene regulatory network inference methods. Firstly, BMNPGRN can find more DNA methylation sites which are associated with gene regulatory network. Such method provide deeper insight into gene regulation mechanism. Secondly, BMNPGRN can effectively control false positives using stability selection strategy. Furthermore, BMNPGRN can more accurately find biological factors using nonconvex penalty based sparse regression. Finally, to the best of our knowledge, BMNPGRN is the first method which is applied to breast cancer data obtained by high-throughput sequencing technology.

In our experiments, we first compared the receiver operating characteristic (ROC) performance of BMNPGRN with well-known gene regulatory network inference methods (DCGRN, ARACNE and CLR) [22, 30, 31] in two kinds of synthetic data sets, experiment results show that BMNPGRN can significantly improve

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2866836, IEEE/ACM Transactions on Computational Biology and Bioinformatics

the performance of detecting gene regulatory network under false positive control. The mean and variance of AUC values obtained from multiple experiment results show that the stability of BMNPGRN is better than state-of-the-art methods of biological network inference. We then applied BMNPGRN on breast cancer data from The Cancer Genome Atlas (TCGA) database and identified several cancer-related gene regulatory network.

2 METHODS

Before introducing our method, we summarize the notations used in this article. Matrices are denoted by boldface uppercase, vectors are denoted by boldface lowercase, and scalars are denoted by lowercase letters. We denote the gene expression matrix by $\mathbf{X} \in \mathbf{R}^{N\times P}$, N represents number of samples and P represents number of genes, $\mathbf{x}_j$ represents the j-th column of gene expression matrix, $\mathbf{x}^i$ represents the i-th row of gene expression matrix, and x_j^i represents the (i, j) entry of matrix. Meanwhile, DNA methylation matrix is denoted by $\mathbf{D} \in \mathbf{R}^{N\times Q}$ with N samples and Q DNA methylation sites, and copy number variation matrix is denoted by $\mathbf{C} \in \mathbf{R}^{N\times P}$ with P CNVs.

We show how to discover gene regulatory network, next. In brief, we first present a new scheme to find gene-related DNA methylation sites and, then, select significant gene regulation-related biological factors (including genes, DNA methylation sites) using nonconvex penalty based sparse regression. In addition, we use nonconvex penalty based sparse regression under stability selection which can control false positives effectively.

2.1 A NEW SCHEME TO FIND GENE-RELATED DNA METHYLATION SITES

Given the datasets containing gene expression data, copy number variation data, and DNA methylation data, we use the heterogeneous multi-omics data to infer gene regulatory network. In general, a dataset include both patient and normal samples. The correlation coefficient of functionally related DNA methylation sites varies greatly from normal sample to patient sample. Based on the characteristics of biological data, we propose a new scheme to find gene-related DNA methylation sites. First we need to calculate the correlation coefficient between expression vectors of DNA methylation sites. Researchers have proposed many methods to measure the correlation coefficient between two variables. Pearson correlation coefficient is a representative method of correlation coefficient approaches [9, 10]. However, Pearson correlation coefficient is sensitive to outliers. Biweight midcorrelation is considered to be a good alternative to Pearson correlation coefficient since it is more robust to outliers. To calculate correlation coefficient between DNA methylation sites, we use biweight midcorrelation, which shall be described in the next section. On the basis of correlation coefficients from patient samples and normal samples, we adopt a differential correlation strategy. This strategy is similar to the case-control sample correlation coefficient measurement step followed by threshold-based variables selection [32-34].

2.1.1 BIWEIGHT MIDCORRELATION COEFFICIENT

In order to introduce the biweight midcorrelation coefficient (BIMC) of two numeric vectors $\mathbf{x}=(x_1,\ldots,x_n)$ and $\mathbf{y}=(y_1,\ldots,y_n)$, $\mathbf{x}$ and $\mathbf{y}$ can be two column vectors of DNA methylation matrix (see Figure 1), u_i, v_i are defined with $i=1,\ldots,n$ as follows:

$$u_i = \frac{x_i - med(\mathbf{x})}{T \cdot mad(\mathbf{x})} \tag{1}$$

$$v_i = \frac{y_i - med(\mathbf{y})}{T \cdot mad(\mathbf{y})} \tag{2}$$

$$mad(\mathbf{x}) = med\left(\left|x_i - med(\mathbf{x})\right|\right) \tag{3}$$

where $med(\mathbf{x})$ and $med(\mathbf{y})$ are the median of vector $\mathbf{x}$ and $\mathbf{y}$ respectively. $mad(\cdot)$ represents the median absolute deviation of numeric vector. Based on u_i and v_i. The weights $w_i^{(\mathbf{x})}$ for x_i and $w_i^{(\mathbf{y})}$ for y_i are defined as follows:

$$w_i^{(\mathbf{x})} = \left(1-u_i^2\right)^2 \mathbf{I}\left(1-|u_i|\right) \tag{4}$$

$$w_i^{(\mathbf{y})} = \left(1-v_i^2\right)^2 \mathbf{I}\left(1-|v_i|\right) \tag{5}$$

where $\mathbf{I}$ is an indicator equation, for equation (5), the indicator equation $\mathbf{I}\left(1-|v_i|\right)$ is 1 if $\left(1-|v_i|\right)>0$ and otherwise equals to 0. The same situation occurs for equation (4). For equation (2) and (5), as the difference between y_i and $med(\mathbf{y})$ gets smaller and smaller, $w_i^{(\mathbf{y})}$ gets closer to 1. If the difference between y_i and $med(\mathbf{y})$ is larger than $T \cdot mad(\mathbf{y})$, $w_i^{(\mathbf{y})}$ equals to 0. The same situation occurs for equation (1) and equation (4). T is a pre-defined parameter. Let us discuss pre-defined parameter T. In practice, T is chosen between 5 and 9; the bigger T, the smaller the number of values to be filtered out. In this article, T is set to 9. For T, we chose the highest valid value to include all potentially interesting values. In addition, users can determine T based on the data characteristics and possible proportion of outliers. The weight values of all outliers are guaranteed to be 0. Based on $w_i^{(\mathbf{x})}$ and $w_i^{(\mathbf{y})}$, we can define BIMC of vector $\mathbf{x}$ and $\mathbf{y}$ as follows:

$$pre(\mathbf{x},\mathbf{y}) = \left(\sqrt{\sum_{i=1}^{n}\left[\left(x_i - med(\mathbf{x})\right)\cdot w_i^{(\mathbf{x})}\right]^2} \cdot \sqrt{\sum_{i=1}^{n}\left[\left(y_i - med(\mathbf{y})\right)\cdot w_i^{(\mathbf{y})}\right]^2} \right)^{-1} \tag{6}$$

$$\mathrm{BIMC}(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n}\left(x_i - med(\mathbf{x})\right)\cdot w_i^{(\mathbf{x})} \cdot \left(y_i - med(\mathbf{y})\right)\cdot w_i^{(\mathbf{y})} \cdot pre(\mathbf{x},\mathbf{y}) \tag{7}$$

where $\mathrm{BIMC}(\mathbf{x},\mathbf{y})$ represents the BIMC of $\mathbf{x}$ and ... It should be noted that, the range of BIMC is from -1 to 1. If there is a strong positive linear relationship between

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2866836, IEEE/ACM Transactions on Computational Biology and Bioinformatics

DNA methylation vectors, the value of BIMC will be close to 1. If there is a strong negative linear relationship between methylation vectors, the value of BIMC will be close to -1. If there is no linear relationship or only a weak linear relationship between methylation vectors, the value of BIMC will be 0 or close to 0.

In order to explain how to use biweight midcorrelation to calculate the correlation coefficient of DNA methylation sites. Let us take DNA methylation matrix as an example. Example of a DNA methylation matrix is shown in Figure 1.

For each sample S_i , we measure level of expression at each DNA methylation site, d_j^i represents the expression level of the j-th DNA methylation site for the i-th sample where $j = 1,\ldots,q$. The i-th column vector of DNA methylation matrix is expression level of i-th DNA methylation site ($DMsi$). Then we can use the first and second column vectors of the matrix to calculate the correlation coefficient between $DMs1$ and $DMs2$.

In this study, the absolute value of the BIMC of DNA methylation sites in the same function module approaches to be 1. The stronger the association, the larger the BIMC value of the DNA methylation sites. In

$$\begin{bmatrix} & DMs1 & DMs2 & \cdots & DMsq \\ S_1 & d_1^1 & d_2^1 & \cdots & d_q^1 \\ S_2 & d_1^2 & d_2^2 & \cdots & d_q^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_n & d_1^n & d_2^n & \cdots & d_q^n \end{bmatrix}$$

Fig. 1. An example of DNA methylation matrix.

order to find DNA methylation sites that are strongly associated with a specific DNA methylation site, we use BIMC with 'differential correlation strategy', which shall be described in the next section.

2.1.2 DIFFERENTIAL CORRELATION STRATEGY

Reasonable use of data characteristics can provide effective prior information. According to the existing relevant biological research results, the value of a methylation site in patient sample differs from that in the control sample; at the same time, the association between methylation sites in the patient sample differs from that in the control sample. In addition, DNA methylation sites are often located in core sequences of gene promoters and transcription start sites (TSS) [35], and they often affect the nearest gene [36]. This prior information helps us find important DNA methylation sites for specific genes. In this section, we introduce 'differential correlation strategy'. This strategy is used to find gene-related DNA methylation sites using two different label samples (disease label samples and normal label samples). This strategy is similar to the gene differential co-expression analysis which find biological pathways or gene function module through measuring gene correlation changes between patient samples and control samples [32-34].

Firstly, for a specific gene G, we determine the corresponding DNA methylation site DMs_G at the promoter or transcription start sit of the gene G by searching relevant public database (e.g. MethDB: http://www.methdb.de/). Secondly, we calculate the BIMCs between DNA methylation site DMs_G and all other DNA methylation sites in the patient samples and control samples, respectively. $\mathrm{BIMC}_{dmps} = [a_1, a_2, \ldots, a_{q-1}]$ represents BIMCs between DMs_G and all other DNA methylation sites in the patient samples, and vector $\mathrm{BIMC}_{dmcs} = [b_1, b_2, \ldots, b_{q-1}]$ represents BIMCs between DMs_G and all other DNA methylation sites in the control samples. For the two elements of the same position in two vectors BIMC_{dmps} and BIMC_{dmcs} (e.g. a_1 and b_1), the two values are calculated from the same pair of DNA methylation sites. Thirdly, in order to find differential correlation methylation sites, we set two thresholds TS_1 and TS_2 ; vector $\mathrm{BIMC}_{dmpc} = [e_1, e_2, \cdots, e_{q-1}]$ contains only 0 and 1 (0 for no correlation and 1 for correlation) indicates whether methylation site is related to the DMs_G. If absolute value of $\mathrm{BIMC}_{dmps}(i)$ is less than or equal to TS_1 and absolute value of $\mathrm{BIMC}_{dmcs}(i)$ is larger than or equal to TS_2, $\mathrm{BIMC}_{dmpc}(i)$ is set to 1; if $\mathrm{BIMC}_{dmps}(i)$ is larger than or equal to TS_2 and absolute value of $\mathrm{BIMC}_{dmcs}(i)$ is less than or equal to TS_1, $\mathrm{BIMC}_{dmpc}(i)$ is set to 1; otherwise $\mathrm{BIMC}_{dmpc}(i)$ is set to 0. $\mathrm{BIMC}_{dmpc}(i) = 1$ represents the i-th DNA methylation site and DMs_G have a strong correlation in patient samples but have a weak correlation in control samples or have a weak correlation in patient samples but have a strong correlation in control samples, which means i-th methylation site and DMs_G function together to affect expression of gene G. TS_1 is set to 0.3 and TS_2 is set to 0.7. Meanwhile, for each methylation site that interacts with DMs_G to affect gene G, we calculate absolute value of difference between the absolute values of its two BIMC values (i.e. $\mathrm{BIMC}_{abs} = \left\| \mathrm{BIMC}_{dmps}(i) \right| - \left| \mathrm{BIMC}_{dmcs}(i) \right\|$). Finally, we select DNA methylation sites based on BIMC_{dmpc}, then sort methylation sites based on BIMC_{abs}, and select high-ranking methylation sites.

2.2 BMNPGRN Model

Our proposed method BMNPGRN is used to discover a directed network that encodes the regulatory relationships over a set of genes and selected methylation sits, these selected methylation sites were obtained by applying biweight midcorrelation coefficient and differential correlation strategy to the raw data. And we focus on the impact of CNV on the gene which in the region of CNV. Let $\mathbf{X} \in \mathbf{R}^{N \times P}$ denotes gene expression matrix of N samples and P selected genes. $\mathbf{D} \in \mathbf{R}^{N \times Q_S}$ denotes DNA methylation matrix of N samples and Q_S DNA methylation sites, $\mathbf{C} \in \mathbf{R}^{N \times P}$ denotes copy number variation matrix of N samples and P CNVs. The three data matrices are defined as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p]$, $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_{q_s}]$, and $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_p]$ where $\mathbf{x}_i, \mathbf{d}_i, \mathbf{c}_i$ are i-th column vector of $\mathbf{X}$, $\mathbf{D}$ and $\mathbf{C}$ respectively. Gene expression vector

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2866836, IEEE/ACM Transactions on Computational Biology and Bioinformatics

affected by biological factors is defined as follows:

$$\mathbf{x}_i = \mathbf{X}\mathbf{f}_i + \mathbf{D}\mathbf{g}_i + \mathbf{C}\mathbf{w}_i + u_i + \varepsilon_i \quad (8)$$

where $\mathbf{f}_i$ and $\mathbf{w}_i$ are *i*-th column vector of adjacency matrix $\mathbf{F} \in \mathbf{R}^{P\times P}$ and $\mathbf{W} \in \mathbf{R}^{P\times P}$ respectively; $\mathbf{g}_i$ is sub-vector of *i*-th column vector of adjacency matrix $\mathbf{G} \in \mathbf{R}^{Q\times Q}$, the number of elements in the $\mathbf{g}_i$ is Q_S, the other elements of the *i*-th column vector in adjacency matrix $\mathbf{G} \in \mathbf{R}^{Q\times Q}$ are 0. u_i is a model bias and ε_i is a residual. f_{ij} represents the regulation modes from *i*-th gene to *j*-th gene: positive for activation, negative for repression, and 0 for no regulation. In this study, we focus on the regulation relationship among genes, thus there is no self-regulation and every diagonal element $\mathrm{f}_{ii}=0$. Meanwhile, there is no two nodes cycle (i.e. both f_{ij} and f_{ji} are non-zero) in BMNPGRN. In addition, it is assumed that a gene can be directly affected by methylation site belong to the gene and other DNA methylation sites that are highly correlated with the methylation site. And a gene can be directly affected by CNV that belong to the gene but no other CNVs. g_{ij} and w_{ii} represent the regulation weight value of DNA methylation site and CNV of *i*-th gene. For equation (8), we want to find $\mathbf{F}$, $\mathbf{G}$ and $\mathbf{W}$ that can best represent weights of genes, DNA methylation sites and CNVs. Our goal is to estimate $\mathbf{f}_i, \mathbf{g}_i$, and $\mathbf{w}_i$ that minimize ε_i, after the bias is removed by mean centering, equation (8) can be restated as follows:

$$\min_{\mathbf{f}_i,\mathbf{g}_i,\mathbf{w}_i} \left\|\mathbf{x}_i - \mathbf{X}\mathbf{f}_i - \mathbf{D}\mathbf{g}_i - \mathbf{C}\mathbf{w}_i\right\|_2^2 \quad (9)$$

where $\|\cdot\|_2$ denotes L2-norm. Generally speaking, L1-norm penalty is often used to avoid overfitting and accurately find genes, DNA methylation sites and CNVs that influence *i*-th gene. L1-norm penalty is applied to all columns of $\mathbf{F}$, $\mathbf{G}$ and $\mathbf{W}$, then equation (9) based on the L1-regularization can be expressed as follows:

$$\min_{\mathbf{f}_i,\mathbf{g}_i,\mathbf{w}_i} \left\|\mathbf{x}_i - \mathbf{X}\mathbf{f}_i - \mathbf{D}\mathbf{g}_i - \mathbf{C}\mathbf{w}_i\right\|_2^2 + \lambda_1\left\|\mathbf{f}_i\right\|_1 + \lambda_2\left\|\mathbf{g}_i\right\|_1 + \lambda_3\left\|\mathbf{w}_i\right\|_1 \quad (10)$$

where λ_1, λ_2 and λ_3 are hyper-parameters for sparsity regularization. Equation (10) can be rewritten as follows:

$$L(\boldsymbol{\beta}_i) = \min_{\boldsymbol{\beta}_i} \left\|\mathbf{x}_i - \mathbf{Y}\boldsymbol{\beta}_i\right\|_2^2 + \lambda\left\|\boldsymbol{\beta}_i\right\|_1 \quad (11)$$

where

$$\mathbf{Y} = \left[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_p, \mathbf{d}_1, \ldots, \mathbf{d}_i, \ldots, \mathbf{d}_{q_s}, \mathbf{c}_i\right] \quad (12)$$

$$\boldsymbol{\beta}_i = \left[\mathrm{f}_{i1}, \mathrm{f}_{i2}, \ldots, \mathrm{f}_{i-1}, \mathrm{f}_{i+1}, \ldots, \mathrm{f}_{ip}, \mathrm{g}_1, \ldots, \mathrm{g}_i, \ldots, \mathrm{g}_{q_s}, \mathrm{w}_{ii}\right] \quad (13)$$

where $\left[\mathbf{d}_1, \ldots, \mathbf{d}_i, \ldots, \mathbf{d}_{q_s}\right]$ represents expression vector of Q_S selected DNA methylation sites and $\left[\mathrm{g}_1, \ldots, \mathrm{g}_i, \ldots, \mathrm{g}_{q_s}\right]$ represent corresponding weights. Equation (11) is a Least Absolute Shrinkage and Selection Operator (LASSO) problem [21]. Nonconvex penalties like smoothly clipped absolute deviation (SCAD) and minimax concave penalty (MCP) are considered as worthwhile alternatives to the LASSO. We show how to use nonconvex penalties and coordinate descent algorithm to find gene related biological factors, next.

MCP penalty is defined on $[0,\infty)$ by:

$$k_{\lambda,\gamma}(\theta) = \begin{cases} \lambda\theta - \dfrac{\theta^2}{2\gamma}, & \text{if } \theta \le \gamma\lambda, \\ \dfrac{1}{2}\gamma\lambda^2, & \text{if } \theta > \gamma\lambda, \end{cases}$$
$$k'_{\lambda,\gamma}(\theta) = \begin{cases} \lambda - \dfrac{\theta}{\gamma}, & \text{if } \theta \le \gamma\lambda, \\ 0, & \text{if } \theta > \gamma\lambda. \end{cases} \quad (14)$$

for $\lambda \ge =0$ and $\gamma > 1$. SCAD is defined on $[0,\infty)$ by:

$$k_{\lambda,\gamma}(\theta) = \begin{cases} \lambda\theta, & \text{if } \theta \le \lambda, \\ \dfrac{-\theta^2 + 2\gamma\lambda\theta - \lambda^2}{2(\gamma-1)}, & \text{if } \lambda < \theta \le \gamma\lambda, \\ \dfrac{\lambda^2(\gamma+1)}{2}, & \text{if } \theta > \gamma\lambda, \end{cases}$$
$$k'_{\lambda,\gamma}(\theta) = \begin{cases} \lambda, & \text{if } \theta \le \lambda, \\ \dfrac{\gamma\lambda - \theta}{\gamma - 1} & \text{if } \lambda < \theta \le \gamma\lambda, \\ 0, & \text{if } \theta > \gamma\lambda. \end{cases} \quad (15)$$

for $\lambda \ge =0$ and $\gamma > 2$. The above LASSO problem (i.e. equation (11)) can be restated as follows:

$$L(\boldsymbol{\beta}_i) = \min_{\boldsymbol{\beta}_i} \left\|\mathbf{x}_i - \mathbf{Y}\boldsymbol{\beta}_i\right\|_2^2 + \sum_{j=1}^{P+Q_S} k_{\lambda,\gamma}\left(\left|\beta_{ij}\right|\right) \quad (16)$$

where $P + Q_S$ represents the sum of genes, selected DNA methylation sites and CNV.

Solving equation (16) depends on the choice of the tuning parameters λ and γ. This is usually accomplished with cross-validation strategy. However, if only cross-validation is used, we often obtain many false positives (i.e. true zero weights are non-zero in estimated weight vector $\boldsymbol{\beta}_i$). In order to effectively control false positives and find more truly related genes and DNA methylation sites. We augment equation (16) with stability selection [29] to determine edges in gene regulatory network. Stability selection strategy is a bootstrapping-type algorithm which can effectively control false positives. Briefly, stability selection strategy works as follows: Firstly, we randomly select half (i.e. $\lfloor N/2 \rfloor$) of the total samples for T_{ss} times. Secondly, for each selected subsamples, equation (16) is run on the samples. Finally, stability selection strategy select factors (i.e. genes, DNA methylation sites and copy number variation) whose weight values are non-zero for $T_{ss} \cdot z_{ts}$ times, where z_{ts} is a user-defined parameters.

We discuss how to determine user-defined parameters T_{ss} and z_{ts}, next. In practices, a large number of experimental results show that $T_{ss} \ge 100$ is sufficient to achieve false positive control [29]. Meanwhile, z_{ts} is often chosen between 0.5 and 1; in theory, under certain conditions, [29] gives the relationship between the number of false positives and z_{ts}. When finding edges between genes, DNA

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2866836, IEEE/ACM Transactions on Computational Biology and Bioinformatics

methylation sites, copy number variation and the *i*-th genes,

$$E(\mathrm{V}_i) \le \frac{1}{2z_{ts}-1} \quad \frac{l^2_{\lambda^*,\gamma^*}}{P+Q_S} \tag{17}$$

where $E(\mathrm{V}_i)$ is the expected number of falsely detected factors (i.e. genes, DNA methylation sites and copy number variation) for the *i*-th gene. $l^2_{\lambda^*,\gamma^*}$ is the number of non-zero weights found by BMNPGRN with λ^* and γ^*. Equation (17) shows that the upper bound on the number of false positives is inversely proportional to z_{ts}.

For the *i*-th gene, stability selection strategy assign each factor (i.e. genes, DNA methylation sites and copy number variation) a weight score, reflecting their degrees of significance. For the association between the *i*-th gene and *j*-th factor, the association score $\mathbf{S}_i^j$ is defined by the proportion of the cases where the *j*-th factor is selected to the total number of randomly selected subsamples.

2.3 EVALUATION CRITERIA

We compare our method BMNPGRN with three methods DCGRN, ARACNE and CLR. In order to strictly evaluate a method on its performance as balanced by true positive rates (TPR) and false positive rates (FPR), we use receiver operating characteristic (ROC) curve, area under receiver operating characteristic (AUROC) and area under precision recall curve (AUPRC), which are calculated based on confusion matrix (Fig. 2).

$$TPR = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{18}$$

$$FPR = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} \tag{19}$$

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{20}$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{21}$$

where true positive (TP), false positive (FP), true negative (TN), and false negative (FN) are defined in Figure 2. We calculate TP, FP, TN, and FN to measure the accuracy criteria, TPR and FPR. The performance of BMNPGRN are compared to other methods with ROC curves and AUROC values and AUPRC values.

True Class \ Predicted Class	Associated	Non-associated
Associated	True Positive (TP)	False Negative (FN)
Non-associated	False Positive (FP)	True Negative (TN)

Fig. 2. Confusion matrix used to evaluate the GRN inference method.

3 RESULTS

In this section, we first compared the performance of BMNPGRN with three state-of-the-art methods (including DCGRN, ARACNE and CLR) in two kinds of synthetic datasets. DCGRN is a method using LASSO with DNA methylation site and CNV [22], ARACNE is an algorithm for reverse engineering of gene regulatory network [30]. CLR is a context likelihood relatedness-based gene regulatory network inference algorithm [31]. We then applied BMNPGRN on breast cancer data from TCGA and identified several significant gene regulatory networks.

3.1 PERFORMANCE COMPARISON ON SYNTHETIC DATASETS

In this section, the BMNPGRN algorithm was tested in terms of different aspects, and the section is organized as follows.

First, in order to demonstrate the superiority of BMNPGRN over three state-of-the-art methods (including DCGRN, ARACNE and CLR) in naive synthetic dataset used by three state-of-the-art methods, a naive experiment is designed to compare BMNPGRN with three methods.

Second, the performance of the BMNPGRN algorithm is compared with three methods (i.e. DCGRN, ARACNE and CLR) in a more practical synthetic dataset which contains relevant factors (i.e. associated genes and associated DNA methylation sites).

3.1.1 NAIVE SYNTHETIC DATASETS

In this section, we compare the performance of the BMNPGRN with three methods (including DCGRN, ARACNE and CLR) in naive synthetic dataset used by three state-of-the-art methods. In this kind of dataset, it is assumed that the *i*-th gene is only affected by DNA methylation site which belong to the *i*-th gene, and biological factors are independent of each other. First we introduce this naive synthetic random networks. We followed the synthetic dataset generation method of [22]. *P* represents the number of genes and is set to 30, 40, and 50. $P \times P$ matrix $\mathbf{F}$ is initialized to zero matrix, then elements of $\mathbf{F}$ are randomly selected avoiding any cycle. The parameter E_G represents the number of inbound edges per gene on average. The larger the parameter E_G, the more complex the network. The weight value f_{ii} is uniformly distributed over $[0.5,1]$ or $[-0.5,1]$. Meanwhile, the adjacency matrix $\mathbf{G}$ and $\mathbf{W}$ are initialized to zero matrix, then diagonal elements (g_{ii} and w_{ii}) are randomly selected. The parameter R_{DC} represents the percentage of nodes that are regulated by DNA methylation sites and copy number variation. For example, if *P* and R_{DC} are 30 and 0.2 respectively, then it means six (30×0.2) DNA methylation sites and copy number variations regulate corresponding genes (i.e. six diagonal elements of $\mathbf{G}$ and $\mathbf{W}$ are non-zero). The selected weight value g_{ii} is uniformly distributed over $[0.5,1]$ or $[-0.5,1]$, and selected w_{ii} is set to 1. d_{ii} from uniform distribution $U[0,1]$. c_{ii} is randomly set as -2, -1, 0, 1, or 2 with the probabilities 0.01, 0.22, 0.55, 0.2 and 0.02 respectively. The design gene expression matrix $\mathbf{X}$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2866836, IEEE/ACM Transactions on Computational Biology and Bioinformatics

can be generated by calculating $\mathbf{X}=(\mathbf{DG}+\mathbf{CW}+\mathbf{E})(\mathbf{F}-\mathbf{I})^{-1}$, where each element of $\mathbf{E}$ is generated from zero-mean Gaussian distribution.

3.1.2 PERFORMANCE EVALUAITION ON NAIVE SYNTHETIC DATASETS

We compare our method BMNPGRN with three methods DCGRN, ARACNE and CLR. Given data $\mathbf{X}$, $\mathbf{D}$ and $\mathbf{C}$, $\hat{\mathbf{F}}$, $\hat{\mathbf{G}}$ and $\hat{\mathbf{W}}$ are inferred, and then they are compared to the true edges of $\mathbf{F}$, $\mathbf{G}$ and $\mathbf{W}$, and calculating TPR and FPR. For the naive synthetic dataset, we generated two sets of datasets based on different parameter combinations. The parameters of first set of naive synthetic datasets (FNSD) are: $N \in \{100, 200, 300, 400\}$, $E_G = 1$, R_{DC}=0.3 and $P = 30$. The parameters of second set of naive synthetic datasets (SNSD) are: $N \in \{100, 200, 300, 400\}$, $E_G = 3$, R_{DC}=0.5 and $P = 50$. For BMNPGRN, we used two different penalty functions MCP and SCAD, and set T_{ss}=100, $z_{ts} = 0.6$ or $z_{ts} = 0.8$. The ROC comparison results for the two sets of datasets are shown in Figure 3 and Figure 4 respectively. The AUROC (Area Under Receiver Operating Characteristic) and AUPRC (Area Under Precision Recall Curve) results for the two sets of datasets are shown in Table S1 and Table S2 respectively. Experimental results show that BMNPGRN outperforms other methods in naive synthetic dataset.

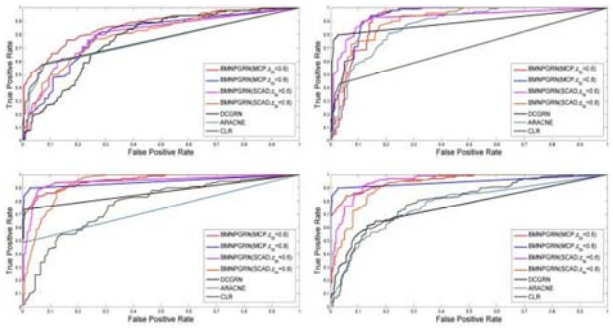


Fig. 3. ROCs of BMNPGRN with DCGRN, ARACNE and CLR: N=100, $E_G = 1$, R_{DC}=0.3 (top left), N=200, $E_G = 1$, R_{DC}=0.3 (top right), N=300, $E_G = 1$, R_{DC}=0.3 (bottom left), N=400, $E_G = 1$, R_{DC}=0.3 (bottom right). For BMNPGRN, we show the results with two settings for z_{ts} 0.6, 0.8 and two penalty functions.

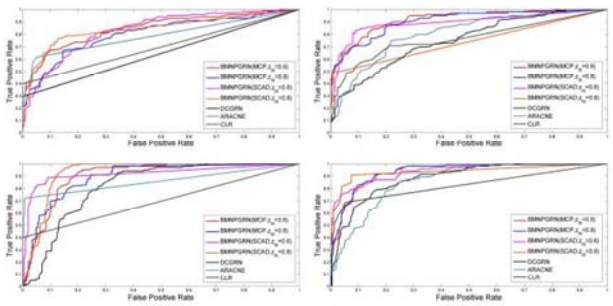


Fig. 4. ROCs of BMNPGRN with DCGRN, ARACNE and CLR: N=100, $E_G = 3$, R_{DC}=0.5 (top left), N=200, $E_G = 3$, R_{DC}=0.5 (top right), N=300, $E_G = 3$, R_{DC}=0.5 (bottom left), N=400, $E_G = 3$, R_{DC}=0.5 (bottom right). For BMNPGRN, we show the results with two settings for z_{ts} 0.6, 0.8 and two penalty functions.

3.1.3 COMPLEX SYNTHETIC DATASETS

In this section, we compare our method BMNPGRN with three methods DCGRN, ARACNE and CLR in a complex synthetic datasets. $\mathbf{C}$, $\mathbf{W}$, $\mathbf{E}$, $\mathbf{F}$ and $\mathbf{I}$ are initialized in the same way as in naive synthetic dataset. We describe how to generate $\mathbf{D}$ and $\mathbf{G}$, next. First, we randomly selected K DNA methylation sites which located in core sequences of gene promoters and transcription start sites (TSS). Second, generating K groups based on these K DNA methylation sites. The number of DNA methylation sites in group is randomly selected from 2, 3, and 4. Third, we used popular value-based co-correlation network method to generate co-correlation expression vectors of $\mathbf{D}$ [12]. The

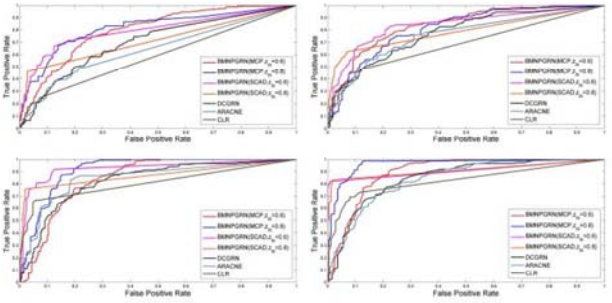


Fig. 5. ROCs of BMNPGRN with DCGRN, ARACNE and CLR: N=100, $E_G = 2$, R_{DC}=0.3 (top left), N=200, $E_G = 2$, R_{DC}=0.3 (top right), N=300, $E_G = 2$, R_{DC}=0.3 (bottom left), N=400, $E_G = 2$, R_{DC}=0.3 (bottom right). For BMNPGRN, we show the results with two settings for z_{ts} 0.6, 0.8 and two penalty functions.

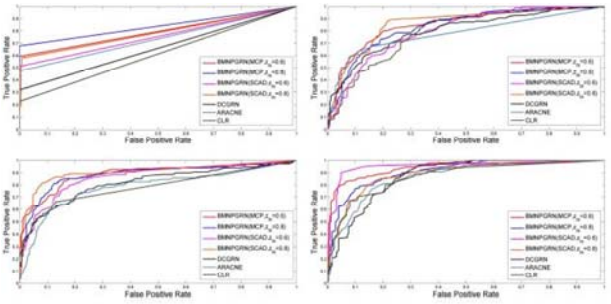


Fig. 6. ROCs of BMNPGRN with DCGRN, ARACNE and CLR: N=100, $E_G = 2$, R_{DC}=0.5 (top left), N=200, $E_G = 2$, R_{DC}=0.5 (top right), N=300, $E_G = 2$, R_{DC}=0.5 (bottom left), N=400, $E_G = 2$, R_{DC}=0.5 (bottom right). For BMNPGRN, we show the results with two settings for z_{ts} 0.6, 0.8 and two penalty functions.

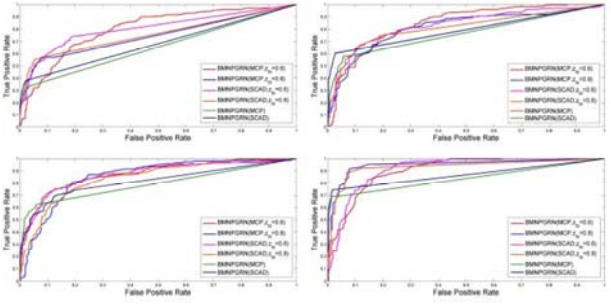


Fig. 7. ROCs of BMNPGRN with DCGRN, ARACNE and CLR: N=100, $E_G = 3$, R_{DC}=0.4 (top left), N=200, $E_G = 3$, R_{DC}=0.4 (top right), N=300, $E_G = 3$, R_{DC}=0.4 (bottom left), N=400, $E_G = 3$, R_{DC}=0.4 (bottom right). For BMNPGRN, we show the results with two settings for z_{ts} 0.6, 0.8 and two penalty functions.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2866836, IEEE/ACM Transactions on Computational Biology and Bioinformatics

correlation coefficient is calculated by the Pearson correlation formula. It should be noted that the sample contains both patient and normal samples. Based on this information, co-correlation networks were constructed. Finally, $\mathbf{d}_{ij}$ from uniform distribution $U[0,1]$. The diagonal elements of $\mathbf{G}$ are 0. The selected weight value g_{ij} is uniformly distributed over $[0.5,1]$ or $[-0.5,1]$.

For the complex synthetic dataset, we generated two sets of datasets based on different parameter combinations. The parameters of first set of complex synthetic datasets (FCSD) are: $N \in \{100, 200, 300, 400\}$, $E_G = 2$, R_{DC}=0.3 and $P = 30$. The parameters of second set of complex synthetic datasets (SCSD) are: $N \in \{100, 200, 300, 400\}$, $E_G = 3$, R_{DC}=0.5 and $P = 50$. For BMNPGRN, we used two different penalty functions MCP and SCAD, and set T_{ss}=100, $z_{ts} = 0.6$ or $z_{ts} = 0.8$. The ROC comparison results for the two sets of datasets are shown in Figure 5 and Figure 6 respectively. The AUROC and AUPRC results for the two sets of datasets are shown in Table S3 and Table S4 respectively. Experimental results show that BMNPGRN outperforms other methods in naive synthetic dataset.

3.2 PERFORMANCE EVALUATION OF BMNPGRN AND BMNPGRN WITHOUT STABILITY SELECTION

In order to verify the effect of the stability selection strategy [29], we generated one set of datasets based on different parameter combinations. The parameters of third set of complex synthetic datasets (TCSD) are: $N \in \{100, 200, 300, 400\}$, $E_G = 3$, R_{DC}=0.4, and $P = 30$. This dataset is used to compare the performance of BMNPGRN and BMNPGRN without stability selection. The ROC comparison results of the datasets are shown in Figure 7.

3.3 Application to Breast Cancer Data

Breast cancer is the most common malignancy in United States women, accounting for >40,000 deaths each year. Discovering cancer-related biological pathway information is one of the key challenges of breast cancer research. Identifying comprehensive gene regulatory network can provide an important resource for studying the underlying mechanisms of breast cancer.

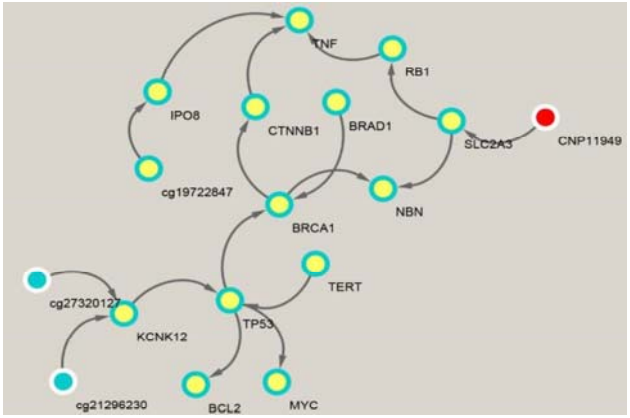


Fig. 8. Gene regulatory network from Breast Cancer Data. DNA methylation sites(blue dot),gene(yellow dot), copy number polymorphism(red dot).

We applied BMNPGRN to a dataset, containing measurement profiles of gene expression, DNA methylation, and copy number variation. Gene expression profiles was measured experimentally using the Illumina HiSeq 2000 RNA Sequencing platform, downloaded from TCGA. DNA methylation profiles were obtained from the ratios of background-corrected methylated and un-methylated probe intensities measured by Illumina Infinium HumanMethylation450k BeadArrays, also downloaded from TCGA. The gene-level CNVs estimated using the GISTIC2 method. There are 760 case and 80 control samples measured from breast tissue. There are 24776 genes and corresponding CNVs in gene expression data and copy number variation data. There are 485578 DNA methylation sites in data.

Figure 8 is the integrated network with 14 genes that have high absolute value of coefficient. There are two DNA methylation sites that are connected to gene KCNK12. Also, there is a CNV that is connected to the gene SLC2A3. Several studies have been reported KCNK12 is associated with breast cancer but there is no report about associations with DNA methylation sites of the gene KCNK12 [37-39]. Several studies have been reported IPO8 is associated with breast cancer but there is no report about associations with DNA methylation sites of the gene IPO8 [40, 41]. It should be noted that CNP11949 is first discovered related to breast cancer [42, 43]. Several studies have been reported BRCA1 and TP53 are associated with breast cancer [44]. Experimental results provide genes that need attention in future work. TNF and CTNNB1 play an important role in gene regulatory network.

4 CONCLUSIONS

In this paper, we propose an approach, which use biweight midcorrelation to measure the correlation between factors and make use of Nonconvex Penalty based sparse regression for Gene Regulatory Network inference (BMNPGRN). BMNPGRN incorporates multi-omics data (including DNA methylation and copy number variation) and their interactions in gene regulatory network model. The experimental results on synthetic datasets show that BMNPGRN outperforms popular and state-of-the-art methods (including DCGRN, ARACNE and CLR) under false positive control.

References

[1] P. A. Jones, J. P. Issa, and S. Baylin, "Targeting the cancer epigenome for therapy," *Nature Reviews Genetics,* vol. 17, no. 10, pp. 630, 2016.

[2] D. S. Huang, and H. J. Yu, "Normalized Feature Vectors: A Novel Alignment-Free Sequence Comparison Method Based on the Numbers of Adjacent Amino Acids," *IEEE/ACM Transactions on Computational Biology & Bioinformatics,* vol. 10, no. 2, pp. 457-467, 2013.

[3] M. R. Aure, S. K. Leivonen, T. Fleischer, Q. Zhu, J.

Overgaard, J. Alsner, T. Tramm, R. Louhimo, G. I. G. Alnæs, and M. Perälä, "Individual and combined effects of DNA methylation and copy number alterations on miRNA expression in breast tumors," *Genome Biology,* vol. 14, no. 11, pp. R126, 2013.

[4] D. S. Huang, "Systematic theory of neural networks for pattern recognition," *Publishing House of Electronic Industry of China, Beijing,* vol. 201, 1996.

[5] D. S. HUANG, "Radial basis probabilistic neural networks: model and application," *International Journal of Pattern Recognition & Artificial Intelligence,* vol. 13, no. 07, pp. 1083-1101, 1999.

[6] D. S. Huang, and J.-X. Du, "A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks," *IEEE Transactions on Neural Networks,* vol. 19, no. 12, pp. 2099-2115, 2008.

[7] C. H. Zheng, L. Zhang, T. Y. Ng, K. S. Chi, and D. S. Huang, "Molecular Pattern Discovery Based on Penalized Matrix Decomposition," *IEEE/ACM Transactions on Computational Biology & Bioinformatics,* vol. 8, no. 6, pp. 1592-1603, 2011.

[8] D. S. Huang, L. Zhang, K. Han, S. Deng, K. Yang, and H. Zhang, "Prediction of protein-protein interactions based on protein-protein correlation using least squares regression," *Curr Protein Pept Sci,* vol. 15, no. 6, pp. 553-560, 2014.

[9] S. D. Bolboac, and L. J?NTSCHI, "Pearson versus Spearman, Kendall's Tau Correlation Analysis on Structure-Activity Relationships of Biologic Active Compounds," *Leonardo Journal of Sciences,* vol. 5, no. 9, pp. 179-200, 2006.

[10] K. Pearson, "Note on Regression and Inheritance in the Case of Two Parents," *Proceedings of the Royal Society of London,* vol. 58, pp. 240-242, 2006.

[11] P. Langfelder, and S. Horvath, "WGCNA: an R package for weighted correlation network analysis," *Bmc Bioinformatics,* vol. 9, no. 1, pp. 559, 2008.

[12] J. K. Choi, U. Yu, O. J. Yoo, and S. Kim, "Differential coexpression analysis using microarray data and its application to human cancer," *Bioinformatics,* vol. 21, no. 24, pp. 4348-55, 2005.

[13] C. H. Zheng, D. S. Huang, L. Zhang, and X. Z. Kong, "Tumor Clustering Using Nonnegative Matrix Factorization With Gene Selection," *IEEE Transactions on Information Technology in Biomedicine,* vol. 13, no. 4, pp. 599-607, 2009.

[14] S. P. Deng, L. Zhu, and D. S. Huang, *Predicting hub genes associated with cervical cancer through gene co-expression networks*: IEEE Computer Society Press, 2016.

[15] D. S. Huang, X. M. Zhao, G. B. Huang, and Y. M. Cheung, "Classifying protein sequences using hydropathy blocks," *Pattern Recognition,* vol. 39, no. 12, pp. 2293-2300, 2006.

[16] A. V. Werhli, and D. Husmeier, "Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge," *Stat Appl Genet Mol Biol,* vol. 6, no. 1, pp. 15-15, 2007.

[17] D. S. Huang, and C. H. Zheng, "Independent component analysis based penalized discriminate method for tumor classification using gene expression data," *Bioinformatics,* vol. 22, no. 15, pp. 1855-1862, 2006.

[18] S. P. Deng, L. Zhu, and D. S. Huang, "Mining the bladder cancer-associated genes by an integrated strategy for the construction and analysis of differential co-expression networks," *Bmc Genomics,* vol. 16, no. S3, pp. S4, 2015.

[19] L. Zhu, W. L. Guo, S. P. Deng, and D. S. Huang, "ChIP-PIT:Enhancing the Analysis of ChIP-Seq Data Using Convex-Relaxed Pair-Wise Interaction Tensor Decomposition," *IEEE/ACM Transactions on Computational Biology & Bioinformatics,* vol. 13, no. 1, pp. 55-63, 2016.

[20] L. Zhu, S. P. Deng, and D. S. Huang, "A Two-Stage Geometric Method for Pruning Unreliable Links in Protein-Protein Networks," *IEEE Transactions on Nanobioscience,* vol. 14, no. 5, pp. 528-534, 2015.

[21] R. Tibshirani, "Regression shrinkage and selection via the lasso: A retrospective," *Journal of the Royal Statistical Society,* vol. 73, no. 3, pp. 273-282, 2011.

[22] D. C. Kim, M. Kang, B. Zhang, X. Wu, C. Liu, and J. Gao, "Integration of DNA Methylation, Copy Number Variation, and Gene Expression for Gene Regulatory Network Inference and Application to Psychiatric Disorders." *IEEE International Conference on Bioinformatics and Bioengineering*, pp. 669-674, 2015.

[23] L. Zhu, Z. H. You, D. S. Huang, and B. Wang, "t-LSE: A Novel Robust Geometric Approach for Modeling

Protein-Protein Interaction Networks," *Plos One,* vol. 8, no. 4, pp. e58368, 2013.

[24] D. S. Huang, and W. Jiang, "A General CPL-AdS Methodology for Fixing Dynamic Parameters in Dual Environments," *IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society,* vol. 42, no. 5, pp. 1489-1500, 2012.

[25] S. P. Deng, and D. S. Huang, "SFAPS: An R package for structure/function analysis of protein sequences based on informational spectrum method," *Methods,* vol. 69, no. 3, pp. 207-212, 2014.

[26] X. Ma, Z. Liu, Z. Zhang, X. Huang, and W. Tang, "Multiple network algorithm for epigenetic modules via the integration of genome-wide DNA methylation and gene expression data," *Bmc Bioinformatics,* vol. 18, no. 1, pp. 72, 2017.

[27] J. Fan, and R. Li, "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Publications of the American Statistical Association,* vol. 96, no. 456, pp. 1348-1360, 2001.

[28] C. H. Zhang, "NEARLY UNBIASED VARIABLE SELECTION UNDER MINIMAX CONCAVE PENALTY," *Annals of Statistics,* vol. 38, no. 2, pp. 894-942, 2010.

[29] N. Meinshausen, and P. Bühlmann, "Stability selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* vol. 72, no. 4, pp. 417-473, 2010.

[30] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, F. R. Dalla, and A. Califano, "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *Bmc Bioinformatics,* vol. 7 Suppl 1, no. Suppl 1, pp. S7, 2006.

[31] J. J. Faith, B. Hayete, J. T. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. J. Collins, and T. S. Gardner, "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles," *Plos Biology,* vol. 5, no. 1, pp. e8, 2007.

[32] O. K. Kathleen, O. Karim, L. Marie-Hélène, B. Sahir, P. N. Tonin, and C. M. T. Greenwood, "Gene Coexpression Analyses Differentiate Networks Associated with Diverse Cancers Harboring TP53 Missense or Null Mutations," *Frontiers in Genetics,* vol. 7, 2016.

[33] D. Amar, H. Safer, and R. Shamir, "Dissection of Regulatory Networks that Are Altered in Disease via Differential Co-expression," *PLoS Computational Biology,9,3(2013-3-7),* vol. 9, no. 3, pp. e1002955, 2013.

[34] L. Yuan, C. H. Zheng, J. F. Xia, and D. S. Huang, "Module Based Differential Coexpression Analysis Method for Type 2 Diabetes," *Biomed Research International,* vol. 2015, pp. 836929, 2015.

[35] P. A. Jones, "Functions of DNA methylation: islands, start sites, gene bodies and beyond," *Nature Reviews Genetics,* vol. 13, no. 7, pp. 484-92, 2012.

[36] T. Dayeh, P. Volkov, S. Salö, E. Hall, E. Nilsson, A. H. Olsson, C. L. Kirkpatrick, C. B. Wollheim, L. Eliasson, and T. Rönn, "Genome-Wide DNA Methylation Analysis of Human Pancreatic Islets from Type 2 Diabetic and Non-Diabetic Donors Identifies Candidate Genes That Influence Insulin Secretion," *PLoS Genet.,* vol. 10, no. 3, pp. e1004160, 2014.

[37] K. C. Johnson, D. C. Koestler, C. Cheng, and B. C. Christensen, "Age-related DNA methylation in normal breast tissue and its relationship with invasive breast tumor methylation," *Epigenetics Official Journal of the Dna Methylation Society,* vol. 9, no. 2, pp. 268-275, 2014.

[38] M. O. Riener, E. Nikolopoulos, A. Herr, P. J. Wild, M. Hausmann, T. Wiech, M. Orlowskavolk, S. Lassmann, A. Walch, and M. Werner, "Microarray comparative genomic hybridization analysis of tubular breast carcinoma shows recurrent loss of the CDH13 locus on 16q," *Human Pathology,* vol. 39, no. 11, pp. 1621, 2008.

[39] K. A. Dookeran, W. Zhang, L. Stayner, and M. Argos, "Associations of two-pore domain potassium channels and triple negative breast cancer subtype in The Cancer Genome Atlas: systematic evaluation of gene expression and methylation," *Bmc Research Notes,* vol. 10, no. 1, pp. 475, 2017.

[40] C. Michaloglou, C. Crafter, R. Siersbæk, O. Delpuech, J. O. Curwen, L. S. Carnevalli, A. D. Staniszewska, U. M. Polanska, A. Cheraghchibashi, and M. Lawson, "Combined inhibition of mTOR and CDK4/6 is required for optimal blockade of E2F function and long term growth inhibition in estrogen receptor positive breast cancer," *Molecular Cancer*

Therapeutics, pp. molcanther.0537.2017, 2018.

[41] S. M. Guichard, Z. Howard, H. Dan, M. Roth, G. Hughes, J. Curwen, J. Yates, A. Logie, S. Holt, and C. M. Chresta, "Abstract 917: AZD2014, a dual mTORC1 and mTORC2 inhibitor is differentiated from allosteric inhibitors of mTORC1 in ER+ breast cancer," *Cancer Research,* vol. 72, no. 8 Supplement, pp. 917-917, 2012.

[42] X. P. Jiang, R. L. Elliott, and J. F. Head, "Exogenous normal mammary epithelial mitochondria suppress glycolytic metabolism and glucose uptake of human breast cancer cells," *Breast Cancer Research & Treatment,* vol. 153, no. 3, pp. 519, 2015.

[43] J. Lester, K. Crosthwaite, R. Stout, R. N. Jones, C. Holloman, C. Shapiro, and B. L. Andersen, "Women with breast cancer: self-reported distress in early survivorship," *Oncology Nursing Forum,* vol. 42, no. 1, pp. 17-23, 2015.

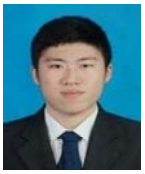
[44] M. Yamamoto, M. Hosoda, K. Nakano, S. Jia, K. C. Hatanaka, E. Takakuwa, Y. Hatanaka, Y. Matsuno, and H. Yamashita, "p53 accumulation is a strong predictor of recurrence in estrogen receptor-positive breast cancer patients treated with aromatase inhibitors," *Cancer Science,* vol. 105, no. 1, pp. 81-88, 2014.

Acknowledgements

This work is partly supported by National Natural Science Foundation of China (Grant nos. 61732012, 61520106006, 31571364, 61532008, U1611265, 61672382, 61772370, 61702371, 61772357 and 61672203) and China Postdoctoral Science Foundation (Grant nos. 2017M611619, and 2016M601646), and supported by "BAGUI Scholar" Program of Guangxi Zhuang Autonomous Region of China.

Competing financial interests

The authors declare no competing financial interests.



Lin Yuan received the M.A. degree in communication and information system from Qufu Normal University, Rizhao, China, in 2014. He is currently studying for the Ph.D. degree in computer application from Tongji University, Shanghai, China. His current research interests include machine learning and bioinformatics.

Le-Hang Guo received the MD degree from Tongji University, Shanghai, China. He is currently studying for the Ph.D. degree in medical imaging from Nanjing Medical University, Nanjing, China. His current research interests include machine learning, artificial intelligence, and tumor imaging.

Chang-An Yuan received the Ph.D. degree in Computer Application Technology from the Sichuan University, China, in2006. He is a professor at Guangxi Teachers Education University. His research interests include Computational intelligence and Data mining.

Youhua Zhang, Male, Born in 1966.11, Ph.D. of University of Science and Technology of China, Professor, Master's tutor. Dean of school of information and computer of Anhui Agricultural University, executive council member of Computer Applications Branch of China Agricultural Society, executive council member of Anhui Association of Agricultural Information. Member of National Bee modern industrial technology and product quality and safety standards system. Expert advisory committee of Anhui provincial human resources and Social Security Department of information technology. Executive Committee of the Hefei branch of the CCF. Mainly engaged in teaching and research in computer applications and logistics engineering. In charge or participate in Provincial science and technology research, National Science and Technology Support Program, 863 project, National Natural Science Foundation, won the First Prize of scientific and technological progress in Anhui Province.

Kyungsook Han is a professor at the Department of Computer Science and Engineering, Inha University, Korea. She received a BS cum laude from Seoul National University in 1983, an MS cum laude in Computer Science from KAIST in 1985, another MS in Computer Science from the University of Minnesota at Minneapolis, USA in 1989, and a PhD in Computer Science from Rutgers University, USA in 1994. Her research areas include bioinformatics, visualization, and data mining.

Asoke K. Nandi received the degree of Ph.D. in Physics from the University of Cambridge (Trinity College), Cambridge (UK). He held academic positions in several universities, including Oxford (UK), Imperial College London (UK), Strathclyde (UK), and Liverpool (UK) as well as Finland Distinguished

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2018.2866836, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Professorship in Jyvaskyla (Finland). In 2013 he moved to Brunel University (UK), to become the Chair and Head of Electronic and Computer Engineering. Professor Nandi is a Distinguished Visiting Professor at Tongji University (China) and an Adjunct Professor at University of Calgary (Canada).

His current research interests lie in the areas of signal processing and machine learning, with applications to communications, gene expression data, functional magnetic resonance data, and biomedical data. He has authored over 500 technical publications, including 200 journal papers as well as four books, entitled Automatic Modulation Classification: Principles, Algorithms and Applications (Wiley, 2015), Integrative Cluster Analysis in Bioinformatics (Wiley, 2015), Automatic Modulation Recognition of Communications Signals (Springer, 1996), and Blind Estimation Using Higher-Order Statistics (Springer, 1999),. Recently he published in Blood, BMC Bioinformatics, IEEE TWC, NeuroImage, PLOS ONE, Royal Society Interface, and Signal Processing. The h-index of his publications is 63 (Google Scholar).

Professor Nandi is a Fellow of the Royal Academy of Engineering and also a Fellow of seven other institutions including the IEEE and the IET. Among the many awards he received are the Institute of Electrical and Electronics Engineers (USA) Heinrich Hertz Award in 2012, the Glory of Bengal Award for his outstanding achievements in scientific research in 2010, the Water Arbitration Prize of the Institution of Mechanical Engineers (UK) in 1999, and the Mountbatten Premium, Division Award of the Electronics and Communications Division, of the Institution of Electrical Engineers (UK) in 1998.

Barry Honig, tenured professor at Columbia University, the US Academy of Sciences. Director of Biology and Bioinformatics computing center of Columbia University. Editorial Board of PNAS, Journal of Molecular Biology, Structure, Biochemistry and Current Opinion in Structural Biology. Research interests include computational biology and bioinformatics. As of 2013, the international high-level SCI journals published more than 300 papers. SCI cited more than 15,600 times, Nature published 17, Science 3, Cell 3, PNAS 22. 2004 was elected to the National Academy of Sciences. In 2007 the US National Academy of Sciences Alexander Hollaender Award.

De-Shuang Huang received the B.Sc., M.Sc. and Ph.D. degrees all in electronic engineering from Institute of Electronic Engineering, Hefei, China, National Defense University of Science and Technology, Changsha, China and Xidian University, Xian, China, in 1986, 1989 and 1993, respectively. During 1993-1997 period he was a postdoctoral research fellow respectively in Beijing Institute of Technology and in National Key Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China. In Sept, 2000, he joined the Institute of Intelligent Machines, Chinese Academy of Sciences as the Recipient of "Hundred Talents Program of CAS". In September 2011, he entered into Tongji University as Chaired Professor. From Sept 2000 to Mar 2001, he worked as Research Associate in Hong Kong Polytechnic University. From Aug. to Sept. 2003, he visited the George Washington University as visiting professor, Washington DC, USA. From July to Dec 2004, he worked as the University Fellow in Hong Kong Baptist University. From March, 2005 to March, 2006, he worked as Research Fellow in Chinese University of Hong Kong. From March to July, 2006, he worked as visiting professor in Queen's University of Belfast, UK. In 2007, 2008, 2009, he worked as visiting professor in Inha University, Korea, respectively. At present, he is the director of Institute of Machines Learning and Systems Biology, Tongji University. Dr. Huang is currently Fellow of International Association of Pattern Recognition (IAPR Fellow), senior members of the IEEE and International Neural Networks Society. He has published over 180 journal papers. Also, in 1996, he published a book entitled "Systematic Theory of Neural Networks for Pattern Recognition" (in Chinese), which won the Second-Class Prize of the 8th Excellent High Technology Books of China, and in 2001 & 2009 another two books entitled "Intelligent Signal Processing Technique for High Resolution Radars" (in Chinese) and "The Study of Data Mining Methods for Gene Expression Profiles" (in Chinese), respectively. His current research interest includes bioinformatics, pattern recognition and machine learning.