BMC Medical Genomics

RESEARCH Open Access

Integrative approach for inference of gene regulatory networks using lasso-based random featuring and application to psychiatric disorders

Dongchul Kim[1], Mingon Kang[2], Ashis Biswas[3], Chunyu Liu[4] and Jean Gao[3]*

From IEEE International Conference on Bioinformatics and Biomedicine 2015
Washington, DC, USA. 9-12 November 2015

Abstract

Background: Inferring gene regulatory networks is one of the most interesting research areas in the systems biology. Many inference methods have been developed by using a variety of computational models and approaches. However, there are two issues to solve. First, depending on the structural or computational model of inference method, the results tend to be inconsistent due to innately different advantages and limitations of the methods. Therefore the combination of dissimilar approaches is demanded as an alternative way in order to overcome the limitations of standalone methods through complementary integration. Second, sparse linear regression that is penalized by the regularization parameter (lasso) and bootstrapping-based sparse linear regression methods were suggested in state of the art methods for network inference but they are not effective for a small sample size data and also a true regulator could be missed if the target gene is strongly affected by an indirect regulator with high correlation or another true regulator.

Results: We present two novel network inference methods based on the integration of three different criteria, (i) z-score to measure the variation of gene expression from knockout data, (ii) mutual information for the dependency between two genes, and (iii) linear regression-based feature selection.
Based on these criterion, we propose a lasso-based random feature selection algorithm (LARF) to achieve better performance overcoming the limitations of bootstrapping as mentioned above.

Conclusions: In this work, there are three main contributions. First, our z score-based method to measure gene expression variations from knockout data is more effective than similar criteria of related works. Second, we confirmed that the true regulator selection can be effectively improved by LARF. Lastly, we verified that an integrative approach can clearly outperform a single method when two different methods are effectively jointed. In the experiments, our methods were validated by outperforming the state of the art methods on DREAM challenge data, and then LARF was applied to inferences of gene regulatory network associated with psychiatric disorders.

Keywords: Gene regulatory network, Psychiatric disorder

*Correspondence: gao@uta.edu
[3]Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, US
Full list of author information is available at the end of the article

Background

Inferring gene regulatory networks (GRN) from biological data is currently the most interesting area of the systems biology research aiming to elucidate cellular and physiological mechanisms. GRN inference, which is often referred to as reverse engineering, is a process in which the network structure that best represents the regulation relationship over gene expression data is estimated. An inferred GRN consists of nodes and edges representing genes and gene-gene regulatory interactions (activation or suppression) respectively. Once the regulation maps are constructed by identifying the interactions of genes from high-throughput data such as gene microarray [1], we can gain insight into complex biological process from the regulatory networks in order to discover biomarkers for a target disease and apply further it to drug design [2, 3].

Basically the inference method should be determined depending on both what kind of data such as gene expression, gene-Transcription Factor (TF) [4], or protein-protein interaction (PPI) [5] are used to infer and which type of network model, such as directed or indirected graph [6], we assume. In addition, we have to consider the case of data integration. Namely, not only individual data but also multiple data types together (i.e. integration of gene expression and gene-TF data [7]) can be used for more reliable inference [8, 9]. As an assumption in this work, we limit our inference methods for directed network with a single data type: gene expression data. In order to decipher regulatory interactions with gene microarray data, which provides the gene expression level regulated by the other genes directly or indirectly, the number of effective network inference methods have been proposed by employing a variety of computational and structural models based on boolean networks [10], Bayesian networks [11], information theory [12], regression model [13], and so on. Depending on the different approaches, however, the results tend to be irregular due to inherently different advantages and limitations of each of the inference solutions [14]. The results of the Dialogue on Reverse Engineering Assessment and Methods (DREAM) project [15] describe well the pros and cons of the different methods as well as how effectively they can work together when the advantages of all methods are integrated (but it does not mean any combination always outperforms any other standalone method). More specifically, we note that they conclude two points through the experiments that (i) there is a limit to a single criterion for continuous improvement of network inference research without the integration and (ii) specifically the bootstrapping (re-sampling) based regression method [16] is required to avoid overfitting in regression-based methods [15].

As the motivation of our first strategy to this end, we focus on an integration of Mutual Information (MI) and L_1 regularized linear regression referred to as lasso [17] but we exclude the learning Bayesian network in the integration. The learning structure of Bayesian networks is somewhat infeasible due to both the discretization problem of a small sample size data and the high cost of computational learning in large scale data. MI is an information theoretic criteria that has been conventionally used for learning large scale network structure [18]. Although MI based approaches such as CLR [19] and ARACNE [20] are limited to reconstructing only an indirected graph unlike linear regression and Bayesian networks, these methods have the popular advantages of computational simplicity and non-linear dependency enabler. In practice, the shortcoming of MI is that it is prone to fail in differentiation between indirect regulation and direct ones. For example, when there is a highly correlated indirect regulation from G1 to G3 like Fig. 1b, MI tends to incorrectly predict feed-forward loops (Fig. 1a) but not cascades (Fig. 1b). Lasso is also frequently used to select the regulators of a given target gene assuming sparseness of GRN in order to avoid the overfitting of the least-squares problem. In contrast to MI, indirect regulation edge in cascades could be pruned away by lasso in which the objective function is penalized for sparsity by a regularization parameter, called the tuning parameter λ. However, a weakness of regression-based method is that only a strong direct regulator is more likely to be selected than another direct regulator in Feed-forward loops. Therefore, the integration of two methods is considered to deal with the trade-off. The motivation of our second strategy is that the property of knockout data allows us to measure statistical variations between wild-type gene expression and perturbed gene expression after knocking them out to provide the cause-effect information between those two genes. However, there is the limitation that the method is only applicable to gene knockout data.

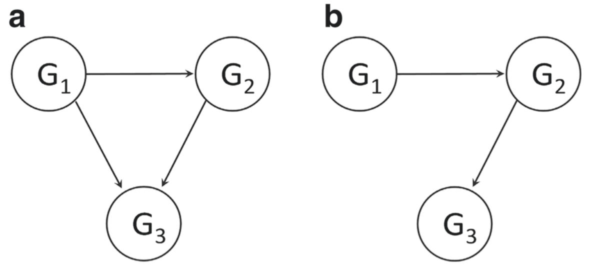


Fig. 1 Example of network structures. **a** Feed-forward loops and **b** Cascades. When G3 is a target gene, G1→G3 and G2→G3 of Cascades are indirect and direct regulations respectively. In MI-based methods, indirect regulations are likely to be selected incorrectly in Cascades. In regression based method, strong direct regulators are more likely to be selected than another direct regulator in Feed-forward loops

In this paper, we propose two methods, IMLARF (integration of MI and LARF) and ISLARF (integration of z-score and LARF). First, IMLARF indicates the integration of MI and LARF and consists of three steps. The first step of IMLARF is to build a matrix where each element is an edge score calculated by MI. In order to overcome the limitation of MI as mentioned above, the second step is to construct another edge score matrix using LARF, then the two edge score matrices are combined as the last step. In LARF, we regard a sparse linear regression as a feature selection since our goal is to identify the regulators that best predict the expression level of target genes. The problem is that features selected by lasso tend to be overfitted to a given tuning parameter λ, and thus the *unstability* problem caused by this overfitting can be solved by using bootstrapping [12, 21] in which data is randomly re-sampled so that a more stable selection can be achieved. However, the limitation of re-sampling is that it could not be effective in the case of a small sample size. Another limitation of bootstrapping is that the true variable (regulator gene) is likely to be missed (false negative) when strong indirect or direct regulators exist. LARF is similar to bootstrapping but LARF selects variables among randomly pre-selected candidate features in each iteration over different tuning parameters of lasso optimization so that true features weakly correlated to the target gene could not be missed, excluding indirect or direct regulators from the feature set. The second method we propose is ISLARF, which integrates two criteria, ZS and LARF. ZS is the name of the criteria that uses the z-score of variation of the knocked out gene expression. Although ISLARF is available only to knockout data, the performance is highly superior to other z-score based similar methods with knockout data in related works.

In the experimental evaluation, we validate the proposed method on a dataset from the DREAM3 challenge [22]. In addition, we explore the gene networks of Psychiatric disease with the related genes. The results shows that the proposed method significantly outperforms the state-of-the art [23, 24] and re-builds the known regulations of genes possibly associated with Psychiatric Disorders.

Methods

Problem definition

We begin with a brief definition of problems and notations. The network we target is a directed graph that consists of n nodes and $n(n-1)$ edges representing genes and regulations respectively. Given a matrix $\mathbf{X}\in\mathbb{R}^{N\times n}$ where N is number of samples, we denote the i-th column by a vector $\mathbf{x}_i$ indicating expression levels of i-th gene over N samples, and we also let $X = \{X_1, \ldots, X_n\}$ be a set of variables (genes, features, node, and variable are interchangeably used in this paper). The goal of our work is to not only identify the regulators given a target gene but also to define the confidence level of regulation as a weight of the edge. In other words, we estimate the weight of all possible regulations, which are directed edges between all pairs of nodes $\{X_i \leftarrow X_j : i, j \in X\}$ in the network , then select only edges that have a higher weight than pre-defined threshold θ. As a final result, therefore, a weight matrix $\mathbf{W}\in\mathbb{R}^{n\times n}$ is returned by the inference method, and W_j^i represents a confidence level of the regulation when target gene i is connected to activator or suppressor gene j. In the following sections, we present how the edge weight is estimated by information theory, the LARF algorithm, and the z-score from knockout data.

Overview

IMLARF and ISLARF

The first method we propose, IMLARF, consists of three steps. Figure 2a describes the overview of the proposed method. First, a symmetric edge weight matrix M is calculated by mutual information assuming that, if two genes have a higher mutual dependency, they are more likely to be in the regulation relationship. Second, another edge weight matrix F is produced by the LARF algorithm that consistently gives higher weight to the true edge from regulator to target gene. Lastly, the two weight matrices are combined by their entry-wise product $M \circ F = \{M_j^i \cdot F_j^i | i, j = 1, \ldots, n\}$. The second method, ISLARF, is similar to IMLARF but using z-score matrix, S, is used instead of MI matrix. If S_j^i has higher value, gene i is more likely to be regulated by gene j. So in the last step S is combined with F by their entry-wise product $S \circ F$

Information theoretic approach

Mutual information matrix

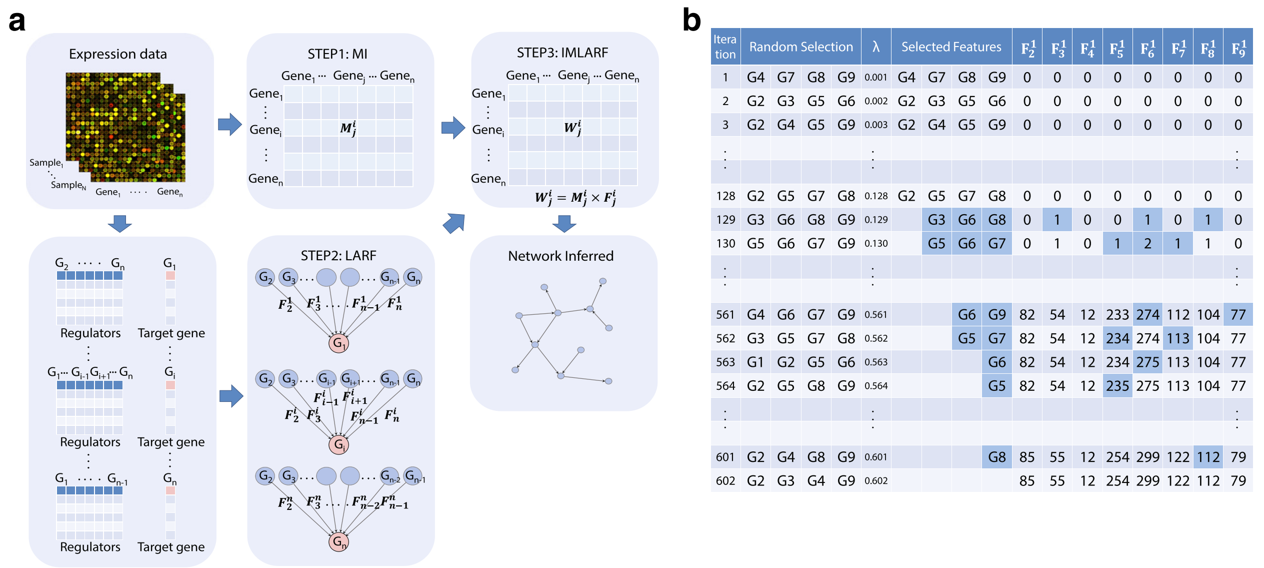
The dependency of two genes, X_i and X_j, can be measured by MI defined as

$$I(X_i, X_j) = \sum_{X_i, X_j} p(X_i, X_j) \log \frac{p(X_i, X_j)}{p(X_i)p(X_j)}, \qquad (1)$$

The strength of MI is the ability to measure non-linear dependencies of genes, but the limitation in practice is that the discretization of gene expression is required to calculate the probability of X_i and X_j. Instead, if we assume the Gaussian distribution of gene expression, MI can be computed with its original continuous values by using Gaussian mutual information [25] defined as

$$I(X_i, X_j) = -\frac{1}{2} \log \frac{|cov(X_i, X_j)|}{|cov(X_i, X_i)||cov(X_j, X_j)|}, \qquad (2)$$

where $cov(X_i)$ is the covariance matrix of variable X_i, and $|cov|$ is the determinant of covariance matrix. The reader is referred to [26] for more details. We build MI matrix in which each element M_j^i indicates the dependency between X_i and X_j which means that X_i and X_j are independent


a
Expression data
Sample$_1$
Sample$_N$
Gene$_1$ ⋯ Gene$_n$
STEP1: MI
Gene$_1$ ⋯ Gene$_j$ ⋯ Gene$_n$
Gene$_1$
Gene$_i$
Gene$_n$
M_j^i
STEP3: IMLARF
W_j^i
$W_j^i = M_j^i \times F_j^i$
Network Inferred
G_2 ⋯ G_n
G_1
Regulators
Target gene
G_1⋯ G_{i-1} G_{i+1}⋯ G_n
G_i
G_1 ⋯ G_{n-1}
G_n
STEP2: LARF
F_2^1 F_3^1 … F_{n-1}^1 F_n^1
F_2^i F_3^i F_{i-1}^i F_{i+1}^i F_{n-1}^i F_n^i
F_2^n F_3^n … F_{n-2}^n F_{n-1}^n
b

Iteration	Random Selection				λ	Selected Features				F_2^1	F_3^1	F_4^1	F_5^1	F_6^1	F_7^1	F_8^1	F_9^1
1	G4	G7	G8	G9	0.001	G4	G7	G8	G9	0	0	0	0	0	0	0	0
2	G2	G3	G5	G6	0.002	G2	G3	G5	G6	0	0	0	0	0	0	0	0
3	G2	G4	G5	G9	0.003	G2	G4	G5	G9	0	0	0	0	0	0	0	0
⋮					⋮												⋮
128	G2	G5	G7	G8	0.128	G2	G5	G7	G8	0	0	0	0	0	0	0	0
129	G3	G6	G8	G9	0.129		G3	G6	G8	0	1	0	0	1	0	1	0
130	G5	G6	G7	G9	0.130		G5	G6	G7	0	1	0	1	2	1	1	0
⋮					⋮												⋮
561	G4	G6	G7	G9	0.561			G6	G9	82	54	12	233	274	112	104	77
562	G3	G5	G7	G8	0.562			G5	G7	82	54	12	234	274	113	104	77
563	G1	G2	G5	G6	0.563				G6	82	54	12	234	275	113	104	77
564	G2	G5	G8	G9	0.564				G5	82	54	12	235	275	113	104	77
⋮					⋮												⋮
601	G2	G4	G8	G9	0.601				G8	85	55	12	254	299	122	112	79
602	G2	G3	G4	G9	0.602					85	55	12	254	299	122	112	79

Fig. 2 Overview of IMLARF and an example of LARF. **a** Overview of IMLARF. The algorithm consists of three steps, the construction of matrix (i) *M* and (ii) *F* and (iii) pairwise product of *M* and *F*. In ISLARF, the matrix *M* in step 1 is simply replaced with the matrix *S* (Section "Statistical approach"). **b** An example of procedures of LARF. It shows how the row vector F^1 of frequency matrix *F* given target gene G1 and 8 other candidate regulators (G2∼G9). By a predefined α, four random features are selected among eight genes in each iteration. In the beginning, F^1 is not increased and four random features are selected without sparsity since λ is not increased enough yet. The more λ is increased, the more the number of selected features (blue-colored cells) is decreased. If no feature is selected due to a highly increased λ, the iteration and frequency measure is finished

if $M_j^i = 0$ or M_j^i is relatively lower than other edges. Networks with the edges whose M_j^i are higher than the heuristic threshold are referred to as relevance networks. Two critical limitations of relevance networks, however, are that firstly, MI does not provide the direction of edges due to $M_j^i = M_i^j$, and secondly, the high co-expression and indirect regulation may cause false positives.

Statistical approach

Z-score and gene knockout data

We note that knockout data implies cause-effect information. The gene expression level after the perturbation of another certain gene provides the chance to observe if the gene is downstream of the perturbed gene. For example, if the variation between wild type of gene *j* (X_j^{wt}) and gene *j* expression measured after gene *i* is knocked out is high, gene *j* is likely to be regulated by gene *i*. The variation matrix *D* is defined as

$$D_j^i = X_j^{-i} - X_j^{wt} \tag{3}$$

$$S_i^j = \left| \frac{D_j^i - \mu_{D_j}}{\sigma_{D_j}} \right| \tag{4}$$

where X_j^{-i} is the expression level of gene *j* after knocking gene *i* out, and μ_{D_j} and σ_{D_j} is mean and standard deviation of *j*-th column vector D_j of variation matrix *D* respectively. As the z-score of D_j^i over D_j is the weight of regulation edge $Gi \rightarrow Gj$, the z-score of D_j^i is equivalent to S_i^j of edge weight matrix S. The limitation of this criterion is the availability only in knockout data.

Algorithm 1 LARF algorithm

```
procedure LARF($X, \alpha, r, stepsize, t$)
    for $i \leftarrow 1, n$ do
        for $h \leftarrow 1, t$ do
            $\lambda \leftarrow stepsize$
            repeat
                $X_{random} \leftarrow RandomFeatures(X^{\setminus i}, (n-1) \times \alpha)$
                $X' \leftarrow RandomSamples(X, N \times r)$
                $X_{selected} \leftarrow Lasso(X'_i, X'_{random}, \lambda)$
                if $0 < |X_{selected}| < n \times \alpha$ then
                    $F^i_{X_{selected}} \leftarrow F^i_{X_{selected}} + 1$
                end if
                $\lambda \leftarrow \lambda + stepsize$
            until $X_{selected} = \emptyset$
        end for
        $F^i \leftarrow Normalization(F^i)$
    end for
    return $F$
end procedure
```

LARF algorithm

The third approach for complementary integration of inference methods is based on L1-regularized linear regression (lasso) defined as

$$argmin_{\beta}||X_i - X^{\setminus i} \cdot \beta_i||_2^2 + \lambda||\beta_i||_1 \quad (5)$$

where coefficient column vector β_i represents regulation relationships between the target gene i and others. More precisely, after β_i is optimized to minimize the objective function (5), then if the j-th element of β_i is zero, gene j does not regulate gene i, otherwise it does. The optimization is performed for each target gene i, $i \in X$. Coefficient matrix $B = \{\beta_1, \ldots, \beta_n\}^T$ is equivalent to adjacency matrix where non-zero B_{ij} is the regulation edge from regulator gene j to target gene i. The tuning parameter λ in lasso is used to enforce network sparsity, so the number of selected (non-zero coefficient) variables varies with different λ. In our works, we regard variable selection of lasso as a feature selection to predict a target gene's expression level.

To overcome the overfitting problem and the strong indirect regulation problem, lasso is iteratively performed over different λ with randomly pre-defined candidate features rather than random samples like bootstrapping. More precisely, the basic idea of LARF is that lasso is iteratively performed with only randomly selected candidate features while increasing the tuning parameter, then giving weight to each feature by counting how many times each feature is selected in the iterations. We predefine the fraction of the number of all possible features as a parameter α ($0 < \alpha < 1$) for the candidate features. For example, when the number of all possible regulators is n=100, α=0.2 means that only 20 random candidate genes are used in a single iteration of lasso. After random featuring, random sampling is performed with parameter r which decides how many samples are used from the original data. For instance, when the original sample size is N=200 and r=0.7, only 140 random samples are used in each iteration of lasso. With randomly (uniform distribution) selected features and samples by parameter α, we iteratively run lasso over increasing tuning parameter λ until lasso does not select any features due to a certain high λ. In each iteration, random candidate features and samples are redefined again. Tuning parameter starts from zero and increases by the parameter *stepsize* that should be small enough, (e.g 0.001). Otherwise, both re-featuring and resampling will be biased. For each iteration, the frequency matrix F is updated. The i-th row of F is the frequency of feature selection for target gene i (F_i^i is supposed to be zero). For example, Fig. 1b describes how the F^i is measured. After finishing the iterations (repeat in line 5), we iteratively perform t times (t=10 in our experiments) of the process from line 5 to 13 again, and then i-th row vector of the frequency matrix is normalized by

$$F_j^i = \frac{(F_j^i - min(F_{-i}^i)}{max(F^i) - min(F_{-i}^i)}, \quad (6)$$

where

$$F_{-i}^i = \{F_j^i, j = 1,\ i-1, i+1, \ldots n\}, \quad (7)$$

and $max(F^i)$ and $min(F_{-i}^i)$ is maximum value of i-th row vector of F and minimum of F_{-i}^i.

Results

We first evaluated the performance of IMLARF and ISLARF on synthetic simulation data as compared to the state of the art, and then explored the inferred networks with real gene microarray data for psychiatric disorders. The synthetic, non-linear expression data is from DREAM3 *In Silico* Network challenge in which the data is created with the subnetworks of well-known reference networks for *Yeast*. To assess the edge weight matrix W elicited by proposed methods, first the matrix is converted to an edge list sorted by the confidence levels (weight), then the top k confidence level edges are selected to measure the accuracy criteria, such as true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The receiver operating characteristic (ROC) curves as a parametric curve were traced over different $k = 1, ..., n(n-1)$ to examine the trade-off between the true positive rate (TPR) and false positive rate (FPR). The criteria to represent the performance are defined as following:

- TPR=TP/(TP+FN)
- FPR=FP/(FP+TN)
- AUROC: the area under ROC curve.

We compared our method to each standalone method without integrations and also other well known the state of the art methods. The abbreviations of algorithms are listed below:

- MI: edge is scored by mutual information
- ZS: relative variation from wild type is measured by z-score.
- LARF: lasso based random featuring and sampling.
- IMLARF: integration of MI and LARF
- ISLARF: integration of ZS and LARF
- ZDR: top rank in DREAM 3 [23]
- GENIE3: top rank in DREAM 4 [24]
- TIGRESS: top rank in DREAM 5 [21]

Evaluation on the DREAM3 benchmarks

Materials

The data for DREAM3 *In Silico* Network challenge consists of three differently sized networks, (10, 50, and 100 genes), and there are five gold-standard networks for each size (total of 15 networks). The five networks are named Ecoli1, Ecoli2, Yeast1, Yeast2, and Yeast3. From each true network, three different data types (knockdown, knockout perturbations, and time series data) are provided, and the knockdown and knockout data includes a single wild type sample. In our experiments, only knockout data is used and 10-gene, 50-gene, 100-gene of Yeast1 networks are mainly tested.

Random sampling vs Random featuring

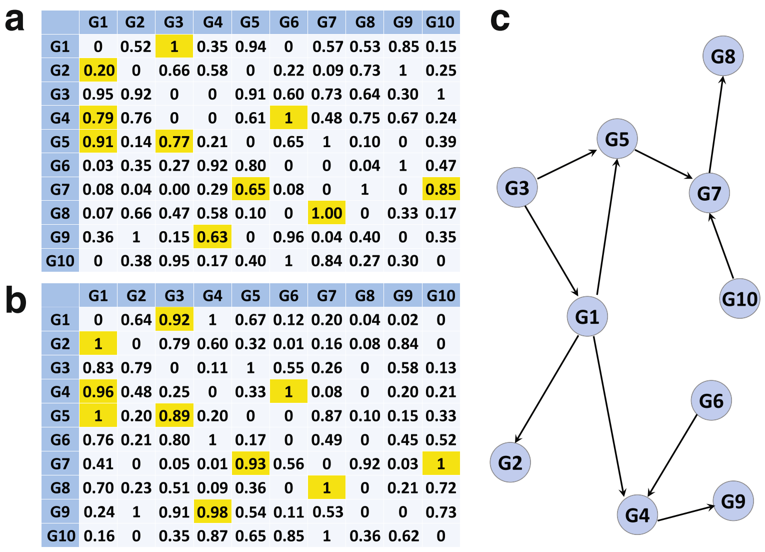
To evaluate how much more effectively LARF selects true edges than random sampling, we compared them with 10-gene Yeast1 network in Fig. 3. Figure 3a is the result of LARF with only random sampling (α=1, r=0.5) and 3b is with only random featuring (α=0.5, r=1). The normalized edge score is the average of 10 experiments and yellow colored cells indicate true edges. In Fig. 3a, though G2's true regulator is G1, G2←G3 is relatively higher than G2←G1 probably because of indirect regulation from G3 to G2 through G1. In Fig. 3b, G2←G1 is correctly estimated as true edge by random featuring. Similarly two true edges (G4←G1 and G5←G1) are inferred with the highest weight in random featuring but random sampling gives only 0.79 and 0.91 to two true edges (G4←G1 and G5←G1) due to another true edges (G4←G6 and G5←G3) have strong direct regulation (1 and 0.99).

Setting parameters

Before we compare our methods to other methods, we explored the optimal parameters that give the best results. As described in Fig. 4, the mean and standard deviation of AUROC are measured after LARF are 10 times performed over different parameters, α and r, for 50-gene Yeast1 network. The range of parameter is 0.2∼1 due to too small number of feature and sample in 10-gene network data. The best result (0.8501±0.0049) is recorded with α=0.4 and r=1 for 50-gene Yeast1 data. This indicates that the random sampling rate does not necessarily need to be applied to avoid overfitting once random featuring is applied. In addition, the figure also shows that the AUROC can be decreased with high standard deviation if both parameters are too small. According to the result of 10-gene and 100-gene Yeast1 data, if the sample size is small (N=10), the deviation is quite high in low α and r though AUROC is high. As the best result for 10-gene and 100-gene Yeast1 data, 0.925±0.0125 and 0.8611±0.0046 were achieved with α=0.5, r=1 and α=0.4, r=1 respectively. It also shows the random sampling could not make an improvement in both small and large sample sizes. Therefore we applied fixed parameters α=0.5, r=1 to all data sets in our experiments.

Effect of integration and performance comparisons

Table 1 presents the performance of integrative approaches compared to a single method. In the case of LARF-based methods, mean and deviation are measured after each method is performed 10 times for Yeast1 network of DREAM3. The integration of more than two methods is simply done by entry-wise product of edge



	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
G1	0	0.52	1	0.35	0.94	0	0.57	0.53	0.85	0.15
G2	0.20	0	0.66	0.58	0	0.22	0.09	0.73	1	0.25
G3	0.95	0.92	0	0	0.91	0.60	0.73	0.64	0.30	1
G4	0.79	0.76	0	0	0.61	1	0.48	0.75	0.67	0.24
G5	0.91	0.14	0.77	0.21	0	0.65	1	0.10	0	0.39
G6	0.03	0.35	0.27	0.92	0.80	0	0	0.04	1	0.47
G7	0.08	0.04	0.00	0.29	0.65	0.08	0	1	0	0.85
G8	0.07	0.66	0.47	0.58	0.10	0	1.00	0	0.33	0.17
G9	0.36	1	0.15	0.63	0	0.96	0.04	0.40	0	0.35
G10	0	0.38	0.95	0.17	0.40	1	0.84	0.27	0.30	0

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10
G1	0	0.64	0.92	1	0.67	0.12	0.20	0.04	0.02	0
G2	1	0	0.79	0.60	0.32	0.01	0.16	0.08	0.84	0
G3	0.83	0.79	0	0.11	1	0.55	0.26	0	0.58	0.13
G4	0.96	0.48	0.25	0	0.33	1	0.08	0	0.20	0.21
G5	1	0.20	0.89	0.20	0	0	0.87	0.10	0.15	0.33
G6	0.76	0.21	0.80	1	0.17	0	0.49	0	0.45	0.52
G7	0.41	0	0.05	0.01	0.93	0.56	0	0.92	0.03	1
G8	0.70	0.23	0.51	0.09	0.36	0	1	0	0.21	0.72
G9	0.24	1	0.91	0.98	0.54	0.11	0.53	0	0	0.73
G10	0.16	0	0.35	0.87	0.65	0.85	1	0.36	0.62	0

Fig. 3 Comparison of random sampling and featuring in LARF. **a** The result of LARF with only random sampling. **b** The result of LARF with only random featuring. **c** True network of 10-gene Yeast1 in DREAM3

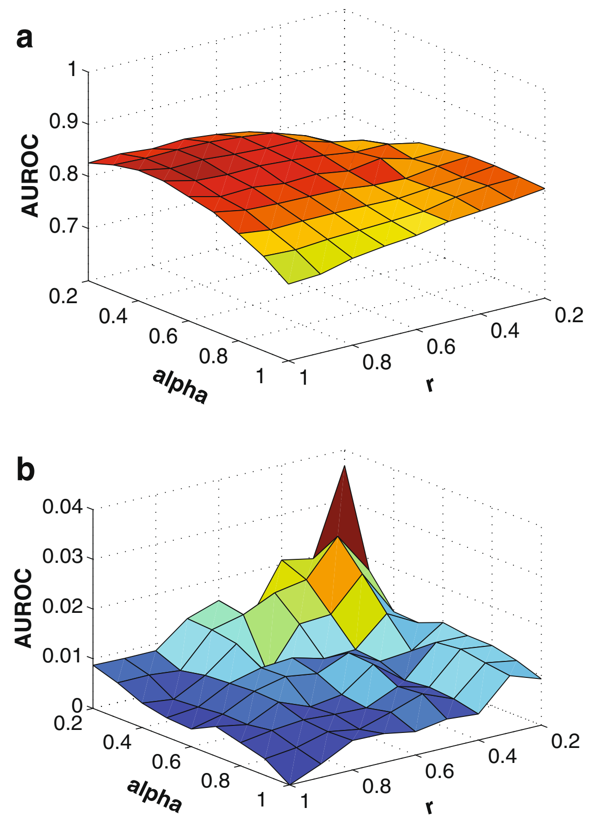


Fig. 4 Mean and standard deviation of AUROC with different parameters and 10 iterations of experiments for 50-gene Yeast1 network. (**a**) Mean and (**b**) standard deviation

score matrix. In TIGRESS-TF, the list of TF is provided as TIGRESS is designed for DREAM5 challenge data in which TF is given. Asterisk(*)-marked methods require knockout data. The integration of MI and LARF outperforms standalone MI and LARF except 50-gene. Similarly the performance of ISLARF is better than other integration such as ZS+MI and standalone ZS. If knockout data is not available, IMLARF will be the best method as ZS is not applicable. Since ZDR is based on knockout data, the result shows that ZDR is quite better than other methods such as IMLARF except in a small size network. In Fig. 5, the AUROC for proposed methods and the state of the art methods with 10-gene Yeast1 data are plotted after only a single experiment. Overall results show that ISLARF is the best method if knockout data is available, otherwise IMLARF is superior to other methods.

Table 1 AUROC of standalone and integrative methods

Method	10-gene	50-gene	100-gene
GENIE3	0.9175	0.8427	0.8631
TIGRESS	0.7044 ± 0.0056	0.8179 ± 0.0025	0.7690 ± 0.0023
TIGRESS-TF	0.8154 ± 0.0037	0.9006 ± 0.0010	0.8777 ± 0.0009
MI	0.9312	0.8329	0.8586
LARF	0.9250 ± 0.0154	**0.8489** ± 0.0038	0.8610 ± 0.0039
IMLARF	**0.9425** ± 0.0047	0.8487 ± 0.0032	**0.8701** ± 0.0012
ZDR*	0.8975	0.9223	0.8876
ZS*	0.9725	0.9204	0.8870
ZS*+MI	0.9775	0.8931	0.8925
ISLARF*	**0.9892** ± 0.0021	**0.9301** ± 0.0049	**0.9065** ± 0.0029

Inference of GRN for psychiatric disorders

In this section, the proposed method is applied to real gene expression data for psychiatric disorders. Through the experiments, we evaluate how the method constructs the network and explore what potential biomarkers of Psychiatric disorders are in the inferred networks. Psychiatric disorders data that are provided from the Stanley Medical Research Institute (SMRI) consist of gene expression data of 25833 genes and 131 samples (43 controls and 88 cases) including bipolar disorder, schizophrenia, major depression as three major psychiatric diseases.

To select genes possibly associated with psychiatric disorders, two statistical tests, t-test and z-test [27], are performed. In Fig. 6a, all genes are plotted by using p-value of t-test for y-axis and z-test value for x-axis, and the plot shows that two tests shows similar results in linear patterns. From these two tests, we selected 1407 genes as cut-off values are set to $-log_{10}(0.01)$ and ± 2.326 for t-test (y-axis) and z-test (x-axis). To find a module of genes that may interact to each other in Psychiatric disorders, we initially built a correlation matrix whose element of ith row and jth column is absolute value of correlation between expressions of ith and jth genes, and then clustering is performed to the estimated correlation matrix as shown in Fig. 6b. Based on the result of clustering, we manually set 8 groups of genes (yellow squares).

To analyze the relationship between clusters, first, IMLARF was applied to all 1407 genes with setting θ to 0.2. Figure 7 shows only the two largest components of the inferred network where node color indicates a cluster number after small components of the network are removed from the figure. The result is consistent with the correlation matrix in Fig. 6b showing the features as follows: (i) cluster 3, 6, and 8 in the network strongly and exclusively interact to each other, (ii) cluster 2, 4, and 5 are complicatedly interacting together, (iii) cluster 7 is widespread over the whole network.

To observe the strong regulation of the network, we inferred network with all the genes again after setting θ to 0.4. As a result, we displayed the second largest component in the inferred network in Fig. 8a. Most nodes of the network are genes of cluster 3 implying that cluster

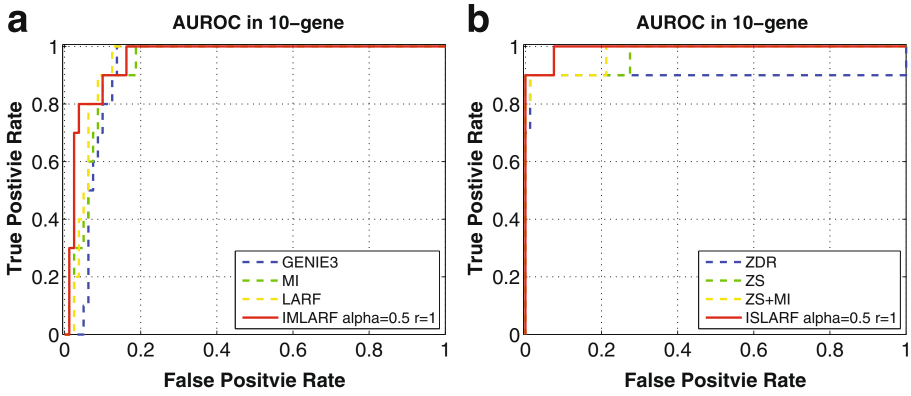


Fig. 5 ROC of the methods (**a**) without and (**b**) with gene deletion information in 10-gene network

3 is most exclusively and strongly interacting within the cluster. It is noted that 7 genes, DAO [28], PRDX6 [29], KCNN3 [30], TCF7L2 [31], RFX4 [32], FYN [33], and B3GAT2 [34] (yellow-colored nodes) , relevant to psychiatric disorders are involved and interestingly these genes except B3GAT2 constitute a connected subgraph. Blue-colored nodes indicate the genes that have more than two connection to yellow nodes supposing that these genes are likely to be susceptible to psychiatric disorders (In this paper we call yellow and blue gene reference gene and susceptible gene respectively. We define a gene as a reference gene if a gene appears with a psychiatric disease in the title of related literatures). There are 4 genes, SOX9, HEPH, AQP1, and SDC3 as susceptible genes, and it was already reported that SDC3 has a weak association with schizophrenia in related GWAS [35].

Figure 8b is the inferred network for cluster 7, and a total of 8 genes known as psychiatric disorder-related genes in related literatures are found as following: TEF [36], NR1D1 [37], KIF13A [38], ADCYAP1R1 [39], MDGA1 [40], GNAZ [41], CNR1 [42], and DCLK1 [43]. Additionally we defined 5 genes, ZBTB20, MAP7, ZBTB16, ANK2, and MRAP2, as susceptible genes, and surprisingly ZBTB20 [44], MAP7 [45], ZBTB16 [46], ANK2 [47] was also reported as schizophrenia disorder-associated genes in SNP and CNV-based studies. So we imply that it is worth to investigate the genes that have only an edge to reference gene as candidate genes associated with psychiatric disorder. In addition, reference genes in the network tend to interact with each other directly or indirectly though susceptible genes but they are not widely spread implying they may work together or may be co-regulated by another unknown biomarker.

The network inference result for the combination of cluster 4 and 5 is shown in Fig. 8c consisting of two components. There are 10 reference genes such as DLG4 [48]], MIF [49], SLC6A5 [50], GAD1 [51], GAD2 [52], GOT2 [53], RGS9 [54], HDAC9 [55], CDH7 [56], and BDNF [57], and 3 susceptible genes such as PRMT8, KIT, and ELAVL2. It is noted that ELAVL2 has connections to

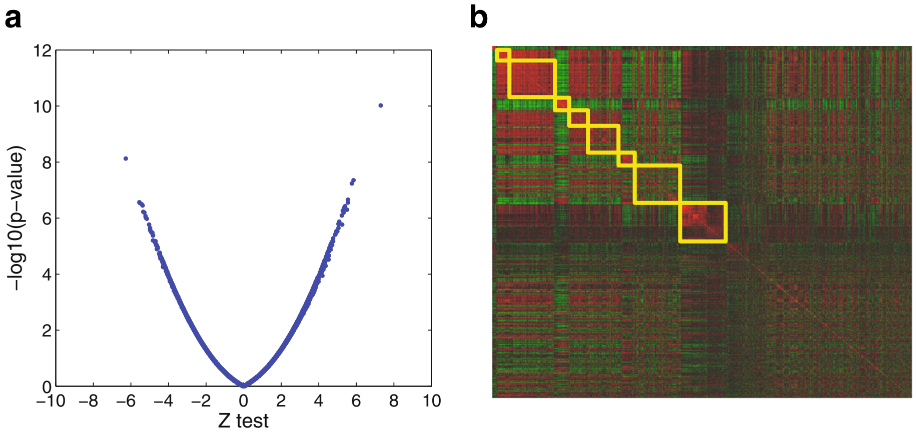


Fig. 6 Statistical test and clustering for gene selection. **a** t-test and z-test **b** clustered correlation matrix and 8 clusters (*yellow squares*)

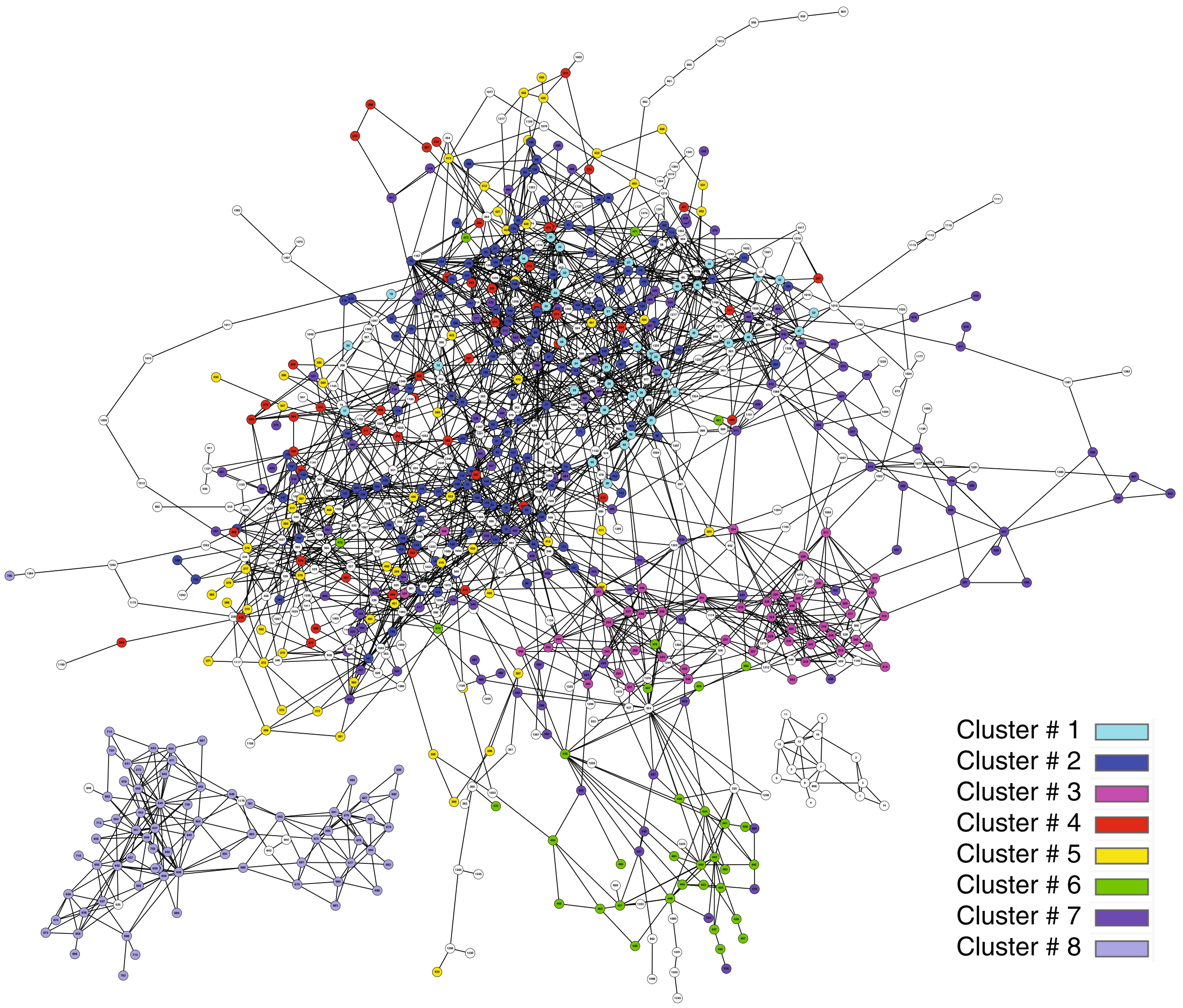


Fig. 7 Large components of network inferred with 1407 genes

three reference nodes and was reported as schizophrenia-related gene in recent GWAS [58].

Discussion

The difference between ZS and z-score of [23] is in whether the absolute value of variation D_j^i is taken before z-scoring or original value of D_j^i is used. In our method, we simply calculate the z-score to measure how many deviations the observed variation is above or below while the absolute value of variation $|D_j^i|$ is used for z-score. Since we want to know how much the variation of a gene is higher than another target gene after knockout of the source gene, the use of D_j^i rather than $|D_j^i|$ is more reasonable and it is not guaranteed to select high-variant genes if absolute value of D_j^i is used. Since random featuring and random sampling are performed in iterations of lasso, the computational time is significantly increased especially in finding optimal parameters. In implementation, the step size, therefore, should be set to a reasonably small value, and parallel processing (i.e. *parfor* in matlab) can reduce the processing time in practice (In our case, eight local cores are used). As a future work, we can integrate TF information additionally in the inference so that we can get more reliable results, and then also apply our method to DREAM5 challenge data for comparison to TIGRESS that utilizes TF information.

Conclusion

We presented two integrative approaches for gene regulatory network inference combining two different algorithms. First, IMLARF that we proposed is based on the integration of MI and LARF, which is a novel regression-based random featuring, to overcome the limitation of random sampling and MI. Secondly, ISLARF is the

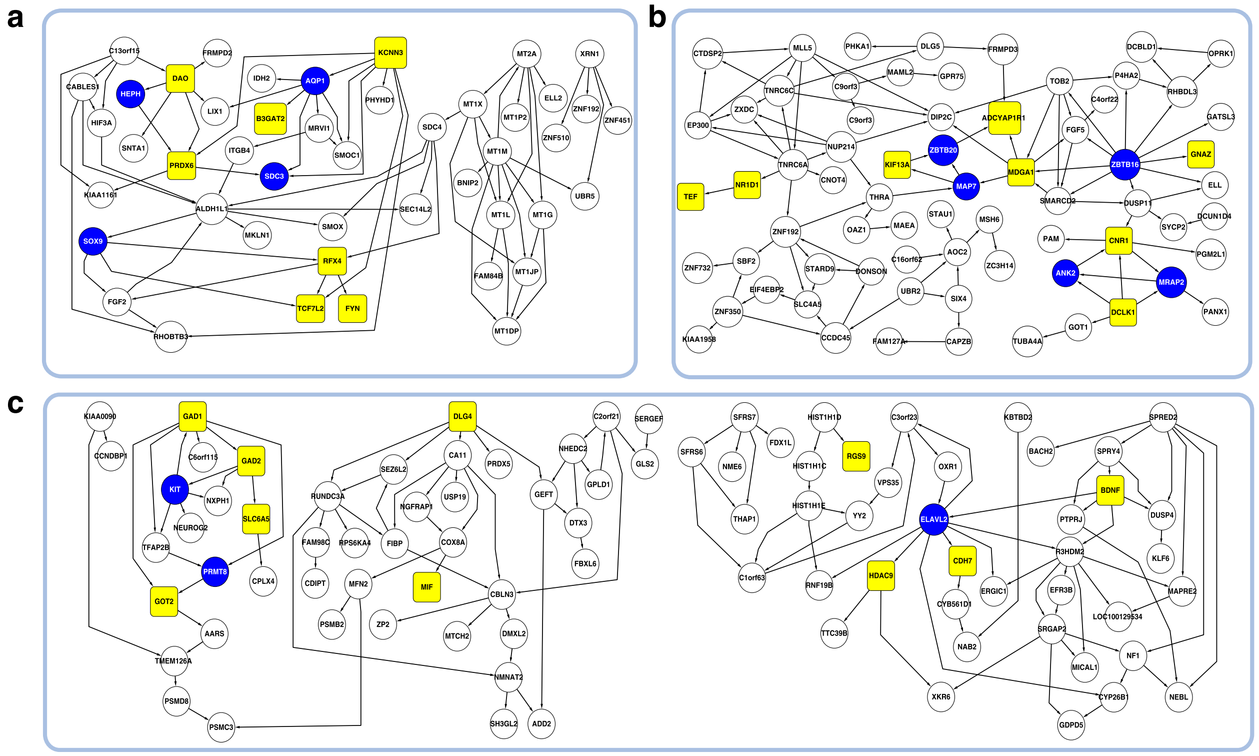


Fig. 8 Inferred gene regulatory networks for psychiatric disorder. **a** Cluster 3 **b** Cluster 7 **c** Cluster 4 and 5. Yellow-colored nodes indicate the genes known as psychiatric disorder genes in the literatures. Blue-colored nodes are the genes that are connected to more than two yellow genes

combination of LARF and ZS that is based on the z-score of variation of expression after the candidate regulator is knocked out. Both integrative methods outperform the standalone methods and the selected state of the art techniques on DREAM3 challenge data. In application to inference of the gene regulation associated with psychiatric disorders, we applied IMLARF to gene expression data and inferred the interactions between genes reported known as psychiatric disorder-associated genes and susceptible genes defined by inferred networks.

Declaration

DK, MK, and JG will pay the publication costs from their research fund.
This article has been published as part of *BMC Medical Genomics* Vol 9 Suppl 2 2016: Selected articles from the IEEE International Conference on Bioinformatics and Biomedicine 2015: medical genomics. The full contents of the supplement are available online at http://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-9-supplement-2.

Availability of data and material

Not applicable.

Authors' contributions

DK conceived the package and wrote the manuscript. MK and AB contributed to data analysis and programming for the experiment. CL generated the experimental data. JG provided overall supervision. All authors reviewed, edited and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Published: 10 August 2016

References

1. Hoheisel JD. Microarray technology: beyond transcript profiling and genotype analysis. Nat Rev Genet. 2006;7:200–10.
2. Madhamshettiwar P, Maetschke S, Davis M, Reverter A, Ragan M. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. Genome Med. 2012;4(5):41. doi:10.1186/gm340.
3. Wang X, Gotoh O. Inference of cancer-specific gene regulatory networks using soft computing rules. Gene Regul Syst Biol. 2010;4:19–34. doi:10.4137/GRSB.S4509.
4. Chen K, Rajewsky N. The evolution of gene regulation by transcription factors and micrornas. Nature Rev Gen. 8:93–103.
5. Bonetta L. Protein-protein interactions: Interactome under construction. Nature. 2010;468:851–4.
6. Tresch A, Beissbarth T, Sultmann H, Kuner R, Poustka A, Buness A. Discrimination of direct and indirect interactions in a network of regulatory effects. J Comput Biol. 2007;14:1217–28.
7. Cheng C, Yan KK, Hwang W, Qian J, Bhardwaj N, Rozowsky J, Lu ZJ, Niu W, Alves P, Kato M, Snyder M, Gerstein M. Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data. PLoS Comput Biol. 2011;7(11):1002190. doi:10.1371/journal.pcbi.1002190.
8. Hecker M, Lambeck S, Toepfer S, Someren Ev, Guthke R. Gene regulatory network inference: data integration in dynamic models-a review. Biosystems. 96:86–103.

9. Zhang S, Li Q, Liu J, Zhou XJ. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules. Bioinformatics. 2011;27(13):401–9. doi:10.1093/bioinformatics/btr206. arxiv http://bioinformatics.oxfordjournals.org/content/27/13/i401.full.pdf+html. Accessed 17 July 2016.
10. Liang J, Han J. Stochastic boolean networks: An efficient approach to modeling gene regulatory networks. BMC Syst Biol. 2012;6(1):113. doi:10.1186/1752-0509-6-113.
11. Xuan N, Chetty M, Coppel R, Wangikar P. Gene regulatory network modeling via global optimization of high-order dynamic bayesian network. BMC Bioinformatics. 2012;13(1):131. doi:10.1186/1471-2105-13-131.
12. de Matos Simoes R, Emmert-Streib F. Bagging statistical network inference from large-scale gene expression data. PLoS ONE. 2012;7(3): 33624. doi:10.1371/journal.pone.0033624.
13. Geeven G, van Kesteren RE, Smit AB, de Gunst MCM. Identification of context-specific gene regulatory networks with gemula–gene expression modeling using lasso. Bioinformatics. 2012;28(2):214–21. doi:10.1093/bioinformatics/btr641. arxiv http://bioinformatics.oxfordjournals.org/content/28/2/214.full.pdf+html. Accessed 17 July 2016.
14. De Smet R, Marchal K. Advantages and limitations of current network inference methods. Nat Rev Immunol. 8:717–29.
15. Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR, Kellis M, Collins JJ, Stolovitzky G. Wisdom of crowds for robust gene network inference. Nature Methods. 2012;9:796–804.
16. Meinshausen N, Buhlmann P. Stability selection. J R Stat Soc Ser B (Stat Method). 2010;72(4):417–73. doi:10.1111/j.1467-9868.2010.00740.x.
17. Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Stat Soc Ser B. 1994;58:267–88.
18. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci. 2000;97(22): 12182–6. doi:10.1073/pnas.220392197. arxiv http://www.pnas.org/content/97/22/12182.full.pdf+html. Accessed 17 July 2016.
19. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007;5(1):8. doi:10.1371/journal.pbio.0050008.
20. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC bioinformatics. 2006;7(Suppl 1):p. S7.
21. Haury AC, Mordelet F, Vera-Licona P, Vert JP. Tigress: Trustful inference of gene regulation using stability selection. BMC Syst Biol. 6:145.
22. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G. Towards a rigorous assessment of systems biology models: The dream3 challenges. PLoS ONE. 2010;5(2): 9202. doi:10.1371/journal.pone.0009202.
23. Pinna A, Soranzo N, de la Fuente A. From knockouts to networks: Establishing direct cause-effect relationships through graph analysis. PLoS ONE. 2010;5(10):12912. doi:10.1371/journal.pone.0012912.
24. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PLoS ONE. 2010;5(9):12776. doi:10.1371/journal.pone.0012776.
25. Gretton A, Herbrich R, Hyvärinen A. Kernel methods for measuring independence. J Mach Learn Res.. 6:2075–129.
26. Cover TM, Thomas JA. Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing): Wiley-Interscience; 2006.
27. Cheadle C, Vawter MP, Freed WJ, Becker KG. Analysis of microarray data using z score transformation. J Mole Diagn. 2003;5(2):73–81.
28. Sacchetti E, Scassellati C, Minelli A, Valsecchi P, Bonvicini C, Pasqualetti P, Galluzzo A, Pioli R, Gennarelli M. Schizophrenia susceptibility and nmda-receptor mediated signalling: an association study involving 32 tagsnps of dao, daoa, ppp3cc, and dtnbp1 genes. BMC Medical Genetics. 2013;14(1):33.
29. Martins-de-Souza D, Gattaz W, Schmitt A, Novello J, Marangoni S, Turck C, Dias-Neto E. Proteome analysis of schizophrenia patients wernicke's area reveals an energy metabolism dysregulation. BMC Psychiatry. 2009;9(1):17.
30. Grube S, Gerchen MF, Adamcio B, Pardo LA, Martin S, Malzahn D, Papiol S, Begemann M, Ribbe K, Friedrichs H, Radyushkin KA, Müller M, Benseler F, Riggert J, Falkai P, Bickeböller H, Nave KA, Brose N, Stühmer W, Ehrenreich H. A cag repeat polymorphism of kcnn3 predicts sk3 channel function and cognitive performance in schizophrenia. EMBO Mole Med. 2011;3(6):309–19.
31. Alkelai A, Greenbaum L, Lupoli S, Kohn Y, Sarner-Kanyas K, Ben-Asher E, Lancet D, Macciardi F, Lerer B. Association of the type 2 diabetes mellitus susceptibility gene, tcf7l2, with schizophrenia in an arab-israeli family sample. PLoS ONE. 2012;7(1):29228.
32. Glaser B, Kirov G, Bray NJ, Green E, O'Donovan MC, Craddock N, Owen MJA. Identification of a potential bipolar risk haplotype in the gene encoding the winged-helix transcription factor rfx4. Mol Psychiatry. 2005;10:920–7.
33. Wu L, Huang Y, Li J, Zhao H, Du H, Jin Q, Zhao X, Ma H, Zhu G. Association study of the fyn gene with schizophrenia in the chinese-han population. Psychiatric genetics. 2013;23:39–40.
34. Kahler AK, Djurovic S, Rimol LM, Brown AA, Athanasiu L, Jonsson EG, Hansen T, Gustafsson O, Hall H, Giegling I, Muglia P, Cichon S, Rietschel M, Pietilainen OPH, Peltonen L, Bramon E, Collier D, Clair DS, Sigurdsson E, Petursson H, Rujescu D, Melle I, Werge T, Steen VM, Dale AM, Matthews RT, Agartz I, Andreassen OA. Candidate gene analysis of the human natural killer-1 carbohydrate pathway and perineuronal nets in schizophrenia: B3gat2 is associated with disease risk and cortical surface area. Biol Psychiatry. 2011;69(1):90–6.
35. Moons T, Claes S, Martens GJM, Peuskens J, Loo KMJV, Schijndel JEV, Hert MD, van Winkel R. Clock genes and body composition in patients with schizophrenia under treatment with antipsychotic drugs. Schizophrenia Research. 2011;125:187–93.
36. Hua P, Liu W, Kuo SH, Zhao Y, Chen L, Zhang N, Wang C, Guo S, Wang L, Xiao H, et al. Association of tef polymorphism with depression in parkinson disease. Mov Disord. 2012;27(13):1694–7.
37. Manchia M, Squassina A, Congiu D, Chillotti C, Ardau R, Severino G, Del Zompo M. Interacting genes in lithium prophylaxis: preliminary results of an exploratory analysis on the role of dgkh and nr1d1 gene polymorphisms in 199 sardinian bipolar patients. Neurosci Lett. 2009;467(2):67–71.
38. Jamain S, Quach H, Fellous M, Bourgeron T. Identification of the human kif13a gene homologous to drosophila kinesin-73 and candidate for schizophrenia. Genomics. 2001;74(1):36–44.
39. Uddin M, Chang S-C, Zhang C, Ressler K, Mercer KB, Galea S, Keyes KM, McLaughlin KA, Wildman DE, Aiello AE, et al. Adcyap1r1 genotype, posttraumatic stress disorder, and depression among women exposed to childhood maltreatment. Depression and anxiety. 2013;30(3):251–258. Wiley Online Library.
40. Li J, Liu J, Feng G, Li T, Zhao Q, Li Y, Hu Z, Zheng L, Zeng Z, He L, et al. The *mdga* 1 gene confers risk to schizophrenia and bipolar disorder. Schizophr Res. 2011;125(2):194–200.
41. Saito T, Papolos DF, Chernak D, Rapaport MH, Kelsoe JR, Lachman HM. Analysis of gnaz gene polymorphism in bipolar affective disorder. AM j medical genetics. 1999;88(4):324–8.
42. Schennach R, Zill P, Obermeier M, Hauer D, Dehning S, Cerovecki A, Opgen-Rhein M, Musil R, Spellmann I, Matz J, et al. The cnr1 gene in depression and schizophrenia-is there an association with early improvement and response? Psychiat Res. 2012;196(1):160.
43. Håvik B, Degenhardt FA, Johansson S, Fernandes CP, Hinney A, Scherag A, Lybæk H, Djurovic S, Christoforou A, Ersland KM, et al. Dclk1 variants are associated across schizophrenia and attentiondeficit/hyperactivity disorder. PloS one. 2012;7(4):35424.
44. Ikeda M, Tomita Y, Mouri A, Koga M, Okochi T, Yoshimura R, Yamanouchi Y, Kinoshita Y, Hashimoto R, Williams HJ, et al. Identification of novel candidate genes for treatment response to risperidone and susceptibility for schizophrenia: integrated analysis among pharmacogenomics, mouse expression, and genetic case-control association approaches. Biological psychiatry. 2010;67(3):263–9.
45. Torri F, Akelai A, Lupoli S, Sironi M, Amann-Zalcenstein D, Fumagalli M, Dal Fiume C, Ben-Asher E, Kanyas K, Cagliani R, et al. Fine mapping of ahi1 as a schizophrenia susceptibility gene: from association to evolutionary evidence. FASEB J. 2010;24(8):3066–82.

46. Sun J, Jia P, Fanous AH, van den Oord E, Chen X, Riley BP, Amdur RL, Kendler KS, Zhao Z. Schizophrenia gene networks and pathways and their applications for novel candidate gene selection. PLoS One. 2010;5(6):11351.
47. Costain G, Lionel AC, Merico D, Forsythe P, Russell K, Lowther C, Yuen T, Husted J, Stavropoulos DJ, Speevak M, et al. Pathogenic rare copy number variants in community-based schizophrenia suggest a potential role for clinical microarrays. Hum Mol Genet. 2013;22(22):4485–4501. Oxford Univ Press.
48. Balan S, Yamada K, Hattori E, Iwayama Y, Toyota T, Ohnishi T, Maekawa M, Toyoshima M, Iwata Y, Suzuki K, et al. Population-specific haplotype association of the postsynaptic density gene dlg4 with schizophrenia, in family-based association studies. PloS one. 2013;8(7):70302.
49. de la Fontaine L, Schwarz MJ, Riedel M, Dehning S, Douhet A, Spellmann I, Kleindienst N, Zill P, Plischke H, Gruber R, et al. Investigating disease susceptibility and the negative correlation of schizophrenia and rheumatoid arthritis focusing on mif and cd14 gene polymorphisms. Psychiatry research. 2006;144(1):39–47.
50. Deng X, Sagata N, Takeuchi N, Tanaka M, Ninomiya H, Iwata N, Ozaki N, Shibata H, Fukumaki Y. Association study of polymorphisms in the neutral amino acid transporter genes slc1a4, slc1a5 and the glycine transporter genes slc6a5, slc6a9 with schizophrenia. BMC psychiatry. 2008;8(1):58.
51. Bharadwaj R, Jiang Y, Mao W, Jakovcevski M, Dincer A, Krueger W, Garbett K, Whittle C, Tushir JS, Liu J, et al. Conserved chromosome 2q31 conformations are associated with transcriptional regulation of gad1 gaba synthesis enzyme and altered in prefrontal cortex of subjects with schizophrenia. The Journal of Neuroscience. 2013;33(29):11839–11851.
52. Arai S, Shibata H, Sakai M, Ninomiya H, Iwata N, Ozaki N, Fukumaki Y. Association analysis of the glutamic acid decarboxylase 2 and the glutamine synthetase genes (gad2, glul) with schizophrenia. Psychiatric genetics. 2009;19(1):6–13.
53. Tsai SJ, Hong CJ, Liou YJ, Liao DL. Association study of got2 genetic polymorphisms and schizophrenia. Psychiatric genetics. 2007;17:314.
54. Herrmann R, Lee B, Arshavsky VY. Rgs9 knockout causes a short delay in light responses of on-bipolar cells. PloS ONE. 2011;6(11):27573.
55. Lang B, Alrahbeni TMA, St Clair D, Blackwood DH, et al. Hdac9 is implicated in schizophrenia and expressed specifically in post-mitotic neurons but not in adult neural stem cells. Am J stem cells. 2012;1(1):31.
56. Soronen P, Ollila H, Antila M, Silander K, Palo O, Kieseppä T, Lönnqvist J, Peltonen L, Tuulio-Henriksson A, Partonen T, et al. Replication of gwas of bipolar disorder: association of snps near cdh7 with bipolar disorder and visual processing. Mole psychiatry. 2010;15(1):4–6.
57. Chang Y-H, Lee S-Y, Chen S-L, Tzeng N-S, Wang T-Y, Lee H-I, Chen PS, Huang S-Y, Yang YK, Ko H-C, et al. Genetic variants of the bdnf and drd3 genes in bipolar disorder comorbid with anxiety disorder. J affective disorders. 2013;151(3):967–972. Elsevier.
58. Yamada K, Iwayama Y, Hattori E, Iwamoto K, Toyota T, Ohnishi T, Ohba H, Maekawa M, Kato T, Yoshikawa T. Genome-wide association study of schizophrenia in japanese population. PloS one. 2011;6(6):20468.