Zhang *et al. BMC Bioinformatics* (2019) 20:12
https://doi.org/10.1186/s12859-018-2560-0

BMC Bioinformatics

SOFTWARE Open Access

GeNeCK: a web server for gene network construction and visualization

Minzhe Zhang[1,2], Qiwei Li[1,2], Donghyeon Yu[3], Bo Yao[1,2], Wei Guo[4], Yang Xie[1,2,5] and Guanghua Xiao[1,2,5*]

Abstract

Background: Reverse engineering approaches to infer gene regulatory networks using computational methods are of great importance to annotate gene functionality and identify hub genes. Although various statistical algorithms have been proposed, development of computational tools to integrate results from different methods and user-friendly online tools is still lagging.

Results: We developed a web server that efficiently constructs gene networks from expression data. It allows the user to use ten different network construction methods (such as partial correlation-, likelihood-, Bayesian- and mutual information-based methods) and integrates the resulting networks from multiple methods. Hub gene information, if available, can be incorporated to enhance performance.

Conclusions: GeNeCK is an efficient and easy-to-use web application for gene regulatory network construction. It can be accessed at http://lce.biohpc.swmed.edu/geneck.

Keywords: Gene network, Gene network, Statistical method, Web server, Correlation, Likelihood, Bayesian, Mutual information, Ensemble, Hub gene, Visualization

Background

A gene regulatory network (GRN) describes biological interactions among genes and provides a systematic understanding of cellular signaling and regulatory processes. It depicts how a set of genes interact with each other to form a functional module and how different gene modules are related. A typical GRN approximates a scale-free network topology with a few highly connected genes (i.e. hub genes) and many poorly connected nodes [1]. These hub genes are master regulators in a gene network, and usually play essential roles in a biological system. Investigations of GRN can facilitate the systematic functional annotation of genes [2] and help identify the hub genes, which may lead to potential clinical applications [3].

Reverse engineering approaches to construct gene networks from transcriptomic data have greatly facilitated biomedical research. Statistical methods proposed for inferring network structure can be categorized into four classes: 1) probabilistic network-based approaches, mainly Bayesian networks (BN); 2) correlation-based methods; 3) partial correlation-based methods; and 4) information theory-based methods [4]. Comparative evaluation among different methods for constructing large scale GRNs revealed the strengths and weaknesses of each method with respect to different scenarios, with no single method outperforming others universally [5]. An ensemble-based network aggregation (ENA) method was proposed to integrate different methods to improve the accuracy of network inference [6]. Recent advancements in statistical methods have extended algorithms to incorporate prior knowledge of hub genes [7]. Besides above statistical methods that aim to infer the latent covariance matrix of all the components in a graph using gene expression data, other algorithms like Petri Nets [8] and ordinary differential equations (ODE) [9] focus more on simulating the dynamics of specific pathways that involve important disease genes.

Despite the development of various computational methods and corresponding R packages for inferring

*Correspondence: Guanghua.Xiao@UTSouthwestern.edu
[1] Quantitative Biomedical Research Center, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas 75390, TX, United States
[2] Department of Clinical Sciences, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, Texas, United States
Full list of author information is available at the end of the article

gene-gene interactions, implementation of those algorithms with graphical interface is still lagging. CoExpNetViz [10] is an online tool developed for constructing co-expression networks in plant research, but its application is limited by simple statistics and compulsory "bait" genes input. To provide easy accessibility for the network construction tool, we introduce a web server called GeNeCK (Gene Network Construction Tool Kit, see Fig. 1) which allows users to upload their own gene expression data and choose their preferred method to infer and visualize the network, as well as integrate different methods to obtain a more confident result.

Implementation

GeNeCK is a web server (http://lce.biohpc.swmed.edu/geneck) with a user-friendly graphical interface. A quick user guide on how to upload data and submit jobs is provided on the website and in the supplementary material (Additional file 1: Figure S9). GeNeCK offers the flexibility for experienced users to select methods and set preferred parameters. Using ENA is more straightforward for most users since it generally performs well in all scenarios, does not require choosing tuning parameters, and can provide a p-value for each connection, which indicates the statistical significance of the connection. The constructed network will be displayed on the website once the job is finished (Fig. 1). Genes with a high degree of connection (i.e. hub genes) will be plotted with different colors. Users can interactively explore the constructed network. Clicking on a specific gene will highlight the gene itself along with its connected neighbors, and the corresponding information will be displayed at the bottom (Fig. 1).Although the current version of GeNeCK does not provide a function for users to download the figure, users can use screenshot software tools to get the figure for the network structure. We recommend that users download and import the constructed network structure into other visualization tools, such as Cytoscape, for further visualization and analysis (Additional file 2: Figure S10).

Methods

GeNeCK allows users to construct network using 11 diffe rent methods (summarized in Additional file 3: Table S1). Readers can refer to Yu et al. [7] for a comprehensive review of the different network construction methods.

Network inference methods

Partial correlation-based methods calculate the inverse covariance matrix $\mathbf{\Omega}$ (also known as the precision matrix) of gene expressions, in which $\omega_{j,h} = 0$ indicates gene j and h given the expressions of all the other genes is conditional independent. GeneNet [11] employs Moore-Penrose pseudoinverse and bootstrap methods to obtain a shrink estimate of $\mathbf{\Omega}$. Meinshausen and Bühlmann [12] proposed the neighborhood selection (NS) method, which converts the precision matrix estimation problem to a regression problem by fitting a LASSO to each gene using others as predictors. Sparse partial correlation estimation (SPACE) is a joint spare regression model developed by Peng et al. [13], which resolves a symmetrically constrained and L_1-regularizated regression problem under high-dimensional settings.

Likelihood-based approaches, such as graphical LASSO (GLASSO [14]) and GLASSO with a reweighted strategy for scale-free networks (GLASSO-SF [15]), optimize a penalized maximum likelihood function to estimate $\mathbf{\Omega}$. Bayesian graphical LASSO (BayesianGLASSO [16]) is a fully Bayesian treatment of GLASSO that uses a double exponential prior and employs a block Gibbs sampler for exploring the posterior distribution.

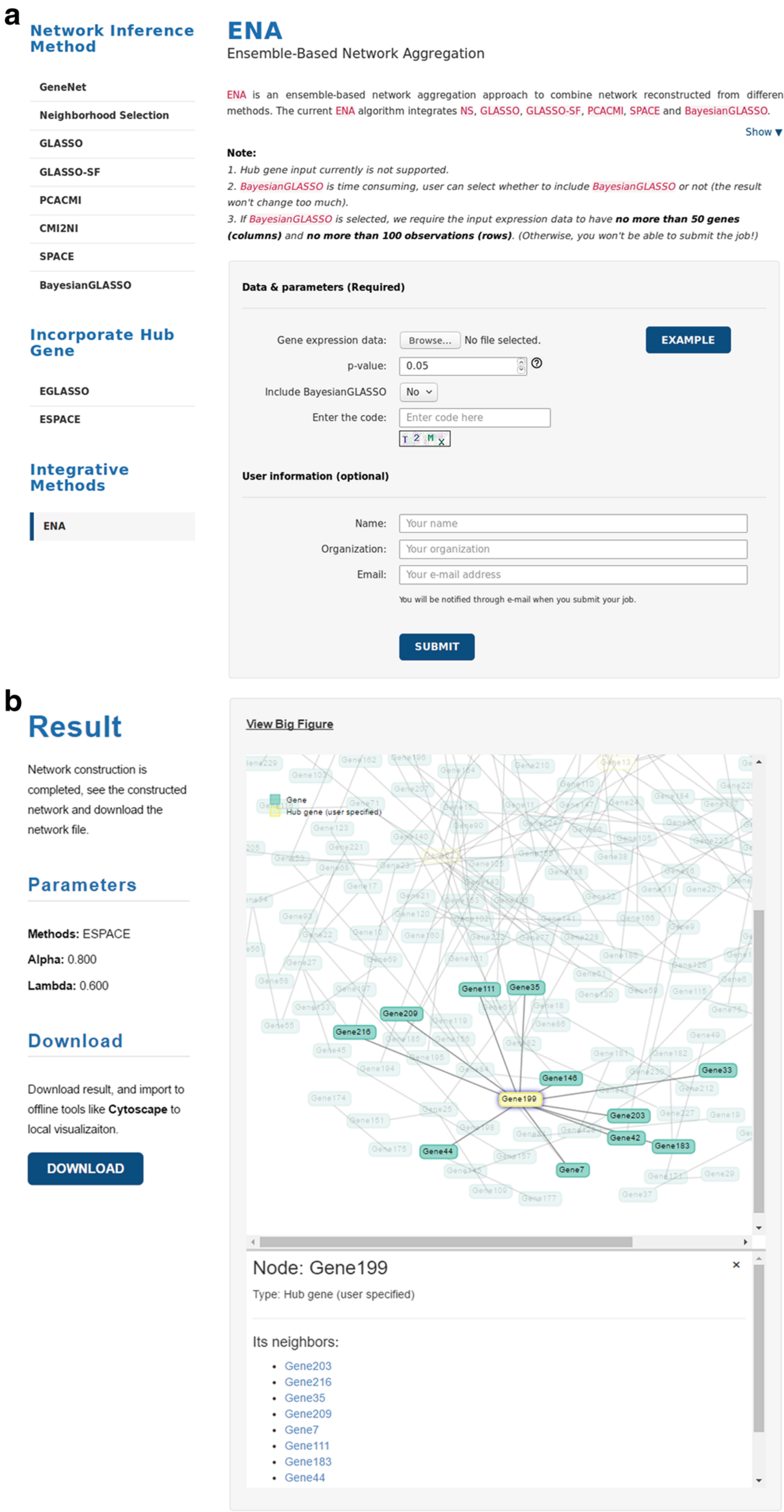
Mutual information (MI) is a measure in information theory of pairwise dependency between two variables. Zhang et al. [17] proposed a path consistency algorithm based on conditional mutual information (PCACMI) to infer graphical structure, and further conditional mutual inclusive information-based network inference (CMI2NI [18]) method that improves the PCACMI method.

Hub gene incorporation

Gene networks usually have scale-free characteristics. In other words, there are usually a few hub genes regulating many others. In practice, most of such hub genes in biological pathways have been well studied and validated through biological experiments. To properly incorporate this prior knowledge, Yu et al. [7] proposed extended sparse partial correlation estimation (ESPACE) and extended graphical LASSO (EGLASSO) methods. In these methods, during the covariance estimation of original SPACE and GLASSO methods, hub gene information can be incorporated to improve the network inferences.

Network integration

An ensemble-based network aggregation (ENA) method [6] combines networks reconstructed from different methods. The original ENA algorithm does not report the confidence level of estimated edges. To derive the p-value of an edge between a pair of genes, we adapted ENA by implementing an additional permutation step to generate the distribution of null hypothesis. We first permute the given gene expression dataset to obtain a resampled dataset $D^{(m)}$. Then we implement the ENA algorithm to get the ensemble rank matrix $\tilde{R}^{(m)}$ for this dataset. This procedure is repeated M times. The empirical null distribution F^{null} of all possible pairwise connection for p genes can be obtained based on all the harmonic means in the M permutations, i.e. $\left\{\tilde{r}_{jh}^{(m)}, m = 1, ..., M, 1 \leq j < h \leq p\right\}$. Then the p-value of the estimated edge between gene j and h is approximated by the quantile of $\tilde{r}_{jh}$ in the null



ENA

Ensemble-Based Network Aggregation

ENA is an ensemble-based network aggregation approach to combine network reconstructed from different methods. The current ENA algorithm integrates NS, GLASSO, GLASSO-SF, PCACMI, SPACE and BayesianGLASSO.

Note:

1. Hub gene input currently is not supported.

2. BayesianGLASSO is time consuming, user can select whether to include BayesianGLASSO or not (the result won't change too much).

3. If BayesianGLASSO is selected, we require the input expression data to have ***no more than 50 genes (columns)*** *and* ***no more than 100 observations (rows)****. (Otherwise, you won't be able to submit the job!)*

Data & parameters (Required)

Gene expression data: No file selected.

p-value: 0.05

Include BayesianGLASSO: No

Enter the code: Enter code here

T 2 M X

User information (optional)

Name: Your name

Organization: Your organization

Email: Your e-mail address

You will be notified through e-mail when you submit your job.

Result

Network construction is completed, see the constructed network and download the network file.

Parameters

Methods: ESPACE

Alpha: 0.800

Lambda: 0.600

Download

Download result, and import to offline tools like **Cytoscape** to local visualizaiton.

Node: Gene199

Type: Hub gene (user specified)

Its neighbors:

- Gene203
- Gene216
- Gene35
- Gene209
- Gene7
- Gene111
- Gene183
- Gene44

Fig. 1 a Web interface of GeNeCK analysis page. **b** Visualization of constructed network in GeNeCK results page distribution F^{null} with Benjamini-Hochberg adjustment [19] to avoid multiple comparison problems.

$$D \xrightarrow{\text{permutate}} \left\{ \begin{array}{ccc} D^{(1)} & \xrightarrow{\text{ENA}} & \tilde{R}^{(1)} \\ \vdots & & \vdots \\ D^{(M)} & \xrightarrow{\text{ENA}} & \tilde{R}^{(M)} \end{array} \right\} \rightarrow F^{\text{null}},$$

$$p - value(jh) = BHadjust\left(\frac{\#\text{ of } \tilde{r}_{jh} \le \text{ permutated } r \text{ value in } F^{\text{null}}}{\text{Total }\#\text{ of } \tilde{r}_{jh} \le \text{ permutated } r \text{ value in } F^{\text{null}}}\right).$$

In the simulation studies, we ensembled the networks constructed by NS, GLASSO, GLASSO-SF, PCACMI, SPACE, and BayesianGLASSO. GeneNet and CMI2NI were excluded because GeneNet performed the worst in all the scenarios (Additional file 4: Figure S1-S8) and CMI2NI produced the exact same results as PCACMI in default settings. We run all the processes in a single node of UT Southwestern BioHPC cluster (Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30GHz, 32GB RAM).

Results

To comprehensively evaulate different models, we simulated co-expression data from four real protein-protein interaction networks (Fig. 2) used in Allen et al. [5], which was selected Keshava Prasad et al. [20]. See the download link for the four real network structure in the Availability of data and materials section. Details of the generative model are discussed below. We investigated the performance of each method for data with various noise levels and sample sizes.

Generative model

We used Gaussian graphical models that are mainly used to infer the gene association network to simulate expression data. Let $\mathbf{y}_i = (y_{i1}, \ldots, y_{ij}, \ldots, y_{ip})$ denotes the collection of expression levels for each gene observed in sample i. This was simulated from a zero-mean multivariate normal distribution $y_i = \text{MN}\left(\mathbf{0}_p,\ \mathbf{\Sigma} + \epsilon^2\mathbf{I}_{p\times p}\right)$, where $\mathbf{0}_p$ denotes the p-dimension zero vector and $\mathbf{I}_{p\times p}$ denotes the p-by-p identity matrix. For the covariance matrix $\mathbf{\Sigma}$,

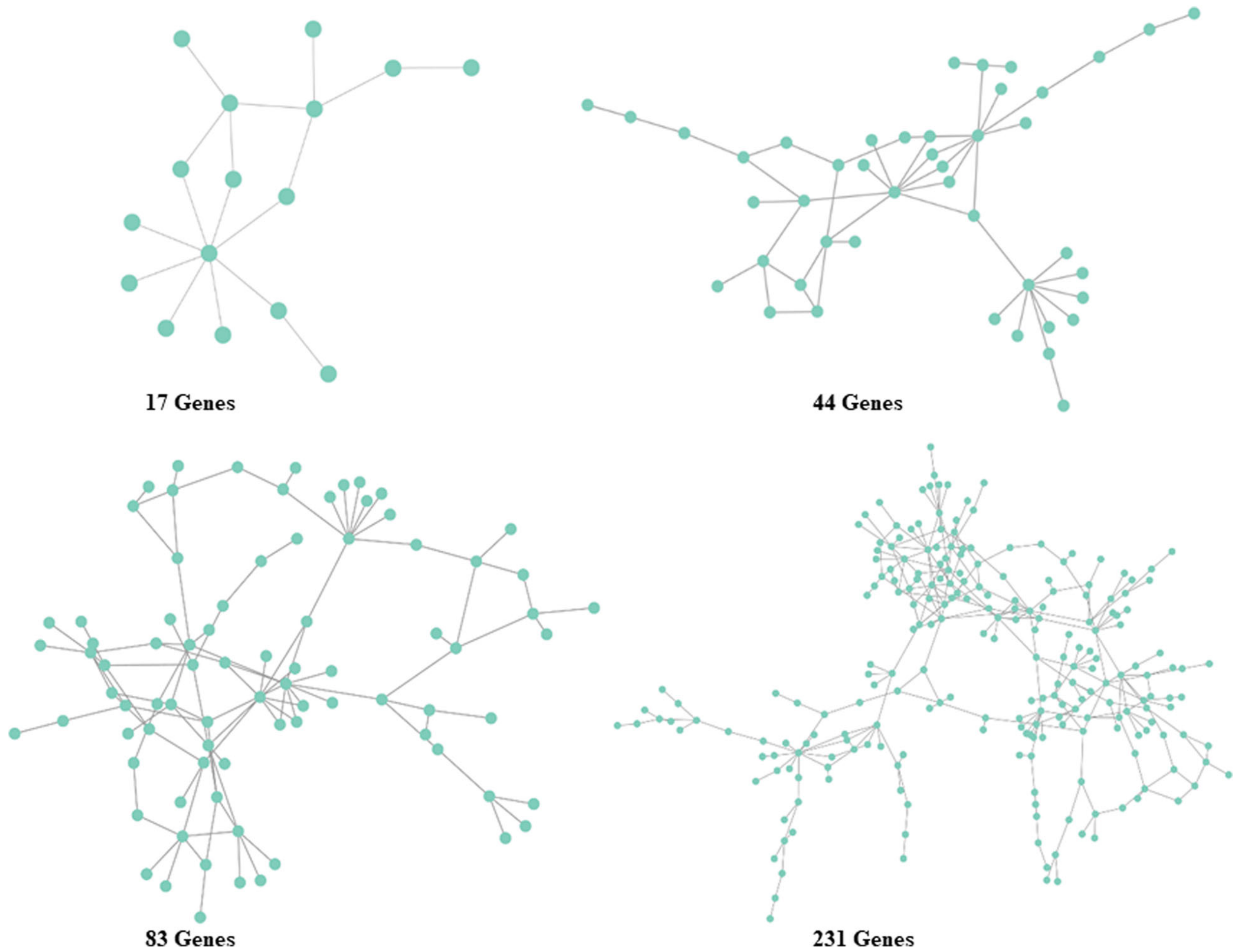


Fig. 2 The four real protein-protein interaction networks used in the simulation study

we generated its concentration matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ following Peng, et al. [13]. The initial matrix $\mathbf{\Omega}$ was created by setting

$$\omega_{jh} = \begin{cases} 1 & , \ j = h \\ 0 & , \ j \neq h, j \nsim h \ , \\ 0.5\text{Uniform}(-1, -0.5) + 0.5\text{Uniform}(0.5, 1) & , \ j \neq h, j \sim h \end{cases}$$

where $Uniform(a, b)$ represents uniform distribution on interval (a, b), $j \sim h$ indicates that there is an edge between gene *j* and $h, j \nsim h$ means otherwise. The network structure was chosen from one of the four real protein-protein interaction networks [20, 21], each of which was approximately scale-free (see Fig. 2). Then, the non-zero elements in $\mathbf{\Omega}$ were rescaled to assure positive definiteness. Specifically, for each row, we first summed the absolute values of the off-diagonal elements, and divided each off-diagonal entry by 1.5-fold their sum. Next, we averaged this rescaled matrix with its transpose to ensure symmetry. We then set 0.1 to those non-zero entries with absolute value smaller than 0.1. After that, the inverse of the final matrix was denoted by $\mathbf{A} = \mathbf{\Omega}^{-1}$. Each element in the covariance matrix $\mathbf{\Sigma}$ was determined by $\delta_{jh} = \alpha_{jh}/\sqrt{\alpha_{jj}\alpha_{hh}}$. For the noise level ϵ, we considered three cases: $\epsilon = 0$, 0.1, 0.5.

Performance metric

We evaluated the result of each method by plotting its operating characteristic curve (ROC) and calculating the area under the ROC curve (AUC). As different methods generate different outputs, we used their corresponding approaches to plot ROC curves for a fair comparison. GeneNet and BayesianGLASSO yield a continuous estimate of each partial correlation ρ_{jh}. They do not require a tuning parameter. Thus, an edge between gene *j* and *h* was determined if the absolute value of ρ_{jh} was greater than a certain threshold. Then the ROC curves were obtained by plotting false positive rates (FPRs) against true positive rates (TPRs) under different thresholds. For mutual information-based methods, we choose the tuning parameter $\alpha = 0.03$ as suggested by the authors [17, 18]. Then, an edge between gene *j* and *h* was determined if the estimated entropy was greater than a threshold. The ROC curves were obtained by plotting FPRs against TPRs under different thresholds. Note that we only included PCACMI in the simulation, since CMI2NI produced the same result as PCACMI did. For the other methods that need a tuning parameter, the ROC curves were obtained by plotting FPRs and TPRs under different choices of the tuning parameter.

Result summarization

As shown in the result of simulation study (Additional file 4: Figure S1-S8), BayesianGLASSO and ENA generally outperform other methods, which is consistent with the literature [6, 16]. Besides, mutual information-based methods also show competitive results. NS, GLASSO, and GLASSO-SF, which share the same strategy, have similar accuracy. As the earliest developed method, GeneNet has significantly lagged performance. Not surprisingly, all methods lose power when either a higher level of noise manifests or a smaller number of samples is generated.

We also logged the computational time of each method in Table S2 (Additional file 5). The Bayesian method consumed several orders of magnitude more time, and it soon went beyond real applicability when the number of genes in the network increased to hundreds. Most other methods shared similar efficacy in the simulation settings, with mutual information-based methods being a little slower.

Discussion

GeNeCK infers a gene-gene connection based on the expression pattern of the two genes. It can provide a hint of their potential functional relationship, but does not necessarily imply a real biological interaction. One should be very cautious when interpreting the result, especially when the tuning parameter is out of a reasonable range (e.g. an almost fully connected network may be a sign of choosing a problematic parameter value). As different methods use different measurements to evaluate the confidence of estimated edges (e.g. partial correlation, mutual information), this may not be easy to interpret for users with little statistical background. We suggest users choose the ENA method, which outputs p-values to indicate the significance of gene-gene connections. More importantly, it generally achieves the best performance. For extended methods (EGLASSO and ESPACE) that allow for the “hub genes” specification, additional attention needs to be paid when choosing the value for the confidence index α. The α value can be selected by different statistical methods, such as the generalized information criterion (GIC) [22]. In practice, we suggest an initial try with no or a very weak prior brief to see if the genes of interest are picked up by the algorithm. Usually a very small α value is not desired, as the influence of hub genes should already be presented in the data if the prior information is correct. Otherwise this can lead to a biased result.

Conclusion

Reconstructions of gene networks from gene expression data greatly facilitate our understanding of underlying biological mechanisms and provide new opportunities for drug and biomarker discoveries. GeNeCK, the online tool kit presented in this paper, enables us to integrate various statistical methods to construct gene networks based on gene expression data. Furthermore, the information of hub genes, which usually play an essential role in gene regulation and biological processes, could be incorporated into GeNeCK to improve the performance of the related methods. It is believed that the tool will cater to a wide audience in the field of biology.

Availability and requirements

Project name: GeNeCK
Project home page: http://lce.biohpc.swmed.edu/geneck/
Operating systems: Windows, Linux and Mac
Programming language: PHP, HTML, JavaScript and R
License: GPL

Additional files

Additional file 1: Figure S9. GeNeCK user guide. A simple tutorial on how to run GeNeCK. (DOCX 195 kb)

Additional file 2: Figure S10. External visulization of GeNeCK inference result. Example of how to import GeNeCK output to Cytoscape for enhanced visulization. (DOCX 326 kb)

Additional file 3: Table S1. Summary of basic information of different methods in GeNeCK. (DOCX 14 kb)

Additional file 4: Figure S1-S8. Comparison of model performance of different methods in simulation studies. Network structures are based on real protein-protein interaction networks. Expression data are simulated under different noise levels. (DOCX 776 kb)

Additional file 5: Table S2. Summary of runtime of different methods in GeNeCK. (DOCX 18 kb)

Abbreviations

AUC: Area under curve; BayesianGLASSO: Bayesian graphical LASSO; CMI2NI: Conditional mutual inclusive information-based network inference; EGLASSO: Extended GLASSO; ENA: Ensemble-based network aggregation; ESPACE: Extended SPACE; GeNeCK: Gene network construction tool kit; GLASSO: Graphical LASSO; GLASSO-SF: GLASSO with reweighted strategy for scale-free network; GRN: Gene regulatory network; MI: Mutual information; NS: Neighborhood selection; PCACMI: Path consistency algorithm based on conditional mutual information; ROC: Operating characteristic curve; SPACE: Sparse partial correlation estimation

Acknowledgments

The authors would like to thank Jessie Norris for helping us in the manuscript.

Funding

This work was supported by the National Institutes of Health [1R01CA172211, 5P50CA070907 and 1R01GM115473], the National Research Foundation of Korea [NRF-2018R1C1B6001108], and the Cancer Prevention and Research Institute of Texas [RP120732]. The funding bodies had no role in the design, collection, analysis, or interpretation of data in this study.

Availability of data and materials

The adjacency matrices corresponding to the four real protein-protein interaction networks, and all the simulated datasets generated based on the four real protein-protein interaction networks used in the simulation study have been deposited in Figshare (https://figshare.com/projects/GeNeCK/36035).

Authors' contributions

MZ have constructed the web server. QL and MZ have collaborated in the simulation study. DY have contributed to the review of different methods. BY and WG have contributed to network visulization of web server. YX and GX have conceived the study and supervised the web application development and the statistical analyses. All authors have contributed to the writing of the manuscript. All authors have read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Quantitative Biomedical Research Center, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas 75390, TX, United States. [2]Department of Clinical Sciences, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, Texas, United States. [3]Department of Statistics, Inha University, Incheon, South Korea. [4]BioHPC team, Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas, Texas, United States. [5]Harold C. Simmons Cancer Center, University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd., Dallas 75390, Texas, United States.

Received: 8 July 2018 Accepted: 5 December 2018
Published online: 07 January 2019

References

1. Barabasi A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5(2):101.
2. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. 2003;34(2):166.
3. Tang H, Xiao G, Behrens C, Schiller J, Allen J, Chow C-W, Suraokar M, Corvalan A, Mao J, White MA, et al. A 12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients. Clin Cancer Res. 2013;19(6):1577–86.
4. Bansal M, Belcastro V, Ambesi-Impiombato A, Bernardo DD. How to infer gene networks from expression profiles. Mol Syst Biol. 2007;3(1):78.
5. Allen JD, Xie Y, Chen M, Girard L, Xiao G. Comparing statistical methods for constructing large scale gene networks. PloS ONE. 2012;7(1):29348.
6. Zhong R, Allen JD, Xiao G, Xie Y. Ensemble-based network aggregation improves the accuracy of gene network reconstruction. PloS ONE. 2014;9(11):106319.
7. Yu D, Lim J, Wang X, Liang F, Xiao G. Enhanced construction of gene regulatory networks using hub gene information. BMC Bioinformatics. 2017;18(1):186.
8. Rohr C, Marwan W, Heiner M. Snoopy—a unifying petri net framework to investigate biomolecular networks. Bioinformatics. 2010;26(7):974–5.
9. Kim J. Validation and selection of ode models for gene regulatory networks. Chemometr Intell Lab Syst. 2016;157:104–10.
10. Tzfadia O, Diels T, Meyer SD, Vandepoele K, Aharoni A, de Peer YV. Coexpnetviz: comparative co-expression networks construction and visualization tool. Front Plant Sci. 2016;6:1194.
11. Schäfer J, Strimmer K. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol. 2005;4(1):1175–1189.
12. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. Ann Stat. 2006;1436–62.
13. Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. J Am Stat Assoc. 2009;104(486):735–46.
14. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9(3):432–41.
15. Liu Q, Ihler A. Learning scale free networks by reweighted l1 regularization. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics; 2011. p. 40–48.
16. Wang H, et al. Bayesian graphical lasso models and efficient posterior computation. Bayesian Anal. 2012;7(4):867–86.
17. Zhang X, Zhao X-M, He K, Lu L, Cao Y, Liu J, Hao J-K, Liu Z-P, Chen L. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. Bioinformatics. 2011;28(1):98–104.

18. Zhang X, Zhao J, Hao J-K, Zhao X-M, Chen L. Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. Nucleic Acids Res. 2014;43(5):31.
19. Benjamini, Yoav, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodological),. 1995289–300.
20. Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, et al. Human protein reference database—2009 update. Nucleic Acids Res. 2008;37(suppl_1): 767–72.
21. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TKB, Chandrika KN, Deshpande N, Suresh S, et al. Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res. 2004;32(suppl_1):497–501.
22. Yu D, Son W, Lim J, Xiao G. Statistical completion of a partially identified graph with applications for the estimation of gene regulatory networks. Biostatistics. 2015;16(4):670–85.