A greedy-network-based approach for human disease module identification

Meng Jin[1, a], Zhiyuan Yang[1, b], Jianwei Lu*[1, 2, c], Tianwei Yu*[3, d]

[1] School of Software Engineering, Tongji University, Shanghai, 201800, China

[2]Advanced Institute of Translational Medicine, Tongji University, Shanghai, 200082, China

[3]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, 30328, United States of America

[a] jinm1992@163.com, [b] yzhuiyi@126.com, [c] jwlu33@gmail.com, [d] tianwei.yu@emory.edu

Keywords: Gene expression, Biological networks, Greedy algorithm, Machine learning, Cancer biology

Abstract. The accurate classification of disease module from gene expression profiles is quite challenging for new biomarkers because of high noise in gene expression measurements and the small sample size [1]. Studies have shown that network-based gene selection is more reliable than individual genes. Because genes related with same or similar disease modules usually reside in the same vicinity of the molecular network [3]. Based on this theory, we propose a greedy-network-based approach for gene identification. In our study, we use this method in a pediatric acute lymphoblastic leukemia (ALL) [4] dataset and a triple-negative breast cancer (TNBC) microarray dataset. The results show our method achieves higher accuracy in the identification of gene makers.

Background

Dys-regulations of biological networks may be the cause of some complex human diseases, such as diabetes, autism and cancer. The normal genetic analysis approaches only focus on individual genetic determinants, which is not straight-forward to describe the network architecture of complex diseases. Thus it would be necessary to gain a comprehensive understanding of how cells using comprehensive genomic, proteomic and environmental information to produce specific cellular functions and how such functions are operated in the disease state if researchers want to create an effective therapy [2]. Recently some studies show that if applied within a network biology framework into some -omics technologies, such as metabolomics, transcriptomics and proteomics, they may have the potential to provide insights into complex disease pathogenesis and heterogeneity. In conclusion, applying system biology and network scientific methods to human disease can create a new field called “Network medicine” [5], and nowadays this new field is developing rapidly.

In the past decade, researchers have done massive work to identify differentially expressed genes in different phenotypes. It can be used as diagnostic markers to classify different disease states or predict clinical outcomes. However, only using the expression data to define genetic markers are unreliable. In order to make up for this shortage, many researchers began to study network medicine in order to obtain a comprehensive understanding of the complex disease process. Comparing to studying individual genes, mapping human disease related genes to interaction data could help researchers understand human disease mechanisms more deeply and comprehensively. Network based methods have a variety of potential biological and clinical applications, including a better understanding of disease genes and pathogenic pathways. In turn, understanding the pathogenesis can also help staff develop more effective drugs. More reliable biomarkers are likely to play a positive role in monitoring the integrity of the network's function affected by the disease.

In recent years, identifying significant sub-network markers with machine learning [11] or data mining strategies has been developed to identify human disease modules. However, most

Copyright © 2016, the Authors. Published by Atlantis Press.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

approaches build sub-networks randomly which barely include a formal topological structure. In this study, we developed a new method based on greedy network search, which is a concept in social network but is also applicative in gene network. The innovation of our approach lies in each center nodes in the sub-networks is the optimal. Experiments show that our method achieves higher accuracy in the identification of gene makers.

Overview of the algorithm

The algorithm takes the gene network and gene expression data as input (Figure 1). Gene network data contains biological network information that can be a gene regulatory network, a signaling pathway network, or a protein-protein interaction network. The gene expression data contains each genetic locus' expression data on each sample as well as a certain biological or clinical outcome of each sample. The outcome of this algorithm is the selected sub-networks and the scores of each network that measure the correlation of the disease.

The workflow of the algorithm is showed in Figure 1. The algorithm first scans through all the genes in the network. Each node with all its neighbors is grouped by as a sub-network. Then we calculate scores of each sub-network. The second step is to scan all the neighbors in each sub-network, adding them into the center of each sub-network, thus the center of the new sub-networks contains two genes and the neighbors is the union set of both center genes' neighbors. Then the scores of each new sub-network are evaluated as the first step. Only the best-performed sub-networks will be reserved to go on next step, as the aim of this algorithm is to find the most likely pathogenic genes. Beyond that, this filtering operation can also improve the algorithm efficiency. The second step iteratively goes until each sub-network's score no longer increases. There is no doubt that the higher scores are, the stronger the correlation of disease is. We can even assume that the genes in the sub-networks with the highest score are the disease-causing genes.

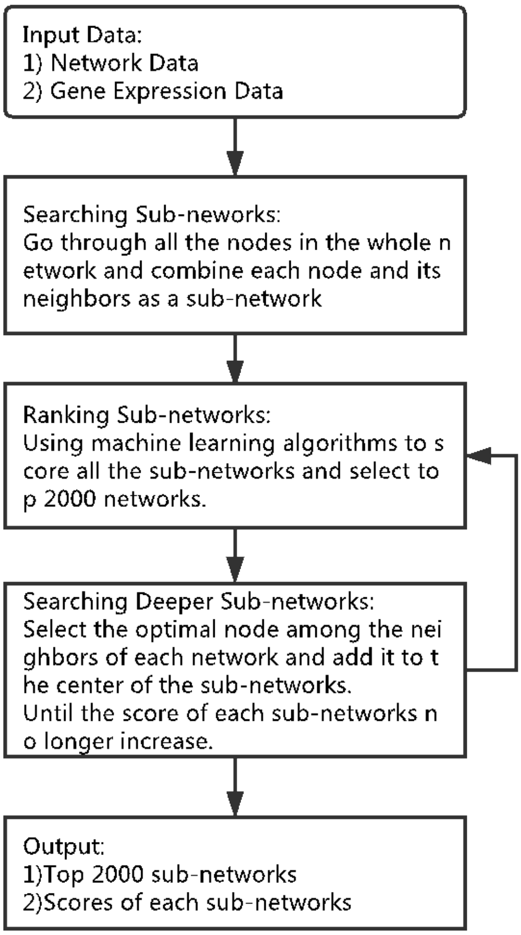
Input Data:
1) Network Data
2) Gene Expression Data
Searching Sub-neworks:
Go through all the nodes in the whole n etwork and combine each node and its neighbors as a sub-network
Ranking Sub-networks:
Using machine learning algorithms to s core all the sub-networks and select to p 2000 networks.
Searching Deeper Sub-networks:
Select the optimal node among the nei ghbors of each network and add it to t he center of the sub-networks.
Until the score of each sub-networks n o longer increase.
Output:
1)Top 2000 sub-networks
2)Scores of each sub-networks

Fig.1. Workflow of the algorithm

Machine learning algorithm

To evaluate the capability of each sub-network to predict the clinical outcome, a machine-learning method needs to be chosen [1]. In this research, SVM is selected to actualize cross-examination of each sub-network.

Support vector machine [6,10] (SVM) based on statistical learning theory is a method of data mining, which can successfully deal with regression (e.g. time series analysis) and pattern recognition (e.g. classification problem, discriminant analysis) problems. SVM can also promote in prediction, comprehensive evaluation and other fields and disciplines. The mechanism of SVM is to find an optimal separating hyper-plane satisfying the requirement of classification, in order to maximize the blank area on both sides of the hyper-plane. In theory, support vector machine can realize optimal classification for linear separable data.

Take two classes for example, given the training sample set: $(y_1, x_1), \dots\dots, (y_l, x_l)$, the optimal hyperplane is $\boldsymbol{w_0} \cdot \boldsymbol{x} + b_0 = 0$. In order to classify all samples correctly and to have the classification interval, constraints need to be followed: $y_i(\boldsymbol{x}_i \cdot \boldsymbol{w} + b) \geq 1,\ i = 1, \dots, l$. To solve this constrained optimization problem, we use a standard optimization technique. We construct a Lagrangian:

$$L(\boldsymbol{w}, b, \boldsymbol{\Lambda}) = \frac{1}{2}\boldsymbol{w} \cdot \boldsymbol{w} - \sum_{i=1}^{l} \alpha_i \, [y_i(\boldsymbol{x}_i \cdot \boldsymbol{w} + b) - 1],$$

where $\Lambda^T = (\alpha_i, \dots, \alpha_l)$ is the vector of non-negative Lagrange multipliers corresponding to the constraints. Calculating the partial derivative of Lagrangian, we can obtain:

$$W(\boldsymbol{\Lambda}) = \boldsymbol{\Lambda}^T \boldsymbol{1} - \frac{1}{2}\boldsymbol{\Lambda}^T \boldsymbol{D} \boldsymbol{\Lambda},$$

where **1** is an l-dimensional unit vector, and $\boldsymbol{D}$ is a symmetric $l \times l$ matrix with elements $D_{ij} = y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j$.

Results

In this study, we compared the results of our method with Yang's method [1]. Figure 2 shows the scores' distribution graphs when using these two methods on ALL. And figure 3 is the scores' distribution graphs on TNBC. Greedy-based network performs better than EGONET on both diseases.

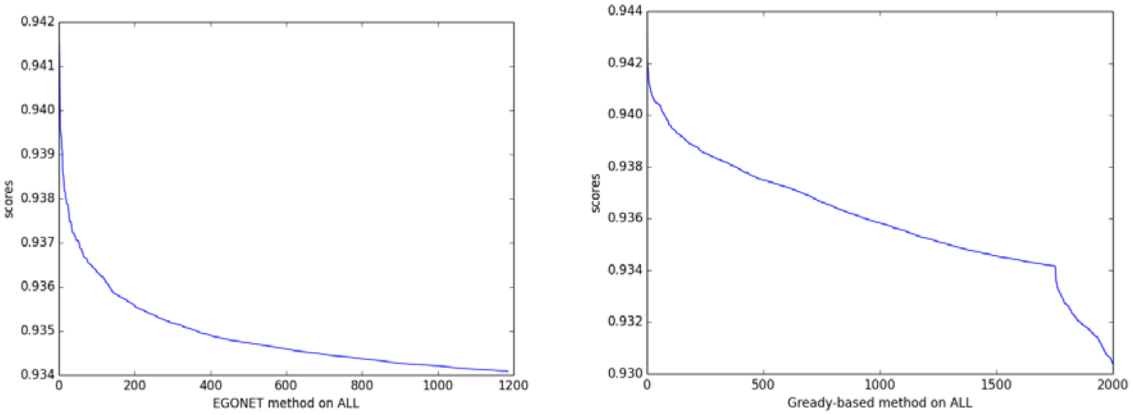

Fig. 2. Line graph of scores using EGONET and Greedy-based method on ALL

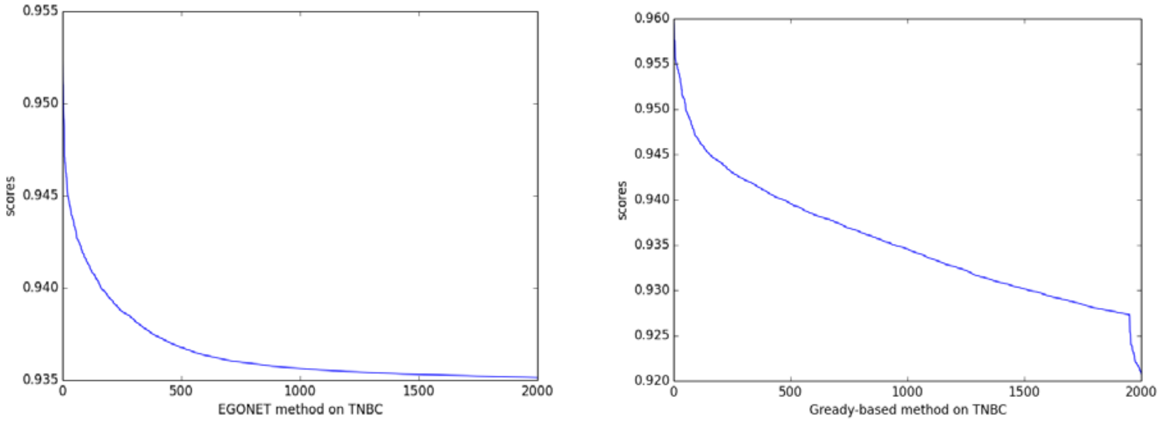

Fig. 3. Line graph of scores using EGONET and Greedy-based method on TNBC

Table 1 is the top 20 genes for classifying ALL (Acute Lymphoblastic Leukemia) patients based on gene ranking metric. For the genes in table 1, there have been literatures reporting EGFR signaling pathways to be related to cancer [9]. EGFR is a member of the epidermal growth factor receptor (HER) family. Both ERBB2 and ERBB3 are HER family members as well, and they are confirmed to have relationship with cancer [10]. Another research, carried out by Michelle L et al, revealed that BCR-ABL fusion gene is one of the main features of Precursor B cell acute lymphoblastic leukemia [11]. What's more, other researches show KIT, GAB1 and JAK2 are the genetic locis closely related to ALL [12-14].

Table 1. Top 20 genes select by greedy-network-based approach of ALL

EntrezID	GeneName	M-value
1956	EGFR	21
2064	ERBB2	17
2065	ERBB3	16
25	ABL1	16
27	ABL2	15
2885	GRB2	15
11184	MAP4K1	15
6655	SOS2	15
3815	KIT	13
2549	GAB1	13
3717	JAK2	12
4233	MET	12
2318	FLNC	11
50807	ASAP1	11
79723	SUV39H2	11
10278	EFS	10
128853	DUSP15	10
1748	DLX4	10
326	AIRE	10
56955	MEPE	10

Conclusion

This study demonstrates that the greedy algorithm for searching sub-networks allows the identification of novel biomarkers and provides a deeper understanding of their roles in complex diseases such as caners or leukemia. It is more reassuring than those methods not taking formal topological structures into consideration. The obtained results prove that when a sub-network contains more potential nodes that compose a cluster instead of one center node, it is more likely to have a potential relationship with the disease.

Acknowledgement

In this paper, the research was supported by Ministry of Science and Technology of China National Key Research Program grant No. 2016YFC0206507, 973 Program grant No. 2013CB967101, Natural Science Foundation of China No.41476120, No.61572362, No.81571347 and No.21477087.

References

[1] Rendong Yang, Yun Bai, et al. EgoNet: identification of human disease ego-network modules [J]. BMC Genomics, 2014, 15(314):1-10

[2] Chuang H-Y, Lee E, Liu Y-T, Lee D, Ideker T: Network-based classification of breast cancer metastasis [J]. Mol Syst Biol 2007, 3:140.

[3] D Schotte, JCK Chau, et al. Identification of new microRNA genes and aberrant microRNA profiles in childhood acute lymphoblastic leukemia [J]. Leukemia Official Journal of the Leukemia Society of America Leukemia Research Fund U.K. 2009, 23(2):313-322.

[4] Zhigang Li, et al. Gene expression based classification and regulatory networks of pediatric acute lymphoblastic leukemia [J]. Blood, 2009, 114(20):4486-93.

[5] Albert-László Barabási, et al. Network medicine: a network-based approach to human disease [J]. Nature Reviews Genetics [J]. 2011, 12(1):56-68

[6] Chih-Chung Chang, Chih-Jen Lin. LIBSVM: A Library for Support Vector Machines[EB/OL]. https://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[7] Cortes C, Vapnik V: Support-vector networks. Mach Learn 1995, 20:273–297.

[8] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning Data Mining, Inference, and Prediction(book), 2008.

[9] Liu X, Wang Q, Yang G, Marando C, Koblish HK, Hall LM, et al. A novel kinase inhibitor, INCB28060, blocks c-MET-dependent signaling, neoplastic activities, and cross-talk with EGFR and HER-3. Clin Cancer Res. 2011;17(22):7127–38. doi:10.1158/1078-0432.CCR-11-1157.

[10] LA Johnson. Characterizing the cancer genome in lung adenocarcinoma [J]. Nature. 2007, 450(450):893-8.

[11] M Churchman, J Low, et al. Efficacy of Retinoids in IKZF1-Mutated BCR-ABL1 Acute Lymphoblastic Leukemia [J]. Cancer Cell. 2015, 28(3):343–356.

[12] W Fang，X Li, et al. Transcriptional patterns, biomarkers and pathways characterizing nasopharyngeal carcinoma of Southern China [J]. Journal of Translational Medicine. 2008, 6(1):1-13.

[13] Nishii K, at el. S. c-kit gene expression in CD7-positive acute lymphoblastic leukemia: close correlation with expression of myeloid-associated antigen CD 13 [J]. Leukemia. 1992, 6(7):662-8.

[14] Patrick Brown, at el. FLT3 inhibition selectively kills childhood acute lymphoblastic leukemia cells with high levels of FLT3 expression [J]. Blood, 2005, 105(2):812-20.