Bioinformatics, 31(14), 2015, 2412–2414
doi: 10.1093/bioinformatics/btv137
Advance Access Publication Date: 8 March 2015
Applications Note

Data and text mining

BMRF-Net: a software tool for identification of protein interaction subnetworks by a bagging Markov random field-based method

Xu Shi[1,†], Robert O. Barnes[1,†], Li Chen[2], Ayesha N. Shajahan-Haq[3], Leena Hilakivi-Clarke[3], Robert Clarke[3], Yue Wang[1] and Jianhua Xuan[1,*]

[1]Bradley Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA 22203, USA, [2]Department of Pathology, Johns Hopkins University School of Medicine, Baltimore, MD 21231, USA and [3]Lombardi Comprehensive Cancer Center and Department of Oncology, Georgetown University, Washington, DC 20057, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Jonathan Wren

Received on October 28, 2014; revised on February 27, 2015; accepted on March 4, 2015

Abstract

Summary: Identification of protein interaction subnetworks is an important step to help us understand complex molecular mechanisms in cancer. In this paper, we develop a BMRF-Net package, implemented in Java and C++, to identify protein interaction subnetworks based on a bagging Markov random field (BMRF) framework. By integrating gene expression data and protein–protein interaction data, this software tool can be used to identify biologically meaningful subnetworks. A user friendly graphic user interface is developed as a Cytoscape plugin for the BMRF-Net software to deal with the input/output interface. The detailed structure of the identified networks can be visualized in Cytoscape conveniently. The BMRF-Net package has been applied to breast cancer data to identify significant subnetworks related to breast cancer recurrence.
Availability and implementation: The BMRF-Net package is available at http://sourceforge.net/projects/bmrfcjava/. The package is tested under Ubuntu 12.04 (64-bit), Java 7, glibc 2.15 and Cytoscape 3.1.0.
Contact: xuan@vt.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Biological systems can be represented as networks of multifunctional elements and modules. The interactions among elements and modules can make it difficult to elucidate biological behavior using any single type analytical approach (Hanash 2004). Data integration across multiple sources is required to solve the problem in different aspects and levels. Network-based analysis becomes an important tool in revealing the underlying modular structure of the signaling driving complex diseases such as cancer. Several methods (Chuang *et al.*, 2007; Dittrich *et al.*, 2008; Ideker *et al.*, 2002) have been proposed to identify protein interaction subnetworks by integrating high-dimensional gene expression data with protein–protein interaction (PPI) data. However, these methods treat proteins/genes in a network independently, which leaves less differentially expressed hub genes likely unidentified. Recently, a bagging Markov random field (BMRF)-based method (Chen *et al.*, 2013), namely BMRF, have been proposed to identify protein interaction networks by modeling the gene dependency in a network explicitly.

Here we describe a BMRF-Net software package that implements the BMRF algorithm, an integrated network analysis of gene expression data and PPI data. The package provides an effective implementation that reduces the execution time of analyzing large-scale biomedical data and makes visualization of PPI subnetworks immediately accessible to users. A user friendly graphic user interface (GUI) is developed as a Java Cytoscape (Smoot *et al.*, 2011) plugin to support ease of use for biomedical researchers in the field of cancer research. To handle high dimensional data, the open-source software package accelerates the analysis by enabling parallel computing on multi-core machines in laboratories or at large computing centers. In addition to providing the statistical information of the identified networks, the detailed structure of the identified subnetworks can be viewed in Cytoscape. The core analytic or computing program is implemented in C++ in the Linux environment, with a simple interface that can be run without the GUI. To evaluate performance, we applied the BMRF-Net software package to breast cancer data for subnetwork identification. The identified subnetworks show a significant enrichment in cancer related signaling pathways.

2 Description

2.1 The BMRF method

BMRF-Net is a computational tool to identify protein interaction subnetworks based on a BMRF framework (Chen *et al.*, 2013). As defined in (Chen *et al.*, 2013), a network score can be estimated based on the discriminative scores and the network connection from PPI data (see Supplementary Material S1.1 for more details). Compared with the network score defined by the average activity score in (Chuang *et al.*, 2007), this network score takes into account the dependency of network genes. The discriminative score of one gene will be re-estimated based on the scores of its neighboring genes. Consequently, the discriminative scores of some weakly differentially expressed hub nodes can be promoted by the differentially expressed genes connected.

Based on the network score as defined, it is nevertheless an NP-hard problem to find the optimal subnetworks. Instead of using an exhaustive search, simulated annealing, a bottom-up searching approach, is used to reduce the computational complexity. The search starts from several candidate genes referred to as ‘seed’ genes. To obtain more confident results, a non-parametric bootstrapping method is used to measure the confidence of the genes in subnetworks. The confidence level of genes is calculated as the frequency of the genes shown in the subnetworks obtained from the bootstrapping process. The final subnetworks are obtained from network genes with confidence values larger than a defined threshold.

2.2 Software and implementation

The BMRF-Net package is implemented as a C++ program together with a Java GUI. Figure 1 shows an overview of the BMRF-Net package. All the inputs and outputs are handled by a user friendly GUI designed and implemented as a Cytoscape plugin in Java (Supplementary Fig. S1 in the supplementary material). In this package, we have also provided a standalone application of BMRF-Net, which can be run independently of Cytoscape (Supplementary Fig. S2). Besides the input data (i.e. gene expression data and protein interaction networks), several options are needed to run the program. As mentioned in Section 2.1, our method searches for subnetworks starting with several seed genes. Thus, users are required to select several genes of interest as the starting points. Another important parameter is the number of bootstrappings (see Supplementary Table S2) used for the confidence analysis that largely determines the computing time of the algorithm. Results with a larger number of bootstraps have greater confidence; therefore, there is a tradeoff between the confidence of the results and computing time. Users can set the number of bootstraps according to their preferences.

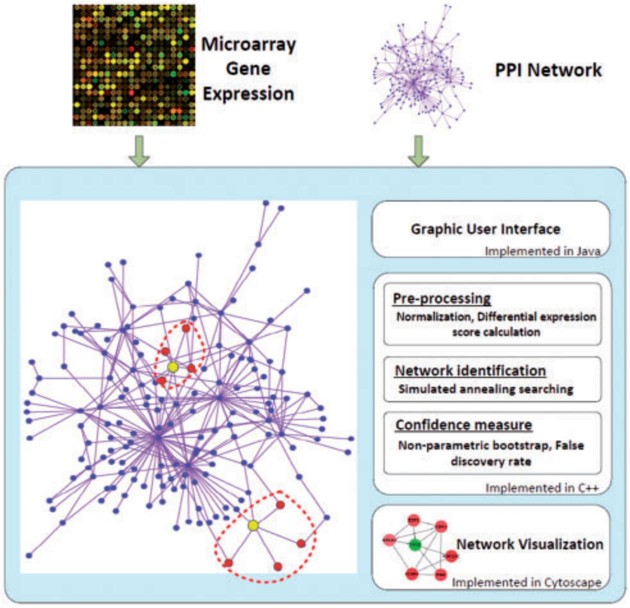

Fig. 1. An overview of the BMRF-Net package

With the input data and options, the C++ core program takes the data to run the core algorithm (please see Supplementary Figs. S3, S4 for more details). Data pre-processing is first performed to standardize the data, and then further calculate differential score by *t*-test. The simulated annealing algorithm then starts to search for a network using the seed genes. As mentioned above, bootstrapping analysis is conducted to infer the confidence level of the results. High computing time is a bottleneck that limits the application to large-scale data analysis. Thus, we have applied concurrency optimization to accelerate the analysis when multi-core machines are available. Iterations of bootstrapping analysis are independent due to the nature of resampling that produces independent data. Network searching will then start independent of the corresponding seed genes. We use boost threads with Philipp Henkel’s thread pool add-on for boost. The number of threads needed depends on the data scale and the number of threads available in the machine, which can be set through the BMRF-Net GUI.

BMRF-Net stores the identified networks in simple interaction format. As a final step, we use Cytoscape to visualize the network (see an example in Supplementary Fig. S5). Users can select the networks of specific genes from the GUI; the network will be automatically loaded into the software.

3 Results

Large computational time is a significant constraint that can limit the application of the BMRF algorithm to large scale genomic data. Here we test the efficiency of BMRF-Net on three PPI networks extracted from the HPRD database release 9 (Keshava Prasad *et al.*, 2009) with different scales. The implementation time is summarized in Supplementary Table S3. As shown in the table, the analysis of a relatively large network with 2545 nodes and 15 094 edges can finish within four minutes when 12 threads are enabled. The speed with 12 threads is about 9 times faster than the single thread implementation, providing the much-needed feasibility for large scale networks analyses in biomedical applications.

The BMRF-Net package was tested on a breast cancer microarray gene expression data set (Loi *et al.*, 2010). The PPI network used was extracted from the HPRD database within two jumps from estrogen receptor, which includes 2794 nodes and 18 202 edges. After mapping the probe sets to the genes, we obtained a gene expression data set with 2794 genes and 48 samples. We further divided the samples into two groups, 'early recurrence' and 'late recurrence', as the samples are divided by 6 years in survival time. We have 20 samples in the 'early recurrence' group and 27 samples in the 'late recurrence' group. The differential score is calculated by *t*-test between two groups and then normalized by standard normal distribution. We further select 204 seed genes which have largest node degrees to start the analysis. It takes about 1.23 h to complete the analysis with 12 threads. Thirty-nine significant subnetworks are selected by a network score threshold of 1.65 and a network size threshold of 5. We then merge the 39 subnetworks to 6 networks, the details of which are shown in Supplementary Figures S6–S11. The genes in the networks are significantly enriched in pathways related to breast cancer progression such as TGF-Beta signaling pathway, Cell cycle, Jak-STAT signaling pathway and ErbB signaling pathway, demonstrating the effectiveness of the BMRF-Net tool in identifying biologically meaningful networks.

4 Conclusion

BMRF-Net is a computational tool to identify biologically meaningful subnetworks from PPI networks. With its user friendly GUI and concurrency options enabled in the core program, the BMRF-Net package can be effectively used to analyze large scale genomic data for subnetwork identification.

Funding

National Institutes of Health, under Grants CA149653, CA149147, CA164384, CA184902 and NS294525-18A1 in part.

Conflict of Interest: none declared.

References

Chen,L. *et al.* (2013) Identifying protein interaction subnetworks by a bagging Markov random field-based method. *Nucleic Acids Res.*, **41**, e42.

Chuang,H.Y. *et al.* (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.

Dittrich,M.T. *et al.* (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics,* **24**, i223–i231.

Hanash,S. (2004) Integrated global profiling of cancer. *Nat. Rev. Cancer,* **4**, 638–644.

Ideker,T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**, S233–S240.

Keshava Prasad,T.S. *et al.* (2009) Human protein reference database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Loi,S. *et al.* (2010) PIK3CA mutations associated with gene signature of low mTORC1 signaling and better outcomes in estrogen receptor-positive breast cancer. *Proc. Natl Acad. Sci. USA,* **107**, 10208–10213.

Smoot,M.E. *et al.* (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics,* **27**, 431–432.