Bioinformatics Advance Access published August 27, 2016

Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw522
Advance Access Publication Date: 8 August 2016
Applications Note

Systems biology

Bayesian network feature finder (BANFF): an R package for gene network feature selection

Zhou Lan[1], Yize Zhao[2], Jian Kang[3,*] and Tianwei Yu[4,*]

[1]Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, USA, [2]Department of Healthcare Policy and Research, Weill Cornell Medical College, New York, NY 10065, USA, [3]Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029, USA, [4]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, USA

*To whom correspondence should be addressed

Associate Editor: Cenk Sahinalp

Received on April 8, 2016; revised on July 12, 2016; accepted on August 3, 2016

Abstract

Motivation: Network marker selection on genome-scale networks plays an important role in the understanding of biological mechanisms and disease pathologies. Recently, a Bayesian nonparametric mixture model has been developed and successfully applied for selecting genes and gene subnetworks. Hence, extending this method to a unified approach for network-based feature selection on general large-scale networks and creating an easy-to-use software package is on demand.

Results: We extended the method and developed an R package, the Bayesian network feature finder (BANFF), providing a package of posterior inference, model comparison and graphical illustration of model fitting. The model was extended to a more general form, and a parallel computing algorithm for the Markov chain Monte Carlo -based posterior inference and an expectation maximization-based algorithm for posterior approximation were added. Based on simulation studies, we demonstrate the use of `BANFF` on analyzing gene expression on a protein–protein interaction network.

Availability: https://cran.r-project.org/web/packages/BANFF/index.html

Contact: jiankang@umich.edu, tianwei.yu@emory.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Feature selection over genome-scale networks has become a very important research question motivated by the needs of incorporating existing knowledge in the analysis of big data in a broad range of biological and biomedical applications. Commonly used genome-scale networks include protein–protein interaction (PPI) network (Rual *et al.*, 2005), transcriptional regulatory network (Licatalosi and Darnell, 2010), signal transduction network (Janes and Yaffe, 2006), metabolomic network (Duarte *et al.*, 2007), etc. Jointly analyzing high-throughput data with the networks can yield robust network markers, i.e. sub-networks critically related to the disease and deeper insights into disease mechanisms.

A Bayesian nonparametric mixture model (Zhao *et al.*, 2014) has been successfully applied to select genes and gene sub-networks under the large-scale simultaneous hypothesis testing framework (Efron, 2004). This method provides a general framework and can be applied to a wide range of biomedical applications. However, its model is relatively restrictive and the computation is not efficient enough on large-scale networks containing tens of thousands of nodes. For improving and implementing this method, we develop an `R` package: Bayes network feature finder (`BANFF`), which provides an efficient implementation of the hierarchical ordered density clustering (HODC), the standard Markov chain Monte Carlo (MCMC) algorithm, the fast computing algorithm based on finite Gaussian mixture approximation, an expectation maximization (EM) algorithm for the posterior mode estimation and the automatic parallel computing implementations of all algorithms. The computational efficiency has been greatly improved (up to 67% reduction in computational time compared to the R code provided by (Zhao *et al.*, 2014). `BANFF` is very user-friendly with a well-written document for illustration of the software. It provides a full package of `R` functions for data preprocessing, efficient Bayesian model fitting with diagnostics, quantitatively and graphically summarizing posterior samples of parameters, along with tutorial examples for the analysis of simulated data and real data.

2 Methods

2.1 Data and models

In the large scale simultaneous hypothesis test framework, a collection of *P*-values or test statistics for many different null hypotheses are considered for the selection of significant features. Our purpose is to incorporate the existing biological network for better feature selection. Suppose we consider n hypothesis tests on n features respectively, where each one tests the association between the feature and the outcome of interest, or certain behavior of the feature. Denote by p_i, the *P*-value and by r_i the test statistic for null hypothesis i, for $i = 1,\ldots,n$. When only *P*-values are available, we can transform it to the test statistic by $r_i = -\phi^{-1}(p_i)$, where $\phi^{-1}(\cdot)$ is the standard normal quantile function. Hence, `BANFF` takes *P*-values $\mathbf{p} = (p_1, \ldots, p_n)^{\mathrm{T}}$ as one of the input data and automatically converts it to $\mathbf{r} = (r_1, \ldots, r_n)^{\mathrm{T}}$ for statistical inference. In addition to $\mathbf{r}$, the network connection information is required, denoted $\mathbf{C} = (C_{i,j})$. Using $\mathbf{r}$ and $\mathbf{C}$ as input data, `BANFF` implements the NETwork enhanced Dirichlet process mixture model (NET-DPM) developed by Zhao *et al.* (2014) to select 'relevant features' under a Bayesian inference framework. Denote by z_i a selection indicator, where $z_i = 1$ indicates feature i is selected, and $z_i = 0$ otherwise. The NET-DPM model makes a common assumption that relevant features are characterized by more extreme test statistics compared to unchanged ones, and it incorporates the network information to improve selection accuracy because it is normally believed that the relevant features tend to be clustered on the network.

Specifically, the NET-DPM model assumes that r_i follows a normal distribution with mean μ_i and variance σ_i^2. The distribution of parameters (μ_i, σ_i^2) given the selection indicator $z_i = k$ is defined by random probability measure G_k, for $k = 0$, 1. The random measure G_k follows a Dirichlet process with base measure G_{0k} and scalar precision τ_k. A weighted Ising prior is assigned to $\mathbf{z} = (z_1, \ldots, z_n)^{\mathrm{T}}$ to incorporate the network information. The probability mass function is given by $\pi(\mathbf{z}|\boldsymbol{\pi}, \boldsymbol{\rho}, \mathbf{C}) \propto \exp\left[\sum_{\mathrm{i}=1}^{\mathrm{n}} (\log(\pi_{\mathrm{z_i}} + \rho_{\mathrm{z_i}} \sum_{\mathrm{i}=\mathrm{j}} \mathrm{C_{ij}} \mathrm{I}[\mathrm{z_i} = \mathrm{z_j}])\right]$ where parameter $\boldsymbol{\pi} = (\pi_0, \pi_1)$ with $0 < \pi_1 = 1 - \pi_0 < 1$ controls the sparsity of $\mathbf{z}$. Parameter $\boldsymbol{\rho} = (\rho_0, \rho_1)$ with $\rho_{\mathrm{k}} > 0$ for $k = 0$, 1 characterizes the smoothness of $\mathbf{z}$ over the network. The method relies on a known network. The selection of the network is outside of the scope of this method. We refer the reader to some reviews of existing biological network databases (Chowdhury and Sarkar, 2015; Szklarczyk and Jensen, 2015).

2.2 Algorithms and implementations

The standard MCMC algorithm (Neal, 2000) has been developed for the NET-DPM model. Please see the algorithm NET-DPM-1 in Section 2.2 and Appendix B1 of Zhao *et al.* (2014) for details. `BANFF` has implemented and optimized this algorithm in R function `Networks.STD()`.

A common issue of all MCMC algorithms is that they are computationally intensive, and thus `Networks.STD()` is fast only for a small scale problem but can be slow for analyzing large data sets. To migrate this problem, `BANFF` takes a two-step approximation approach for posterior inference on $\mathbf{z}$: the first step approximates the marginal posterior distribution of r_i using a finite Gaussian mixture model, for which both DPM density fitting and the EM algorithm are implemented. Each of them can be combined with the HODC algorithm to generate the approximation to the two posterior marginal distributions of 'selected features' and 'unselected features', respectively; The second step simulates the posterior distribution of $\mathbf{z}$ given the approximated marginal posterior distributions using the Gibbs sampling. `BANFF` integrates those algorithms into one R function `Networks.Fast()`.

Table 1. Network marker selection accuracy and computational time

Methods	STD-DPM	STD-EM	NET-DPM-Fast	NET-EM-Fast	NET-STD
TPR	0.64	0.73	0.94	0.99	0.98
FPR	0.24	0.03	0.02	0.01	0.01
FDR	0.27	0.09	0.03	0.01	<0.01
Time (h)	0.20	0.10	1.00	0.40	6.00

TPR, true positive rate; FPR, false positive rate; FDR, false discovery rate.

A parallel computing procedure is implemented when applying DPM density fitting for approximating the marginal posterior distribution: simultaneously obtaining a certain posterior samples of the parameters of the marginal density of r. It greatly reduced the computational time compared to that without parallel computing. `BANFF` can automatically select the hyperparameters $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ by optimizing the Bayesian factor, which are implemented in functions `Network.STD()` and `Network.Fast()`. Please refer to *A guide for BANFF* in the Supplementary Materials for more details on the usages and example for the aforementioned R functions.

3 Examples: simulations and applications

To show the performance of `BANFF`, we performed a simulation study on a 1000-node scale-free gene network. Table 1 summarizes the selection accuracy and computational time for different algorithms that are implemented in `BANFF`. We refer to the standard MCMC algorithm as NET-STD, which is implemented in function `Network.STD()`; and refer to the EM algorithm and the DPM model fitting as NET-EM-Fast and NET-DPM-Fast, respectively, which are implemented by function `Network.Fast()`. We use the DPM model fitting or the EM algorithm combined with the HODC algorithm, referred as DPM-STD or EM-STD accordingly for comparison, both of which do not consider the network information. Apparently, `BANFF` substantially improves the selection accuracy and computational efficiency.

To demonstrate the utility of `BANFF` on high-throughput biological data, we conducted the analysis of a human breast cancer dataset, GSE18864, together with the PPI network obtained from the HINT database. The results were biologically meaningful and shed new lights on the data. In addition, `BANFF` can be applied for imaging data analyses where the spatial structure of the image is considered as a network. For details of these studies, please see *A guide for BANFF* in the Supplementary Materials.

4 Conclusion

In summary, BANFF implements a Bayesian nonparametric approach for large-scale multiple hypothesis testing over the network, motivated by the need for network marker selection on genome-scale networks. Compared to the existing software, BANFF achieves a high feature selection accuracy, controls the false discovery rate very well, and is computationally efficient for large scale network. It has a broad range of applications in biomedical sciences.

Funding

This work was partially supported by National Institutes of Health (Grants 1R01MH105561-02 and U19 AI090023).

Conflict of Interest: none declared.

References

Chowdhury,S. and Sarkar,R.R. (2015) Comparison of human cell signaling pathway databases–evolution, drawbacks and challenges. *Database (Oxford)*, **2015**, doi: 10.1093/database/bau126.

Duarte,N.C. *et al*. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl. Acad. Sci.* USA, **104**, 1777–1782.

Efron,B. (2004) Large-scale simultaneous hypothesis testing. *J. Am. Stat. Assoc.*, **99**, 96–104

Janes,K.A. and Yaffe,M.B. (2006) Data-driven modelling of signal-transduction networks. *Nat. Rev. Mol. Cell. Biol.*, **7**, 820–828.

Licatalosi,D.D. and Darnell,R.B. (2010) Rna processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.*, **11**, 75–87.

Neal,R.M. (2000) Markov chain sampling methods for dirichlet process mixture models. *J. Comp. Graph. Stat.*, **9**, 249–265.

Rual,J.-F. *et al*. (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, **437**, 1173–1178.

Szklarczyk,D. and Jensen,L.J. (2015) Protein–protein interaction databases. *Methods Mol. Biol.*, **1278**, 39–56.

Zhao,Y. *et al*. (2014) A bayesian nonparametric mixture model for selecting genes and gene subnetworks. *Ann. Appl. Stat.*, **8**, 999–1021.