Journal of Theoretical Biology 362 (2014) 75–82

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

A co-expression modules based gene selection for cancer recognition

Xinguo Lu [a,b,*], Yong Deng [a], Lei Huang [a], Bingtao Feng [a], Bo Liao [a]

[a] School of Information Science and Engineering, Hunan University, Changsha 410082, China
[b] College of Mechatronics and Automation, National University of Defense Technology, Changsha 410073, China

ARTICLE INFO

Available online 15 January 2014

Keywords:
Gene expression data
WGCNA
Cancer recognition

ABSTRACT

Gene expression profiles are used to recognize patient samples for cancer diagnosis and therapy. Gene selection is crucial to high recognition performance. In usual gene selection methods the genes are considered as independent individuals and the correlation among genes is not used efficiently. In this description, a co-expression modules based gene selection method for cancer recognition is proposed. First, in the cancer dataset a weighted correlation network is constructed according to the correlation between each pair of genes, different modules from this network are identified and the significant modules are selected for following exploration. Second, based on these informative modules information gain is applied to selecting the feature genes for cancer recognition. Then using LOOCV, the experiments with different classification algorithms are conducted and the results show that the proposed method makes better classification accuracy than traditional gene selection methods. At last, via gene ontology enrichment analysis the biological significance of the co-expressed genes in specific modules was verified.

1. Introduction

The rapidly developed microarray technique generates a huge amount of large-scale gene expression profiles. These gene expression data are applied to cancer diagnosis and therapy. In these gene expression profiles, there is a lot of noisy and redundant information. The research of cancer recognition using gene expression data is usually plagued with "curse of dimensionality". How to reduce these useless features is the key step to cancer recognition (Wang and Gotoh, 2010; Lu et al., 2006).

Dimension-reduction is crucial to address the "curse of dimensionality" problem in molecular classification of cancer (Golub et al., 1999; Cho et al., 2003). Feature transformation has been applied to processing gene features. The transformation includes principal components analysis (PCA) (Raychaudhur et al., 2000), independent components analysis (ICA), and clustering based transformation (Conde et al., 2002). Feature transformation is able to discover the underlying informative factors for classification. However, the noisy and redundant information cannot eliminate efficiently.

Gene selection (Kuramochi and Karypis, 2005) is also an effective dimension-reduction method. Numerous methods of selecting informative gene groups to conduct cancer classification have been proposed. Most of the methods ranked the genes based on certain criteria firstly, and then selected a small set of informative genes for classification from the top-ranked genes. The most used gene ranking approaches include the signal to noise ratio, *t*-score, chi-square, information entropy-based, Relief-F, symmetric uncertainty, etc. (Parsons et al., 2004; Shipp et al., 2002). One of the weaknesses of these ranking methods is that they only consider the individual features in isolation and ignore their possible interactions.

Correlation networks are increasingly being used in bioinformatic applications. A protein-network-based approach (Chuang et al., 2007) that identifies subnetwork markers is more reproducible than individual marker genes selected without network information. A novel network-based systems biology framework (Liu et al., 2011a) was present to identify and analyze differentially activated pathways, and found that these dysfunctional pathways provide insights into understanding the dynamics of AD progression in six brain regions. Here, a spatio-temporal analysis on type 2 diabetes mellitus (Sun et al., 2013) was performed by developing a new form of molecular network.

Weighted gene co-expression network (Zhang and Horvath, 2005) analysis is a systems biology method for describing the correlation patterns among genes across microarray samples. Instead of relating thousands of genes to the physiologic trait, it focuses on the relationship between a few modules and the trait. Using the weighted gene co-expression network analysis the hiding and biologic patterns can be explored.

In this paper a co-expression modules based gene selection method for cancer recognition is proposed. A weighted correlation network (Langfelder and Horvath, 2009; Fuller et al., 2007) is constructed and the gene co-expression modules are detected. These cancer related modules are chosen and their biologic significance are analyzed. Then the informative genes from these modules are selected for cancer recognition. Via these steps the

* Corresponding author at: School of Information Science and Engineering, Hunan University, Changsha 410082, China.
E-mail address: hnluxinguo@hnu.edu.cn (X. Lu).

0022-5193/$ - see front matter Crown Copyright © 2014 Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2014.01.005

most distinguishing gene features are picked out. At last classification methods are used to test the recognition performance with these genes. Using LOOCV the experiments with different classification algorithms are conducted on DLBCL dataset, colon dataset and breast cancer. The experimental results show that the proposed method makes better precision than other usual gene selection algorithms (Li and Ruan, 2005).

2. Pre-requirements

2.1. Gene expression data

The gene expression data is usually represented by the expression matrix. Let X be a gene expression matrix, where rows of elements represent genes, columns of elements represent various samples. Thus cell x_{ij} is the ith gene expression level in the jth sample.

2.2. Classification methods

2.2.1. SVM

Support vector machine (SVM) estimates the function classifying the data into two classes. Via building a hyperplane SVM makes a decision surface to maximize the margin of separation between positive and negative examples. SVM achieves this by the structural risk minimization principle that the error rate of a learning machine on the test data is bounded by the sum of the training-error rate and a term that depends on the Vapnik–Chervonenkis (VC) dimension. For notional simplicity, we consider the case of a linear classifier function

$$f(X)=\omega^{\tau}X+b \tag{1}$$

where ω and b are unknown and determined from the training dataset. With the kernel notion $K(X_i,X_j)=\Phi(X_i)^{\tau}\Phi(X_j)$, the resulting SVM classifier function can be rewritten as

$$f(X_i)=\sum_{i=1}^{M} y_k\alpha_k K(X_i.s_k)+b \tag{2}$$

where s_k, $k=1,\ldots,N_s$, are the so-called support vectors, which correspond to those training samples that are either inside or on the decision margin of the classifier, which are determined during the training step.

2.2.2. Decision tree (DT)

The construction of J48 consists of tree construction and tree pruning. In tree construction, the predictable variables are identified and divided into two child nodes. The split maximizes the homogeneity of the sample population in each child node. For example, most of the samples in one child node are the cancer samples, and most of the samples in the other child node are non-cancer samples. Then, these child nodes are split further until samples are in one category or the quality of this DT model cannot be improved. To avoid over-fitting the tree is pruned to a desired size using a cost complexity pruning method.

3. A weighted correlation network based gene selection for cancer recognition

3.1. Gene correlation network construction

Gene correlation (Witten and Frank, 2005; Langfelder and Horvath, 2008) is described as a network in which the relationship between the connected genes is represented by the weight. In a gene correlation network, as a undirected network, each gene is described as a node. The edge weight between the connected nodes is the pairwise Pearson coefficient. The higher degree of topological overlap between two nodes in the gene correlation network implicates the same biological function or pathway. The topological overlap measure can be calculated as follows:

$$\begin{aligned}\omega_{ij}&=\frac{l_{ij}+a_{ij}}{\min\{k_i,k_j\}+1-a_{ij}}\\ l_{ij}&=\sum_u a_{iu}a_{uj}\\ k_i&=\sum_u a_{iu}\\ k_j&=\sum_u a_{uj}\end{aligned} \tag{3}$$

$$a_{ij}=|s_{ij}|^{\beta}=|cor(x_i,x_j)|^{\beta} \tag{4}$$

where ω_{ij} is the similarity between topology matrix nodes, k_i is the connectivity similarity of ith gene, k_j is the connectivity similarity of jth gene. l_{ij} represents the sum of products that the adjacency coefficient of the node gene i and gene j common connects. s_{ij} is the Pearson coefficient between the expression of ith gene and jth gene. To emphasize the close correlation between genes, the power β is used to acquire adjacency function a_{ij}, $\beta\geq 1$. We use the scale free topology criterion to choose the soft threshold β.

3.2. Co-expression modules identification

In gene correlation network construction, an adjacency matrix and an adjacency function are defined. Using adjacency function the co-expression similarity between genes can be described as connection strength. The adjacency function consists of different statistical or biological criteria. The resulted adjacency matrix is used to define a measure of node dissimilarity d_{ij}^{ω} which is calculated as follows:

$$d_{ij}^{\omega}=1-\omega_{ij} \tag{5}$$

The node dissimilarity is input to hierarchical clustering to define network modules. From the clustering tree many gene co-expression modules are discovered. In the construction of hierarchical clustering tree, a dynamic shear algorithm based on tree branch shape is used.

3.3. Co-expression modules based gene selection (CMGS)

After the modules are identified, the relevance between the gene modules and the cancer class can be calculated. First, the eigenvalues and eigenvectors of gene modules are derived. Then the correlation between the module eigenvectors and the cancer classes of samples is obtained. Finally, T test is used to analyze significantly differently expressed genes between different groups and gene significance (GS) is derived. The average gene significance in a module is defined as the module significance (MS). These modules are ranked according to their module significance coefficients. The top gene module significance coefficients are accumulated until the accumulated coefficients is greater than a threshold and these corresponding modules are selected.

However, the number of selected genes is too larger to achieve better recognition performance. So information gain (IG) is applied to obtaining the more distinguished genes. Information gain is an entropy-based feature evaluation method. As IG is used in feature selection, it is defined as the amount of information provided by the feature genes for the sample classification. Information gain is calculated by how much of a gene can be used for classification of information, in order to measure the importance of genes for the classification. The formula of the information gain is described as