GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Classification of colon cancer based on the expression of randomly selected genes

X.H. Tan, R. Cheng, H.P. Hu and Y.P. Bai

School of Science, North University of China, Taiyuan, Shanxi, China

Corresponding author: Y.P. Bai
E-mail: baiyp666@163.com

Genet. Mol. Res. 14 (4): 12628-12635 (2015)
Received March 12, 2015
Accepted August 14, 2015
Published October 19, 2015
DOI http://dx.doi.org/10.4238/2015.October.19.6

ABSTRACT. In order to ascertain the relationship between gene expression and colon cancer localization, a classification method based on random gene selection and a self-organizing map network is proposed. Different numbers of genes were selected randomly from 54,675 genes of 53 colon cancer patients in stage union for international cancer control II. These patients were then divided into two sets: a training set of 36 and a validation set of 17 patients. In this study, we randomly selected 1000, 100, 50, 30, 10, 5, and 3 genes, 1000 times, respectively. The minimum misclassification ratio of each gene group was 3/17 to 4/17, and the percentage of gene groups that were less than 0.25 was approximately 1-7%. Moreover, the misclassification ratio of most gene groups (about 82-89%) was lower than 0.4. Through the analysis of these low misclassification ratio gene groups, we found that there were few common genes between them. This revealed that colon cancer localization is not associated with a single gene group but with many gene groups. Furthermore, K-fold cross validation was used to test the reliability of the possible informative genes, and the results indicated that using gene expression to classify colon tumor localization was not feasible.

Key words: Informative genes; Colon cancer localization;
Self-organizing maps; Random genes; Tumor classification;
Nonlinear principal component analysis

INTRODUCTION

The incidence of cancer increases annually, and the desire for improved treatments and diagnostics is always at the forefront of science. In recent years, studies on the relationship between gene expression profiles and different cancer outcomes of cancer have been particularly illuminating (Pomeroy et al., 2002; Vant't Veer et al., 2002; Ntzani and Ioannidis, 2003; Moreaux et al., 2013). However, the challenge in this process is the analysis of a large amount of gene expression data. Indeed, there are typically only a small number of key genes that affect cancer classification, while the others are often less significant or irrelevant. Therefore, it is necessary that a specific group of genes be identified to classify different types of cancer effectively, and these genes are called informative genes. Several methods have been proposed for identifying such genes. For instance, neighborhood analysis was used for the classification of acute myelocyticleukemia and acute lymphocytic leukemia, by which 1100 genes were selected and the accuracy rate was 89.47% (Golub et al., 1999). Moreover, 64 genes selected through support vector machine were used for colon cancer classifications, and were 98% accurate (Guyon et al., 2002). These studies indicate that the use of gene expression data for tumor classification has become an important method.

Conversely, some studies have indicated that random gene sets as predictors of prognosis are highly unreliable (Michiels et al., 2005). For instance, most random gene expression signatures are significantly associated with breast cancer (Venet et al., 2011). These studies, therefore, dispute whether random gene sets allow for adequate classification, and whether specific gene groups can be informative genes. The purpose of this paper is to explore this issue in further detail by evaluating the accuracy of self-organizing map (SOM) neural network in the classification of colon cancer localization. In section 2, with the absence of priori information, random 1000, 100, 50, 30, 10, 5, and 3 genes are put into the SOM network as possible informative genes, get classification results; section 3 analyzes the statistical characteristics and stability of the classification results further; and section 4 concludes the paper.

MATERIAL AND METHODS

Clinical data

The dataset we used was from National Center for Biotechnology Information (NCBI). It consisted of expression profile arrays of 53 colon cancer patients in sporadic stage union for international cancer control II . There were 54,675 gene expression features in this group, and according to the localization of the tumor, the 53 patients were divided into two groups: 25 distal colon cancer and 28 proximal colon cancer.

Research methods

SOM is an effective method for predicting cancer outcomes (Golub et al., 1999; Gohari et al., 2011, Biglarian et al., 2012; Valarmathi and Radhakrishna, 2013). In this study, we formed a system based on a KohonenSOM neural network to predict the localization of colon tumors. KohonenSOM network (Kohonen, 2001) is one of SOM networks, it arranges neurons in a two-dimensional grid, and the competitions among neurons make each neuron represents a class of input patterns. Competition among neurons of the output layer makes the one with the highest value win, and then the winning unit weight is adjusted to make the network represent those input patterns.

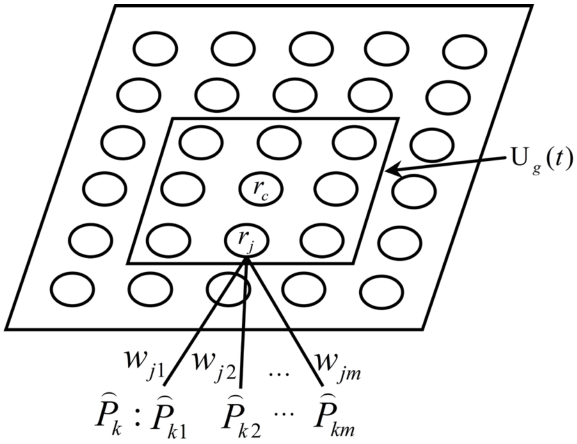

Figure 1. Structure of Kohonen self-organizing map.

As shown in Figure 1, P = (P_1,..., Pn) is a set of training data with size n, and the dimension of the input space is m. $W_j = (w_{j1} = w_{j2}... = w_{jm})$ (j = 1, 2,..., s) are neuron weights connecting neuron j and the components of input vector. The training process for the SOM algorithm is given in the following four steps:

Step 1: Initialize the network. Normalize each input vector P_k into $\hat{P}_k$ subject to $\|\hat{P}_k\|=1$,

$$\hat{P}_k = \frac{P_k}{\|P_k\|} = \frac{P_k}{\left[\sum_{i=1}^{m}(P_{ki})^2\right]^{1/2}}, \quad (k=1,2,\cdots,n) \qquad \text{(Equation 1)}$$

Give neuron weights W_j ($j = 1, 2,..., s$) equal to a part of the normalized input vector $\hat{P}_{k_l}$ (l = 1, 2,..., s), and $\|W_j\| = 1$.

Step 2: Calculate the inner product between the normalized input vector $\hat{P}_k$ and each neuron weight W_j. Identify the winning neuron W_c:

$$\hat{P}_k \cdot W_c = \max_j(\hat{P}_k \cdot W_j) = \max_j(\left\|\hat{P}_k\right\|\left\|W_j\right\|\cos\theta_{jk}) = \max_j(\cos\theta_{jk}) \qquad \text{(Equation 2)}$$

where Θ_{jk} is the angle between $\hat{P}_k$ and W_j.

Step 3: Adjust the weights of the winning neuron W_c and its neighbor unit W_j:

$$W_{ji}(t+1) = W_{ji}(t) + \eta(t)[\hat{P}_k^{\,i} - W_{ji}(t)], \quad j \in \mathrm{U}_g(t) \qquad \text{(Equation 3)}$$

where $\eta(t)$ is a decreasing learning rate as a function of time t, and U_g (t) is a decreasing neighborhood kernel with Gaussian function:

$$\mathrm{U}_g(t)=\exp(\frac{\left\|r_c - r_j\right\|^2}{2\sigma^2})\,\eta(t) \qquad \text{(Equation 4)}$$

where r is the locations of the neuron on the two dimensional map grids, r_c and r_j are the locations of the winning neuron and neuron j, σ is a smoothing factor.

Step 4: Repeat steps 2 and 3 until the convergence criterion is satisfied.

In practical computation, similarity can be replaced by distance, and thus, the similarity measurement for the expression of two genes is converted to the distance between the expressions of two genes. Thus, the smaller the distance, the more similar the expression patterns. Under the assumption that there are informative genes regarding colon cancer localization, we formed SOM neural networks to classify gene groups that included different numbers of genes. By comparing the results to the classification of colon cancer localization, we can test the validity of the classification and prediction.

According to the training rule of SOM, we used the gene groups of different patients as training vectors. When multiple similar gene vectors are put into the network, the final training results make the network weights similar to the average value of the input vector. Thus, the trained network achieves the function of classification. This is suitable for the principle of gene classification. If the selected genes are strongly associated with tumor categories, then the gene set vectors of the same tumor categories will be similar and fall into one group and the error rate of classification would be low. Conversely, weak association will lead to high error rates of classification. Given this fact, we aimed to find appropriate gene groups and suitable SOM models that can provide low error rates of classification, and thereby produce the most possible informative genes for the classification of colon tumor localization.

Two important aspects to finding informative genes are the number and type of genes in a gene set. In order to find the possible informative genes, different numbers of genes were randomly selected in this study (i.e., 1000, 100, 50, 30, 10, 5, and 3 genes), and this random selection was done 1000 times for each grouping. As shown in Figure 2, the components of the input vectors are based on a random combination of genes each time. For the 53 patients, 36 were used as a training set (in which 16 were distal colon tumor and 20 were proximal colon tumor) and 17 (in which 9 were distal colon tumor and 8 were proximal colon tumor) were used as the test set. Each combination of random genes is used to train a SOM neural network, and each trained SOM gives a test result for the corresponding gene group.

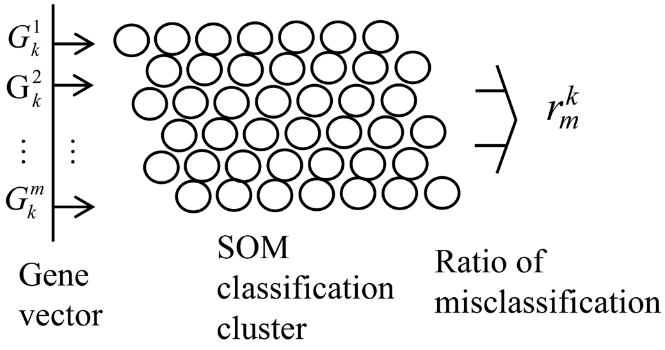

Figure 2. SOM classification cluster, where G_k^m is the *m* random selected gene in the *k* experiment, and r_m^k is the corresponding misclassification ratio, *k* = 1, 2,…, 1000; *m* = 1000, 100, 50, 30, 10, 5, and 3.

One thousand genes were selected randomly from the 54,675 genes in total. The genes selected from the 36 patients of the training set were used as the input vectors of a SOM network to train the network and get the classification. The randomly selected genes from the other 17 patients were then used to validate the network. At the same time, the misclassification ratio of the 17 patients was recorded together with the 1000 genes. This process of selecting 1000 genes randomly, training the network, validating the network, and recording the misclassification ratio was performed 1000 times. The process was repeated for sets of 100, 50, 30, 10, 5, and 3 genes.

RESULTS

By analyzing the misclassification ratio, the gene combinations that make a good classification effect on colon tumor localization were recorded for further analysis. The results of this experiment show that most gene groups were associated with the localization of colon cancer, but to different degrees. In each 1000 times simulation, the distributions of the misclassification ratios with different numbers of genes were similar. For instance, the percentage of groups that were less than 0.25 was about 1-7%, and most gene groups (about 78-89%) had a weak association with the colon cancer localization (their misclassification ratios were more than 0.4; Figure 3). The best misclassification ratios were 3/17 to 4/17 (Table 1), and the percentage of the corresponding gene groups was less than 1%. In this study, the 1000- and the 30-gene group provided the most accurate classifications.

The general view is that the high classification accuracy gene groups should be (or contain) the possible informative genes, and thus, those sets were analyzed according to the-following methods.

Remove interference with nonlinear principal component analysis (NLPCA)

The dataset used in this study was from 53 colon cancer patients, and the samples used to train the SOM network was 37. The number of genes was large, and this will cause an outlier disturbance on the classification results. Principal component analysis (PCA) is an effective method for feature extraction and dimensionality reduction of data sets, and it has been used to discover the association between genes and cancer (Venet et al., 2011). However, this method is very sensitive to outliers in the data, and may be replaced by NLPCA, which is less sensitive to outliers (Verboon, 1991; Scholz et al., 2002). For instance, in the 30-gene group, for example, NLPCA was used to reduce the dimension of the data from 30 to 10 and then, the 10-dimensional data was used to train and test the SOM network. The results indicate that NLPCA was an effective method for feature extraction of a gene set, and the ratio of misclassification was reduced from 3/17 to 2/17. This suggests that the accuracy of classificationcan be improved by using NLPCA.

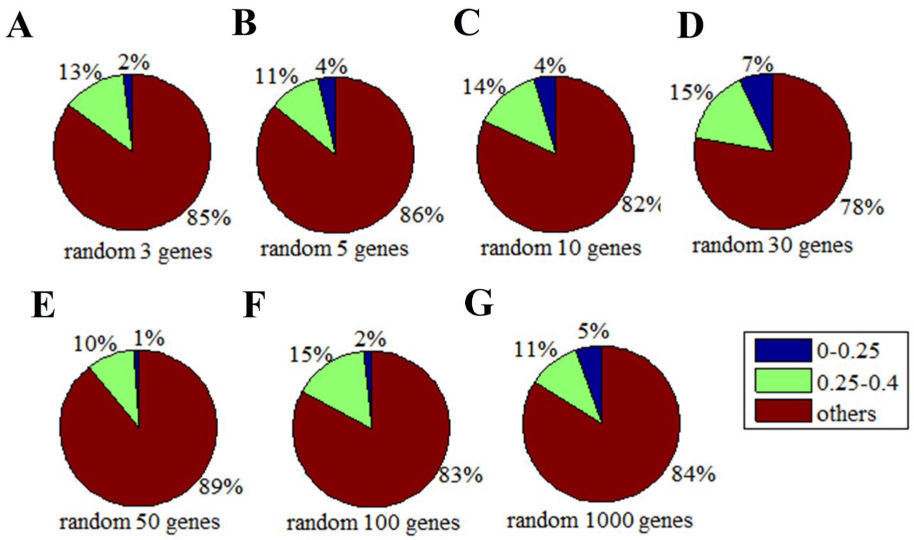

Figure 3. Distributions of misclassification ratio in 1000 experiments. **A. B. C. D. E. F. G.** correspond to 3, 5, 10, 30, 50, 100, 1000 gene groups, respectively, and the distributions of misclassification ratio in each 1000 experiments.

Table 1. Minimum misclassification results of different gene groups and their percentage in 1000 times.

	1000 genes	100 genes	50 genes	30 genes	10 genes	5 genes	3 genes
Minimum misclassification ratio	3/17	4/17	4/17	3/17	4/17	4/17	4/17
Percentage of misclassification ratio <0.25	5.9%	1.8%	1.0%	6.8%	3.5%	4.1%	1.9%

Search for informative genes

According to the SOM results, there were many gene groups that were strongly associated with colon cancer localization. To further explore these relationships, the gene groups whose misclassification ratios reached 3/17 were taken for further analysis. Specifically, there were four groups of 1000 genes (A_1, A_2, A_3, A_4) and four groups of 30 genes (B_1, B_2, B_3, B_4) that were taken for further study. In order to study the similarities between them, their intersections were calculated. The intersection between two groups of 1000 genes contained approximately 20-30 genes, but there was no common gene when three groups of genes were compared. The six intersections found were used to train a SOM network, and the results indicate that the misclassification ratio was similar to the individual groups. Similar to the 1000-gene groups, intersections between groups of 30 genes contained, at most, only one gene in common, and no common genes between three groups of 30 genes. Moreover, there were very few genes contained in the intersection between 1000- and 30-gene groups (Figure 4). This reveals that although there are many gene groups associated with colon cancer localization, there are few common genes among them. Thus, the potential gene groups that can be used to classify the colon cancer localization are not unique.

	A_1	A_2	A_3	A_4
A_1	1000	26	20	24
A_2	26	1000	10	27
A_3	20	10	1000	25
A_4	24	27	25	1000

	B_1	B_2	B_3	B_4
B_1	30	1	0	1
B_2	1	30	0	0
B_3	0	0	30	1
B_4	1	0	1	30

Figure 4. Number of genes in the intersection of 1000- and 30-gene groups.

K-fold cross validation

To determine how reasonable the above method was for determining gene associations with colon cancer localization, we utilized k-fold cross validation (K-CV) to further analyze the data. K-CV can avoid over-training and under-training a network. Specifically, we divided the raw data into k groups, and made each subset of data a validation set, and the rest k-1 subsets of data were used as training sets. From these sets of data k SOM networks were developed. The average ratio of misclassification for the k validation set using this k model was used as the performance index of the K-CV classifier. If the gene groups are strongly associated with colon cancer localization, the average ratio of misclassification will be low, and the networks perform good stability at the same time. In this study, six gene groups that had good classification effects (misclassification ratios less than 0.2) were chosen for K-CV. Fifty-three patients were divided into three groups (18, 18, and 17), the validation results indicated that the results changed substantially with the change in gene array. For the six gene groups the 6 average misclassification ratios were above 0.2 and less than 0.4, worse than original experiment.

Table 2. K-CV of some gene groups.

Gene groups	Ratio of misclassification	3-fold cross validation Ratio of misclassification	Average ratio of misclassification
1	0.177	0.333, 0.389	0.300
2	0.177	0.333, 0.444	0.318
3	0.226	0.222, 0.278	0.242
4	0.226	0.389, 0.444	0.353
5	0.235	0.111, 0.333	0.226
6	0.235	0.167, 0.333	0.245

DISCUSSION

In this study, we proposed a method to find informative genes that were strongly "associated" with colon tumor localization. The randomly selected gene groups were first trained by the SOM network, and from this, the gene groups with low error classifications were chosen for further improvement by NLPCA. Analysis of these select gene groups indicated that there were few common genes between them, and the result of K-CV shows weak association between the six gene groups and colon tumor localization. Overall, the results indicate that utilizing gene expression to classify colon tumor localization is not feasible.

Due to the data limitations, this study involved only 53 patients, all of which had colon cancer. Future research should concentrate on utilizing a higher number of samples across a range of different outcomes and tumors.

Conflicts of interest

The authors declare no conflicts of interest.

ACKNOWLEDGMENTS

Research supported by the National Science Foundation of China (#61275120).

REFERENCES

Biglarian A, Bakhshi E, Gohari MR and Khodabakhshi R (2012). Artificial neural network for prediction of distant metastasis in colorectal cancer. *Asian Pac. J. Cancer Prev.* 13: 927-930.

Gohari MR, Biglarian A, Bakhshi E, Pourhoseingholi MA, et al. (2011). Use of an artificial neural network to determine prognostic factors in colorectal cancer patients. *Asian Pac. J. Cancer Prev.* 12: 1469-1472.

Golub TR, Slonim DK, Tamayo P, Huard C, et al. (1999). Molecular classification of cancer: Class prediction by gene expression monitoring. *Science* 286: 531-537.

Guyon I, Weston J and Barnhillet S (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.* 46: 389-422.

Kohonen T (2001). Self-Organizing Maps. 3rd edn. Springer, Berlin.

Michiels S, Koscielny S and Hill C (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 365: 488-492.

Moreaux J, Reme T, Leonard W, Veyrune JL, et al. (2013). Gene expression-based prediction of myeloma cell sensitivity to histone deacetylase inhibitors. *Br. J. Cancer* 109: 676-685.

Ntzani EE and Ioannidis JP (2003). Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet* 362: 1439-1444.

Pomeroy SL, Tamayo P, Gasenbeek M, Sturia LM, et al. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415: 436-442.

Scholz M, Fraunholz M and Selbig J (2002). Nonlinear Principal component analysis: neural networks models and applications. In: Principal Manifolds for Data Visualization and Dimension Reduction (Groban AN, Kégl B, Wunsch DC and Zinovyev A, eds.). Springer Berlin Heidelberg, 44-46.

Valarmathi P and Radhakrishna V (2013). Tumour Prediction in Mammogram Using Neural Network. *Global J. Comp. Sci. Technol. Neural Artificial Intelligence* 13: 19-24.

Vant't Veer LJ, Dai H, van de Vijver MJ, He YD, et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415: 530-536.

Venet D, Dumont JE and Detours V (2011). Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Comput. Biol.* 7: 1-8.

Verboon P (1991). Nonlinear Principal Components Analysis: Overview and New Developments with Respect to Resistance Properties. Department of Data Theory, University of Leiden, Netherlands.