DNA 마이크로어레이를 이용한 암 관련 유전자 탐색

박진현 [1,2], 이동권[2], 최상욱[3], 조지훈[2], 김명수, 김덕희[2], 이인범[2,3]
[1]㈜피앤아이컨설팅, [2]포항공과대학교 화학공학과, [3]포항공과대학교 환경공학부

Searching for Cancer Related Genes with DNA Microarray

Jin Hyun Park[1, 2], Dongkwon Lee[2], Sang Wook Choi[3], Ji-Hoon Cho[2], Myengsoo Kim[1], Duk-Hee Kim[2], and In-Beum Lee[2, 3]
[1]P&I Consulting Co., Ltd., pcanda@postehc.ac.kr,
[2]Department of Chemical Engineering, POSTECH
[3]School of Environmental Science and Engineering, POSTECH

Introduction

Upcoming the era of postgenomic, various important data are produced from biological and medical experiments such as DNA microarray, 2D-PAGE, blotting and so on. Although these data includes many significant biological meaning and concepts, it is hard to find out useful and significant knowledge from multidimensional raw data because these data contain noises and faults from the non-automatic experiment and the change of various experiment conditions. Then, bioinformatics (especially data analysis) is spotlighted in all area of biology because data analysis technique can give statistical and biological annotation with more systematical approach. For the gene expression data, the following systematical data analysis steps are needed. 1)Collection of DNA chip data 2)Check of statistical significance of differential gene expression 3)Normalization of DNA microarrays 4)Selection of the significant genes 5)Elucidation of biological phenomena from the gene expression patterns

Cancer has many subtypes and each subtype requires a different therapy regimen. Hence an exact discrimination of the cancer subtype is crucial for successful treatment [1-5]. The appearance of the cDNA microarray and oligonucleotide chip technologies brought about a revolution in the distinction of cancer subtypes and the discovery of subtype specific genes that can be used for diagnostic purposes. It is very inspiring to work with these technologies, which enable the possibility of analyzing several thousand human genes simultaneously. From the clinical viewpoint, it is also an important issue to uncover the reason for clinical heterogeneity and analyze the therapy response. Alizadeh *et al.* looked into the cause of clinically heterogeneous behavior and identified the relationship of a clinical prognostic indicator and survival time in diffuse large B-cell lymphoma (DLBCL). Veer *et al.* [6] predicted clinical outcome of breast cancer using unsupervised two-dimensional cluster analysis. And also, Shipp *et al.* [7] extracted informative genes more systematically and take steps to predict clinical outcome in DLBCL patient data. However, all these studies do not address the relationship between therapy response and related genes.

In this research, we suggest biplot analysis, which is based on PCA and DPLS, for the clinical outcome analysis without use of a prior biological knowledge. During tumor therapy, this kind of analysis provides therapeutic guideline and graphical information about the relationship between genes. And the relative position of genes and patients, which is plotted on the same domain, can give additional information. That is, these approaches make it possible to analyze the drug effect and provide high quality treatment with a viewpoint of rational molecular biology. The biplot based on PCA and DPLS algorithm, and gene (variable) selection methods are described below, after which the ability of these methods to elucidate the relationship between therapy responses and genes is evaluated.

Method

The biplot provides plot of the *m*-dimensional variables, and simultaneously gives plot of the relative positions of the *n*-dimensional sample data. When these types of plots are overlapped, this gives additional insight about relationship between variables and samples. There are numerous other methods for biplot analysis, but we mainly used the principal component analysis (PCA) and discriminant partial least square (DPLS) method because it is easy to use and gives good graphical representation.

Principal Component Analysis (PCA)

The principal component analysis (PCA) can transform the high dimensional problem into lower dimensional problems without excessive information loss. It transforms a set of correlated variables into a new set of data which are uncorrelated to each other. It can be calculated by singular value decomposition of the covariance matrix of original variables. Let **X** be *m*-dimensional variables and *n*-dimensional sample data. The matrix **X** can be decomposed into a loading matrix **P** showing the influence of variables and a score matrix **T** that summarizes the **X** variables.

$$\hat{\mathbf{X}} = \mathbf{T}\mathbf{P}^T \qquad (1)$$

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \qquad (2)$$

The **E** denotes the residual matrix, the deviations between the original values and the projections. Here, the superscript *T* denotes the operation of matrix transposition and $\hat{\mathbf{X}}$ denotes the projection value of **X**.

Discriminant Partial Least Square (DPLS)

Discriminant partial least square (DPLS) is used to extract the informative genes and produce a calibration equation to predict clinical outcome from gene expression data. In this section, we shortly review the concept of DPLS and then show the variable selection procedure based on DPLS regression coefficients.

DPLS models the relationship between the predictor (independent) block **X** and the predicted (dependent) block **Y** using a series of local least-square fits. Unlike the typical PLS method, the predicted block $\mathbf{Y} \in R^{n \times c}$ in the DPLS model is made to contain information about class memberships of objects. Each row, $\mathbf{y}^T$, in the **Y** matrix has the following structure

$$y_{ij} = \begin{cases} 1, \text{ if the } i\text{th sample belongs to class } j \\ 0, \text{ otherwise} \end{cases} \qquad (3)$$

where, $\mathbf{y}_j$ is the *j*th column in **Y** and $j = 1, 2, \ldots, c$.

Consequently, the **Y** matrix has the binary variable form [8].

The matrices **X** and **Y** should be autoscaled (zero mean and unit variance). The predictor block **X** is decomposed into a score matrix $\mathbf{T} \in R^{n \times lv}$, a loading matrix $\mathbf{P} \in R^{m \times lv}$ and a residual matrix **E** (same size as **X**) as follows, where *lv*, which determines the order of the PLS model, is the number of latent variables.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \qquad (2)$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \qquad (4)$$

Using the nonlinear iterative partial least square (NIPALS) algorithm, the regression matrix **B** that maximizes the covariance of **X** and **Y** can be obtained and the DPLS model is given by

$$\mathbf{Y} = \mathbf{T}\mathbf{B}\mathbf{Q}^T + \mathbf{F}^* \qquad (5)$$

where, $\mathbf{F}^*$ is the prediction error matrix whose norm is minimized by the regression matrix, **B**.

When the final DPLS model is used in prediction, the estimated $\hat{\mathbf{Y}}$ matrix does not have the same structure **Y**. That is, the predicted values in $\hat{\mathbf{Y}}$ are not binary (0 and 1) but are real numbers. The merit of the DPLS method is the orthogonality of the predictor block (genes). Using DPLS modeling, the orthogonal transform of the predictor block makes each gene independent and hence we need not consider gene-gene correlations. Note that DPLS is a variation on PLS regression and therefore represents an application of the regression method to pattern classification.

AML(Acute Myeloid Leukemia)

The AML dataset from the oligonucleotide chip [1] has a large number of variables (7129 human genes) but few samples (14 patients). Prior to multivariate statistical analysis, it is necessary to remove some genes which cannot affect therapeutic response. To extract the informative gene, we performed the leave-one-out *T-test* method because this method gives a more robust result than the general *T-test* for gene selection. First of all, *T*-values of all genes are calculated except the first sample and then the large negative and positive 50 genes are selected, respectively. After leave-one-out procedure, we can obtain common 20 genes that always belong to relevant 100 genes. The constructed PCA model for AML patients' clinical outcome monitoring has 2 principal components which explain 70% variance. Using the above methods, we revealed the biological and medical connectivity of genes and clinical outcome of AML patients successfully [9].

DLBCL (Diffuse Large B-Cell Lymphoma)

In this study, we finally obtained 35 variables through leave-one-out VIP and made DPLS model with 7 misclassifications in the cross validation. Although the stepwise-DPLS may improve the classification power, it is not applied because classification performance is not the main focus of this research. The aim of this paper is to identify what genes are key factors to determine the clinical outcome (cured or fatal). The clinical outcome and its related genes are elucidated by looking at both the score plot and loading one [9]. Shipp *et al.* [7] showed that L20971 (PDE4B), M18255 (PKC-β protein) and U12767 (MINOR) were clearly correlated with clinical outcome and the same results were extracted from the suggested biplot analysis. The proposed procedure gives a more powerful class prediction performance and therapeutic guideline than the previous research.

Conclusion

We modified the typical VIP and *T*-test for more rigorous gene selection in order to remove unusual genes that are highly expressed in some patients, but not expressed in other patients. Then, the selected genes were highly informative in a view point of molecular biology. We used two biplot analysis methods and these methods are successfully implemented to elucidate the relationship between therapy response and its related genes. The PCA gives a good graphical representation between the AML patients' clinical outcome and related genes. In DLBCL data, the regressive property of DPLS makes it possible to analyze not only classification performance but also consider the interrelationship between genes in the model during the gene selection procedure. The prediction performance of the proposed method was estimated by cross-validation and the suggested method showed more satisfactory result than that of Shipp *et al.* [7]. This kind of analysis seems to be a promising method because it is easy to use and gives additional information. Also, these methods can also be used to identify the therapy efficacy and monitor the patient's status.

Acknowledgement

This work was supported by the Brain Korea 21 project.

References

(1) Golub, T. R.; Slonim, D. K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J. P.; Coller, H.; Loh, M. L.; Downing, J. R.; Caligiuri, M. A.; Bloomfield, C. D.; Lander, E. S., *Science*, 1999, **286**, 531

(2) Alizadeh, A. A.;Eisen, M. B.; Davis, R. E.; Ma, C.; Lossos, I. S.; Rosenwald, A.; Boldrick, J. C.; Sabet, H.; Tran, T.; Yu, X.; Powell, J. I.; Yang, L.; Marti, G. E.; Moore, T.; Hudson Jr., J.; Lu, L.; Lewis, D. B.; Tibshirani, R.; Sherlock, G.; Chan, W. C.; Greiner, T. C.; Weisenburger, D. D.; Armitage, J. O.; Warnke, R.; Levy, R.; Wilson, W.; Grever, M. R.; Byrd, J. C.; Botstein, D.; Brown, P. O.; Staudt, L. M., *Nature*, 2000, **403**, 503

(3) Lee, D.; Choi, S.; Park, J. H.; Lee, I., *AIChE Annual Meeting*, 2001

(4) Choi, S.; Lee, D.; Park, J. H.; Lee, I., *AIChE Annual Meeting*, 2001

(5) Lee, D.; Park, J. H.; Choi, S.; Kim, M.; Lee, I.; Kim, Y. H.; Im, S. U.; Chung, E. J.; Kim, M.; Kim, J., *KIChE 2001 Fall Meeting*, 2001

(6) van 't Veer L. J.; Dai H.; van de Vijver M. J.; He Y. D.; Hart A. A.; Mao M.; Peterse H. L.; van der Kooy K.; Marton M. J.; Witteveen A. T.; Schreiber G. J.; Kerkhoven R. M.; Roberts C.; Linsley P. S.; Bernards R.; Friend S. H., *Nature*, 2002, **415**, 530

(7) Shipp, M. A.; Ross, K. N.; Tamayo, P.; Weng, A. P.; Kutok, J. L.; Aguiar. R. C. T.; Gaasenbeek, M.; Angelo, M.; Reich, M.; Pinkus, G. S.; Ray, T. S., Koval, M. A.; Last, K. W.; Norton, A; Lister, T. A.; Mesirov, J.; Neuberg, D.; Lander, E. S.; Aster, J. C.; Golub, T. R., *Nature Medicine*, 2002, **8**, 68

(8) Cho, J.; Lee, D.; Park, J. H.; Kim, K.; Lee, I., *Biotechnol. Prog.* 2002, **18**, 847

(9) Lee, D.; Choi, S. W.; Cho, J.; Park, J. H.; Lee, I., *Biotechnol. Prog. submitted* 2002