A Novel Feature Selection Method to Improve Classification of Gene Expression Data

Liang Goh, Qun Song, and Nikola Kasabov
Knowledge Engineering and Discovery Research Institute
Auckland University of Technology
Private Bag 92006, Auckland 1020, New Zealand
Email: liang.goh@aut.ac.nz; qsong@aut.ac.nz; nkasabov@aut.ac.nz

Abstract

This paper introduces a novel method for minimum number of gene (feature) selection for a classification problem based on gene expression data with an objective function to maximise the classification accuracy. The method uses a hybrid of Pearson correlation coefficient (PCC) and signal-to-noise ratio (SNR) methods combined with an evolving classification function (ECF). First, the correlation coefficients between genes in a set of thousands, is calculated. Genes, that are highly correlated across samples are considered either dependent or co-regulated and form a group (a cluster). Signal-to-noise ratio (SNR) method is applied to rank the correlated genes in this group according to their discriminative power towards the classes. Genes with the highest SNR are used in a preliminary feature set as representatives of each group.

An incremental algorithm that consists of selecting a minimum number of genes (variables) from the preliminary feature set, starting from one gene, is then applied for building an optimum classification system. Only variables, that increase the classification rate in each of the validation iteration, are selected and added to the final feature set. The results show that the proposed hybrid PCC, SNR and ECF method improves the feature selection process in terms of number of variables required and also improves the classification rate. The classification accuracy of the ECF classifier is tested through the leave one out method for validation.

Keywords: feature selection, gene expression, microarray, connectionist classification systems.

1. Introduction

A problem with gene expression analysis, or with any large dimensional data set, is often the selection of significant variables (feature selection) within the data set that would enable accurate classification of the data to some output classes. These variables may be potential diagnostic markers too. There are good reasons for reducing the large number of variables: (a) an opportunity to scrutinise individual genes for further medical treatment and drug development; (b) reducing the number of redundant and unnecessary variables can improve inference and classification (Ramsey and Schafer, 2002).

In genomics expression, the data set is usually plagued with large number of variables versus the small number of records or vectors (the problem is known as the 'curse of dimensionality'). Genes are clustered first, and usual methods used are K-means clustering and hierarchical clustering (Lukashin and Fuchs, 2001, Gad et al., 2000, Eisen et al., 1998), Singular Value Decomposition or Principal Component Analysis (Alter et al., 2000), supervised clustering and fuzzy clustering methods (Gasch and Eisen, 2002, Futschik, 2002), self-organizing maps (Tamayo et al., 1999, Alizadeh et al., 2001). Feature selection methods are then applied (e.g. signal noise ratio, correlation coefficient, p-values, etc) (Goh and Kasabov, 2003, Magrath, 2002, Miller et al., 2002, Shipp et al., 2002a, Yeoh et al., 2002, Ramaswamy et al., 2001) to extract the significant variables.

Feature selection is the process of choosing the most appropriate features when creating the model of the process (Kasabov, 2002). Most of the feature selection methods are applied across the entire data set. In Van Veer approach (Veer et al., 2002), the genes were correlated against the disease outcome and then ranked. With increment of 5 genes for each classification model, a set of 70 genes was then identified. In Shipp's approach, signal noise ratio was used to identify 30 genes (Shipp et al., 2002a). The set of variables were then used for building an outcome prognosis model. While there has been success in the use of these approaches, no clear theoretical advantage exists for any given algorithm. Some methods apply multiple algorithms (Ochs and Godwin, 2003).

In this paper, we propose that multiple passes of different feature selection methods can be applied to a data set to remove redundant variables and therefore improve the effect of each feature selection method at each pass. We demonstrate the effects of applying multiple pass of feature selection using PCC and SNR in the first pass and using SNR for the second pass before the ECF

classification method is applied to build a classifier with the selected variables. Note that the method is not restricted to just two passes of feature selection, neither to using only PCC and SNR. Other feature selection methods can be used at each level and the number of passes can be adjusted

The rationale of the approach is based on the idea that most feature selection methods are based on some assumptions of the data. Using a hybrid approach could overcome the shortcomings of each method. The hybrid method introduced here attempts to select the features for each level of its pass, so that after each pass, the redundant and unnecessary variables in the data are reduced to allow an improved set of significant variables to be extracted. With the use of a large number of variables the apparent (training) error rate of the classifier may decrease, but the generalisation ability of the model may decrease too (Ambroise and McLachlan, 2002). A feature selection that aims to reduce the number of genes could improve the generalization of the classifier.

The computation time of calculating the PCC matrix increases exponentially with the size of the data set. In this paper, we also present a novel way of reducing the computation time of PCC calculation.

The validation of the hybrid method is done using an evolving classifier ECF, a supervised clustering method (Kasabov, 2002). To avoid bias in the validation, a hold-out approach was adopted for a second set of experiments as explained in the next sections.

The cancer data set used in the experiments is from (Shipp et al., 2002a) (classification of DLBCL and Follicular Lymphoma). The lymphoma data set contains 77 vectors: 58 for DLBCL, and 19 for Follicular Lymphoma. There are 7129 gene variables.

2. Brief Introduction to PCC and SNR

Linear correlation coefficient is a measurement of the strength of a linear relationship between a dependent variable (i.e. the output class, *y*) and the independent variable (i.e. the genes, *x*).

$$r = \Sigma(x-x_{mean})(y-y_{mean})/ (\sigma_x + \sigma_y)$$

When *x* increases and if *y* also tends to increase or decrease, there is a mathematical linear dependency between *y* and *x*. How strong is this dependency? The *r* calculated by PCC gives a quantitative idea of the dependency. The correlation value varies from –1 to 1. A value of 0 suggests no linear correlation, while values nearer to –1 or 1 means negatively or positively correlated. PCC is for bivariate analysis, and provides a quick way to estimate linear relationship for data that has a normal distribution. However, for large data set, the computation time to calculate the PCC matrix is very long. Here we propose a novel way to calculate the matrix in a shorter time.

SNR is a calculated ranking number for each variable to define how well this variable discriminates two classes. The following formula is used:

$$(\mu_{class\ 1} - \mu_{class\ 2})/(\sigma_{class\ 1} + \sigma_{class\ 2})$$

where: $\mu_{class\ 1}$ and $\mu_{class\ 2}$ are the mean values for this variable for the samples from class 1 and class 2 respectively and $\sigma_{class\ 1}$ and $\sigma_{class\ 2}$ are the corresponding standard deviations.

This method is combined with a weighted voting classifier in (Shipp et al., 2002b). SNR measurement is affected by the size of the variables, as can be seen from the formula. When there are more variables, the mean and variance of the rest of variables of other classes are dependent on the data dispersion and the number of variables, which affects SNR ranking of the significant variables due to the general increase of noise in the data. If the number of variables can be reduced significantly, the SNR method is more capable of detecting and ranking a smaller number of significant variables.

3. Introduction to Evolving Classifier Function

Evolving connectionist systems (ECOS) are systems that evolve their structure and functionality over time from incoming information (Kasabov, 2002). The ECF (Evolving Classifier Function) model used here is a connectionist system for classification tasks that consists of four layers of neurons (nodes). The first layer represents the input variables; the second layer – the fuzzy membership functions; the third layer represents clusters centers (prototypes) of data in the input space; and the four layer represents classes (Kasabov, 2002).
The learning algorithm of the ECF is as follows:

1. If all vectors have been input, finish the current iteration; otherwise, input a vector from the data set and calculate the distances between the vector and all rule nodes already created using Euclidean distance by default.
2. If all distances are greater than a max-radius parameter, a new rule node is created. The position of the new rule node is the same as the current vector in the input data space and the radius of its receptive field is set to the min-radius parameter; the algorithm goes to step 1; otherwise it goes to the next step:
3. If there is a rule node with a distance to the current input vector less then or equal to its radius and its class is the same as the class of the new vector, nothing will be changed; go to step 1; otherwise:
4. If there is a rule node with a distance to the input vector less then or equal to its radius and its class is different from those of the input vector, its influence field should be reduced. The radius of the new field is set to the larger value from the two numbers: distance minus the min-radius; min-radius. New node is created as in (2) to represent the new data vector.
5. If there is a rule node with a distance to the input vector less than or equal to the max-radius, and its class is the same as of the input vector's, enlarge the influence field by taking the distance as a new

radius if only such enlarged field does not cover any other rule nodes which belong to a different class; otherwise, create a new rule node in the same way as in step 2, and go to step 1.

The recall procedure (classification of a new input vector) in the trained ECF is performed in the following way:

1. If the new input vector lies within the field of one or more rule nodes associated with one class, the vector is classified in this class;
2. If the input vector lies within the fields of two or more rule nodes associated with different classes, the vector will belong to the class corresponding to the closest rule node.
3. If the input vector does not lie within any field, then there are two cases: (i) one-of-n mode: the vector will belong to the class corresponding the closest rule node; (ii) m-of-n mode: take m highest activated by the new vector rule nodes, and calculate the average distances from the vector to the nodes with the same class; the vector will belong to the class corresponding the smallest average distance.

The ECF model used in the paper has the following parameter values: MaxField=1.0, MinField= 0.02, number of membership functions MF=1 (no fuzzy membership functions); number of rule nodes used to calculate the output value of the ECF when a new input vector is presented MofN=1 (number of neighbors to consider when evaluating nearest node); number of iterations for presenting each input vector Epochs=5.

4. PCC and SNR Hybrid Method for Preliminary Feature Selection

The algorithm for the hybrid feature selection is as follows:

1. Set a threshold for the Pearson Correlation Coefficient, e.g. $P_{threshold}$ = 0.6. For each iteration, $i=1,\ldots n$ (where n is the number of variables in the data set) calculate the correlation coefficient r_{ij} for x_i with respect to x_j, $j=i,..,n$.
2. Select those variables where r_{ij} is greater than $P_{threshold}$, i.e. $PCC(x_{ij}) > P_{threshold}$, where $x_{ij} \in S_i$, and S_i is the set that contains the correlated variables for variable x_i at iteration i.
3. Apply SNR on S_i and select the highest ranked variable to represent this group of correlated variables: $x_i = max(SNR(x_{ij}))$, $x_{ij} \in S_i$. This will remove redundant variables that are 'similar' to each other in the current iteration, selecting only the best variable to represent S_i, therefore reducing the size of the original matrix for future PCC calculation. At the end of all iterations, a much reduced data set D is obtained. Note that there are three variations to selecting the representative variables that will be mentioned later in this section.
4. Perform SNR on D and rank the variables according to their SNR value. Set thresholds for variable cut-off, $V_{threshold}$ and classification threshold $C_{threshold}$ e.g. $V_{threshold}$ = 100, $C_{threshold}$ = 1 (100% classification rate). This will set the limit in which the validation process will stop, based on maximum number of SNR-ranked variables that are investigated or given classification accuracy is achieved. Select the increment step in which number of variables are selected, e.g. $V_{step} = 1$. This will determine how many variables to increment for each step of validation.
5. Start with the highest ranked SNR variables, i.e. v_k, $v_k=\{L, rank(SNR(D),k)\}$, $k=1,..m$, where m is the number of variables in D, and L is the list of successful variables which is null in the beginning. Validate the variables using ECF leave-one-out method. If the average classification rate is less than $C_{threshold}$, or less than the $V_{threshold}$, add variable to the list, L= $\{L, v_k\}$. Add the next set of variables from the ranked SNR by the amount of V_{step}.
6. If the set average classification rate is achieved (e.g.100%) or variable threshold is reached, stop the process. The set of variables in L will be the set that has the highest classification rate amongst the iterations.

In the algorithm, the computation time is greatly reduced as variables that behave 'similarly' are grouped and removed for further selection, thereby decreasing the size of data for further correlation coefficient calculation. In this paper, three approaches are used: Method 1: selecting the first correlated variable in S_i; Method 2: selecting the variable highly correlated with output class in S_i; and Method 3: selecting a variable ranked highest in SNR in S_i (presented in the algorithm). In another implementation, instead of selecting the next variable incrementally in step (5) of the algorithm, only variables that improve the classification rate in each validation iteration are selected and added to the final variable set. We shall denote this variation to the algorithm as Method 1a, Method 2a, and Method 3a with respect to the three methods described above.

In our experience, the calculation of the entire correlation matrix for gene expression data is computationally intensive that often takes a long time to calculate. Using this approach of reducing the data by removing highly correlated variables (which are redundant) we are able to get a reasonable matrix of correlation coefficient of the remaining variables. Selecting the set of variables incrementally based on the classification rate makes it possible that the selected set of variables that can classify the entire data set will be the optimum as every variable in the list L is selected only if it increases the classification rate.

The criterion of setting a maximum number of variables, $V_{threshold}$ as cut-off for the validation process is also an advantage feature of the method introduced here. We selected 100 as the number of the cut-off point, based on

the fact that previous research has shown good classification results with genes less than this number.

To avoid selection bias in the validation (Ambroise and McLachlan, 2002), another experiment was done where the data set was divided into training and testing sets, D_{train} and D_{test}. The hybrid feature selection method was run on D_{train} where a set of variables was selected. This set of variables was used to build an ECF model and D_{test} was then tested on the model. As the data set was not large in terms of the number of sample vectors, leave one out cross validation method was selected, i.e. one sample vector was selected for the testing set in each iteration of the validation and the rest of the data samples were used for training an ECF model.

5. Analysis of Results

Results of the lymphoma cancer data are shown in Table 1. The first column indicates the data set for the experiment and the method that was applied, second column shows the correlation coefficient threshold set, third column shows the number of variables that was selected using SNR, and the last column shows the validation classification rate (leave one out) using ECF. Other experiments using different correlation coefficient settings ranging from 0.2 to 0.9 have been done. The best performance of 0.6 is presented here.

Data	PCC threshold	# Var selected	Classific. Rate
Lymphoma – Method 1	0.6	15	100%
Lymphoma – Method 2	0.6	60	100%
Lymphoma – Method 3	0.6	61	100%
Lymphoma –Method 1a	0.6	9	100%
Lymphoma– Method 2a	0.6	11	98.7%
Lymphoma– Method 3a	0.6	10	98.7%
Lymphoma – using variables highly correlated to output	0.7	282	74.03%
Lymphoma – using top variables according to SNR	-	205	100%

Table 1 Experimental results of lymphoma cancer data using different correlation coefficient settings. The number of variables selected and classification rate are shown in the last 2 columns. Method 1: selecting first correlated variable in S_i, Method 2: selecting variable most correlated to output class in S_i; Method 3: selecting variable ranked highest by SNR. The methods with suffix 'a' mean the alternative algorithm of selecting variable only when it improves the validation rate. The traditional approaches were to use variables most correlated to output class or to apply SNR on the entire data set.

The results demonstrate that for the lymphoma data, less number of variables and higher classification rate is achieved with the use of the introduced here hybrid method. The proposed hybrid PCC and SNR method improves the feature selection process in terms of number of variables required and also improves the classification rate as compared with the standard procedure of: (1) calculate the correlation coefficients between genes and classes and choose the highly correlated as a set of features, or/and (2) apply SNR on the whole set of genes and rank them with a consecutive selection of the top genes to build a classifier(Shipp et al., 2002a, Shipp et al., 2002b).

In the case of lymphoma data, when using 0.6 for the correlation coefficient threshold, 15 variables were selected from the 7129 variables using Method 1. This gave a classification rate of 100%, which is higher than Shipp's experiment, and with less number of genes identified. In Shipp et al, 30 genes were identified and classification rate of 91% was obtained. In the alternative algorithm 1a, 9 variables were selected with a classification of 100%, which is even better. The list of genes is shown in Table 3. The results show the potential of the hybrid method in identifying a lower number of variables. Using SNR on its own (i.e. traditional approach), the number of variables was much higher in comparison (205 for lymphoma), with about similar classification rate.

Table 2 shows the accuracy of the model when unbiased (hold-out) experiment is conducted. In this approach, the PCC are calculated not on the whole data set, as it was done so far, but at every leave-one-out iteration. We were interested to achieve a lower number of variables with highest classification rate and the experiments show that the alternative algorithm gave a more consistent performance in terms of less number of variables selected as well as highest classification rate.

Data	PCC	Average Number of Variables Extracted	Average Rate
Lymphoma	0.6	10	91%

Table 2 Experiment results of lymphoma data with hold-out.

The hold-out classification rate of 91% is significant compared with Shipp's bias rate of the same. It is known that classification rate for un-biased selection would yield a lower accuracy than applying the biased approach (Ambroise and McLachlan, 2002). In addition, the number of variables selected was much less than the 30 variables selected by Shipp. In cases where a low number of variables are desirable, e.g. drug targeting of the disease or development of blood test kit on proteins, it is easier to work with less number of genes, which means less number of proteins to be tested.

Fig. 1 shows the principal component analysis (PCA) plot of the first two components for the extracted data set of 15 variables for method 1 (see table 1). The plot shows that the two classes of DLBCL and FL are separable. The blue circle represents the DLCBL class and red circle FL class. Fig 2 and 3 show the PCA plots of the various methods from table 1. Fig 4 and 5 show the PCA plots with sample vectors that were incorrectly classified.

Methods 1, 2 and 3 yield 100% classification but at the expense of a larger set of variables. Methods 1a, 2a and 3a has slightly lower classification rate but the rates are consistent and the variables selected are much less. The results show that the variables selected by the hybrid approach were significant in its expression within each vector, such that they exhibit similar trends, enough to be well clustered by ECF. This results in the higher classification rate obtained in the experiments as compared to the traditional approaches.

Generally, the individual feature selection approach would probably also yield vectors with similar trend, good enough for clustering, but because a larger number of variables are selected, the analysis becomes increasingly tedious and complicated. As mentioned, simplicity in terms of less number of variables can improve inference and classification.

'Bcl-2 related (Bfl-1) mRNA'
'TFRC Transferrin receptor (p90, CD71)'
'ADA Adenosine deaminase'
'SLC'
'PLOD Lysyl hydroxylase'
'TYMS Thymidylate synthase'
'SERUM PARAOXONASE/ARYLESTERASE'
'(clone GPCR W) G protein-linked receptor gene m (GPCR) gene, 5' end of cds'
'Tryptophan hydroxylase (Tph) mRNA'

Table 3. The list of 9 genes selected in the modelling experiment with the use of Method1a that gives 100% accuracy of classification between the DLBCL and the Follicular Lymphoma data (Shipp et al., 2002a).

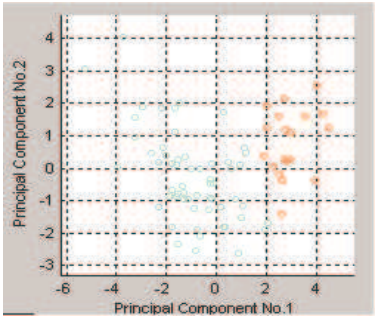

Figure 1 PCA plot of DLBCL (blue) versus FL (red) for the 15 variables selected using Method 1

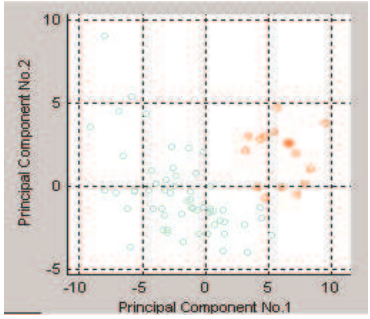

Figure 2 PCA plot of DLBCL (blue) versus FL (red) for the 60 variables selected using Method 2.

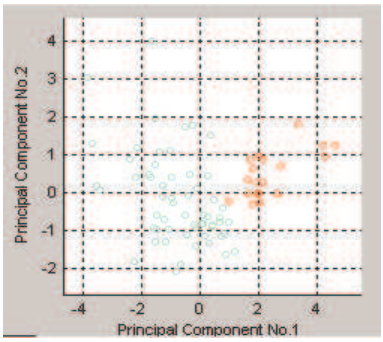

Figure 3 PCA plot of DLBCL (blue) versus FL (red) for the 9 variables selected using Method 1a

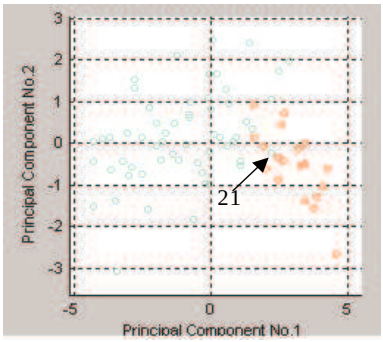

Figure 4 PCA plot of DLBCL (blue) versus FL (red) for the 11 variables selected using Method 2a. Sample vector 21 was classified incorrectly which corresponds to the blue circle amidst the red circles on the centre right side.

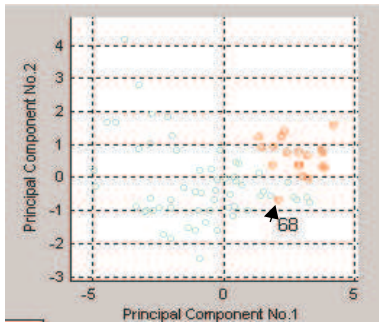

Figure 5 PCA plot of DLCBL (blue) versus FL (red) with variables using Method 3a. Vector sample 68 was classified incorrectly and corresponds to the red circle amidst the blue ones on the lower right.

6. Conclusion

The hybrid feature selection method described in this paper has demonstrated that the approach is able to reduce the number of genes selected as well to increase the classification rate on a case study Lymphoma classification data. Like ensemble of experts, each feature selection method at each stage of processing removes the redundant variables and thereby reducing the noise in the data to allow for a better set of features to be selected in the next stage. The method has been successfully tested

on the lymphoma data set. Though the data set is related to cancer, the method is generic and can be applied on other large data sets that require feature selection.

7. Future Work

There is a potential in using this approach, which is a divide and conquer approach, to analyse clusters of correlated genes, and discover the functional genomics of these groups. In our experiment, a gene is selected out of every correlated cluster in the hybrid approach. This can be further explored to discover if there is any common functional genomics within these groups, and if the selected gene can be better represented by another one – more appropriate for clinical applications.

8. Acknowledgement

The research presented in the paper is funded by the New Zealand Foundation for Research, Science and Technology under the grant: NERF/AUTX02-01.

9. References

Alizadeh, A. A., Ross, D. T., Perou, C. M. and van de Rijn, M. (2001) *Journal of Pathology*, **195,** 41-52.

Alter, O., Brown, P. O. and D, B. (2000) *Proceedings of the National Academy of Sciences of the United States of America*, **97,** 10101-10106.

Ambroise, C. and McLachlan, G. J. (2002) *Proceedings of the National Academy of Sciences of the United States of America*, **99,** 6562.

Eisen, M. B., Spellman, P. T., Brown, P. O. and D, B. (1998) *Proceedings of the National Academy of Sciences of the United States of America***,** 14863-14868.

Futschik, M. E. K., N.K. (2002) In *Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on*, Vol. 1, pp. 414-419.

Gad, G., Erel, L. and Eytan, D. (2000) *Proceedings of the National Academy of Sciences of the United States of America [H.W. Wilson - GS]*, **97,** 12079.

Gasch, A. P. and Eisen, M. B. (2002) *Genome Biol.3;RESEARCH0059.*

Goh, L. and Kasabov, N. (2003) *International Joint Conference on Neural Networks*.

Kasabov, N. K. (2002) *Evolving Connectionist Systems, Methods and Applications in Bioinformatics, Brain Study and Intelligent Machines*, Verlag Springer.

Lukashin, A. V. and Fuchs, R. (2001) *Bioinformatics*, **17,** 405-414.

Magrath, I. (2002) *The New England Journal of Medicine*, **346,** 1998.

Miller, L. D., Long, P. M., Wong, L., Mukherjee, S., McShane, L. M. and Liu, E. T. (2002) *Cancer Cell*, **2,** 353-361.

Ochs, M. F. and Godwin, A. K. (2003) *BioTechniques*, **34,** 4-15.

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S. and et al. (2001) *Proceedings of the National Academy of Sciences of the United States of America*, **98,** 15149.

Ramsey, F. L. and Schafer, D. W. (2002) *The Statistical Sleuth, a course in methods of data analysis*, Duxbury Learning Publishing.

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. S., Lander, E. S., Aster, J. C. and Golub, T. R. (2002a) *Nature Medicine*, **8,** 68-74.

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. S., Lander, E. S., Aster, J. C. and Golub, T. R. (2002b) *Nature Medicine*, **8,** 68-74.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R. (1999) *Proceedings of the National Academy of Sciences of the United States of America*, **96,** 2907-2912.

Veer, L. J. v. t., Dai, H., Vijver, M. J. v. d., He, Y. D. and et al. (2002) *Nature*, **415,** 530.

Yeoh, E. J., Ross, M. E., Shurtleff, S. A., Williams, W. K., Patel, D., Mahfouz, R., Behm, F. G., Raimondi, S. C., Relling, M. V., Patel, A., Cheng, C., Campana, D., Wilkins, D., Zhou, X. D., Li, J. Y., Liu, H. Q., Pui, C. H., Evans, W. E., Naeve, C., Wong, L. S. and Downing, J. R. (2002) *Cancer Cell*, **1,** 133-143.