The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010
Copyright © 2010 ORSC & APORC, pp. 43–50

Improving MSVM-RFE for Multiclass Gene Selection[*]

Yan-Mei Zhao[2] Zhi-Xia Yang[†1]

[1]College of Mathematics and System Science, Xinjiang University, Urumuchi China,830046
[2]College of Science, China Agricultural University, Beijing, China,100083

Abstract Along with the advent of DNA microarray technology, gene expression profiling has been widely used to study molecular signatures of many diseases and to develop molecular diagnostics for disease prediction. In class prediction problems using expression data, gene selection is essential to improve the prediction accuracy and to identify informative genes for a disease. In this paper we improve the multi-class support vector machine-recursive feature elimination (MSVM-RFE) by combining minimum redundancy maximum relevancy (mRMR) criterion and introducing the kernel. The result is the better performance with a smaller number of irredundant genes for multi-class datasets.

1 Introduction

Microarray technology allows us to measure the expression levels of thousands of genes simultaneously resulting in a vast pool of data. Normally, the gene expression dataset contains small number of samples and very large number of genes. This characteristic makes gene selection a necessary procedure to improve the classification accuracy. Furthermore, the selected handful important genes can be used to design less expensive experiments for disease prediction.

Among the existing numerous gene selection methods, support vector machine-based recursive feature elimination (SVM-RFE) [1]is a widely used method conducting gene selection in a backward elimination procedure. It was initially proposed for binary classification. The gene selection for multi-class gene expression data has begun to draw more and more attention from researchers. SVM-RFE has been extended into multi-class MSVM-RFE to solve multi-class problems using one-versus-all [2, 3] or all together techniques [4]. However, the genes selected using MSVM-RFE may have a degree of redundancy which may reduce the classification performance.

In this paper we improve MSVM-RFE by combining minimum redundancy maximum relevancy (mRMR) criterion [5], resulting a new mixed method. We also introduce kernel for MSVM-RFE to overcome the limitation of linear MSVM-RFE. The proposed mixed method provides minimum redundancy genes which broadly represent whole gene expression data leading to more accurate classification.

[*]This work is supported by the National Science Foundation of China (No.10801112) and the Ph.D Graduate Start Research Foundation of Xinjiang University Funded Project (No.BS080101).

[†]Corresponding author. E-mail: xjyangzhx@sina.com

2 Background

2.1 Crammer and Singer multiclass SVM (CSSVM)

The CSSVM[8] is a multiclass method of “all-together" implementation by solving one single optimization problem. Given the training set $T = \{(x_1, y_1), \cdots, (x_l, y_l)\}$, where $x_i \in R^n$ is the input, and $y_i \in \{1, \cdots, k\}$ is the output or the class label. The input x is mapped into a Hilbert space $\mathscr{H}$ by a function $\mathrm{x} = \Phi(x) : x \in R^n \to \mathrm{x} \in \mathscr{H}$. It is required to find k hyperplanes to construct the decision function $f(x) = \arg\max_{r=1,\cdots,k} \sum_{i=1}^{l} \alpha_i^r K(x_i, x)$, where $K(x, x') = (\Phi(x) \cdot \Phi(x'))$ is the kernel function.

To get $\alpha_i^r, r = 1, \cdots, k, i = 1, \cdots, l$ in the decision function, we need to solve the dual problem:

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} K(x_i \cdot x_j)\alpha_i^{\mathrm{T}}\alpha_j - \sum_{i=1}^{l}\alpha_i^{\mathrm{T}} e_i, \tag{1}$$

$$\text{s.t.} \quad \sum_{r=1}^{k} \alpha_i^r = 0, i = 1, \cdots, l, \tag{2}$$

$$\alpha_i^r \leqslant 0, \ \ if \ \ y_i \neq r, i = 1, \cdots, l.r = 1, \cdots, k, \tag{3}$$

$$\alpha_i^r \leqslant C, \ \ if \ \ y_i = r, i = 1, \cdots, l.r = 1, \cdots, k, \tag{4}$$

where $\alpha = (\alpha_1, \cdots, \alpha_l), \alpha_i = (\alpha_i^1, \cdots, \alpha_i^k)^{\mathrm{T}}, i = 1, \cdots, l, e_i = (e_i^1, \cdots, e_i^k)^{\mathrm{T}}, e_i^r = 1 - \delta_{y_i, r}, i = 1, \cdots, l, r = 1, \cdots, k$, and $K(x, x') = (\Phi(x) \cdot \Phi(x'))$ is the kernel function.

2.2 Vector Output Support Vector Machine(VOSVM)

The VOSVM[9] is another multi-class method which comes from a simple reinterpretation of the normal vector of the separating hyperplane. Given the training set $T = \{(x_1, y_1), \cdots, (x_l, y_l)\}$, where $x_i \in R^n$ is the input, $y_i \in \{1, \cdots, k\}$ is the output. In order to establish the decision function, the output y_i is mapped into $\hat{y}_i \in R^k$, and we define $\hat{y}_i = ((\hat{y})_1, \cdots, (\hat{y})_k)^T$ as follows

$$(\hat{y}_i)_t = \begin{cases} 1, & \text{if item } i \text{ belongs to } t\text{-th class}, t = 1, \cdots, k; \\ -\dfrac{1}{k-1}, & \text{if others.} \end{cases} \tag{5}$$

The decision function is presented as $f(x) = \arg\max_{r=1,\cdots,k} \hat{y}_r^{\mathrm{T}} \sum_{i=1}^{l} \alpha_i y_i K(x_i, x)$, where $K(x, x') = (\Phi(x) \cdot \Phi(x'))$ is the kernel function. To get $\alpha_i, i = 1, \cdots, l$ in the decision function, the following QP problem is solved

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i\alpha_j(y_i \cdot y_j)K(x_i, x_j) + \sum_{i=1}^{l}\alpha_i, \tag{6}$$

$$\text{s.t.} \quad 0 \leqslant \alpha_i \leqslant C, i = 1, \cdots, l, \tag{7}$$

where $\alpha = (\alpha_1, \cdots, \alpha_l)^T$ and $K(x, x') = (\Phi(x) \cdot \Phi(x'))$ is the kernel function.

2.3 SVM-RFE and its extension

The SVM-RFE([1])is a simple and efficient method which conducts gene selection in a backward elimination procedure for the standard SVM. This method starts from all features and eliminates one feature at a time. The square coefficient $\frac{1}{2}w_j^2, j = 1,\cdots,n$ of weight vector $w = (w_1,\cdots,w_n)^T$ obtained from the primal problem of the standard SVM can be used as a criterion for feature ranking. In fact,it removes the feature with smallest ranking criterion. Repeat this procedure until obtaining the smaller feature subset. Afterward, [4] has extended SVM-RFE to linear multi-class SVM. Similar to the derivation of ranking criterion for binary SVM-RFE, $\frac{1}{2}\sum_{r=1}^{k}(w_{rj})^2, j = 1,\cdots,n$ is obtained as an appropriate ranking criterion for linear "all-together" SVMs. In addition, [1] has introduced kernel for the binary SVM-RFE but there is no corresponding research for MSVM-RFE. So we concentrate on the kernel MSVM-RFE in this paper.

2.4 The minimum Redundancy and Maximum Relevancy (mRMR) criterion

The mRMR criterion is to find the feature subset in which features have maximal relevancy to class and minimum redundant among themselves. The detail can be seen in [5].

3 Kernel MSVM-RFE combined with mRMR criteria

In this section, we propose a variant of MSVM-RFE, which extends MSVM-RFE to non-linear case by introducing kernel function and combines with mRMR criterion.

Let us recall the optimization problems (1)∼(4) and (6)∼(7), both of their objective functions can be formulated as follows:

$$J = \frac{1}{2}\alpha^{\mathrm{T}}H\alpha - e^{\mathrm{T}}\alpha, \tag{8}$$

where α is the solution of dual problem (1)∼(4)(or (6)∼(7)) and H is a matrix with the kernel.

To compute the weightiness of the j-th feature, which is deleted. One leaves the α unchanged and one re-computes matrix $H(-j)$, where the notation $(-j)$ means that component j has been removed. Then the following equation evaluate the change in objective function (8) caused by removing the j-th feature

$$DJ(j) = \frac{1}{2}\alpha^{\mathrm{T}}H\alpha - \frac{1}{2}\alpha^{\mathrm{T}}H(-j)\alpha. \tag{9}$$

The input corresponding to the smallest $DJ(j)$ will be eliminated. This criterion can be used in the kernel MSVM-RFE. Specially in the linear case, $\frac{1}{2}\alpha^{\mathrm{T}}H\alpha = \frac{1}{2}\sum_{r=1}^{k}||w_r||^2$, therefore $DJ(j) = \frac{1}{2}\sum_{r=1}^{k}{w_{rj}}^2$. So the method is identical to the linear MSVM-RFE.

Now we give the detailed algorithm by combining the kernel MSVM-RFE with mRMR criterion (K-MSVM-RFE+mRMR) for the feature selection as follows.

Algorithm: K-MSVM-RFE+mRMR

(1)Given the training set $T = \{(x_1,y_1),\cdots,(x_l,y_l)\}$, $x_i = ([x_i]_1,\cdots,[x_i]_n) \in R^n$ is the input, $y_i \in \{1,2,\cdots,k\}$, $i = 1,\cdots,l$ is the output, $g_j = ([x_1]_j,\cdots,[x_l]_j)^{\mathrm{T}}$ is the feature

Table 1: The information of four datasets

Dataset	Samples	Number of features		Classes
		Original	Pre-processed	
NCI Staunton	58	7129	3144	8
MLL	72	12582	10930	3
11Tumors	174	12533	9700	11
GCM	198	16063	14122	14

vector, which comprises the j-th feature of all inputs, where $j = 1,\cdots,n$, $c = \{1,\cdots,k\}$ is the target class;

(2)Initialize $Z = [1,\cdots,n]$ be the subscript index of candidate feature set $S = [g_1,...,g_n]$, Ranked feature set $R = [\quad]$;

(3)Solve the optimization (1)$\sim$(4)(or (6)$\sim$(7))using the features in Z, and get the optimal solution α;

(4)Compute $DJ(j)$ given by (9);

(5)Evaluate the relevancy of each feature to class and mutual information to features in Z $m_j = \dfrac{I(g_j;c)}{\frac{1}{|Z|}\sum\limits_{i\in Z} I(g_j;g_i)}$, where $I(x;x')$ is the mutual information of x and x';

(6)Calculate h_j the score of each feature by $h_j = \dfrac{|DJ(j)|}{\max\limits_{j\in Z}(|DJ(j)|)} + \dfrac{m_j}{\max\limits_{j\in Z}(m_j)}$;

(7)Remove the feature with the minimum score $e = \arg\min\limits_{j\in Z} h_j$;

(8)Eliminate e from Z, then add it into R, $R = [e;R]$;

(9)If $Z = [\quad]$, stop; otherwise, go to step (3).

Note: the MATLAB sentence $[\quad]$ is used as an empty set.

4 Numerical Experiments

In order to test the effectiveness of our feature selection methods, we choose four multiclass gene expression datasets: NCI Staunton ([10]), MLL ([11]), 11Tumors ([12]) and GCM ([13]). Our experiments are executed on the pre-processed data. Some basic information of the four datasets are summarized in Table 1.

Before calculating the mutual information, we discretized the training dataset using threshold decided by the mean μ and the standard deviation σ of each gene expression. In the expression of each gene, data largerd than $\mu + 0.5\sigma$ will be changed into 2; data smaller than $\mu - 0.5\sigma$ will be changed into -2; data between $\mu - 0.5\sigma$ and $\mu + 0.5\sigma$ will be changed into 0.

Normally, the number of genes is very larger (several thousands) in gene expression data. In order to speed up the selection procedure, we remove 10% genes of the remainder genes in the first few iterations, and then remove one gene at a time when the number of gene is equal to or less than 200.

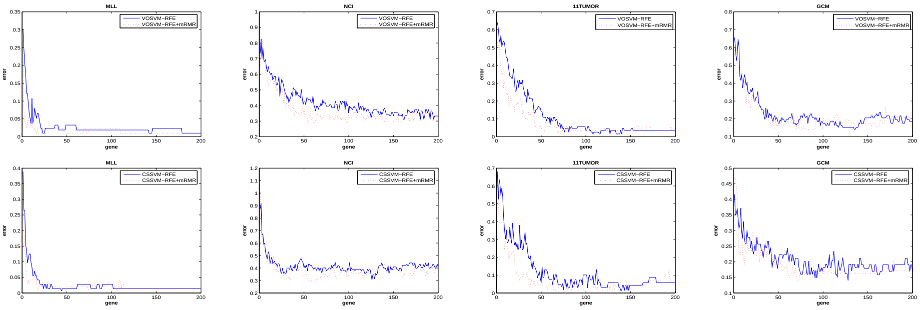

Figure 1: The first row figures are the comparisons of the VOSVM-RFE and K-VOSVM-RFE+mRMR on MLL, NCI, 11TUMOR and GCM; the second row figures are the comparisons of K-CSSVM-RFE and K-CSSVM-RFE+mRMR on MLL, NCI, 11TUMOR and GCM.

Table 2: MLL

Method	Gene#	Error
L-VOSVM-RFE	26	2.56%
K-VOSVM-RFE	24	0.99%
K-VOSVM-RFE+mRMR	**16**	**0.69%**
CSSVM-RFE	41	2.78%
K-CSSVM-RFE	44	0.69%
K-CSSVM-RFE+mRMR	**17**	**0.69%**

Table 3: NCI

Method	Gene#	Error
L-VOSVM-RFE	152	38.3%
K-VOSVM-RFE	161	31.16%
K-VOSVM-RFE+mRMR	**55**	**28.76%**
L-CSSVM-RFE	143	33.14%
K-CSSVM-RFE	126	30.65%
K-CSSVM-RFE+mRMR	**53**	**29.17%**

In our experiments, RBF kernel is choosed $K(x\cdot x') = \exp(-\gamma||x-x'||^2)$. The penalty parameter C is fixed as 100. We employ 4-fold stratified cross validation [14] to choose the optimal parameter γ and evaluate the feature selection performance.

In the Figure 1, the classification error rates on 1-200 genes of the K-MSVM-RFE+ mRMR criterion are plotted, where the K-MSVM is CSSVM or VOSVM. These 200 genes are in the front rank in the algorithm, so they are the more important genes. It can be seen from the Figure 1 that the K-MSVM-RFE+mRMR gives low classification error in most part of gene subset when the number of genes is less than 100. This method selects the genes based on their effect on classification accuracy and makes sure that they are least redundant among themselves. So we may say that our algorithm has superior when small number genes are selected.

In the Table 2 $\sim$ 5, we report the numbers of selected genes and the corresponding error rates by implementing the linear MSVM-RFE (L-MSVM-RFE), the kernel MSVM-RFE (K-MSVM-RFE) and the K-MSVM-RFE+mRMR on the four datasets: MLL, NCI, 11TUMOR and GCM, where the MSVM is CSSVM or VOSVM. And “Gene#" is the numbers of selected genes. “Error" is the error rate of the 4-fold stratified cross validation. As shown by the table, the K-MSVM-RFE gives better performance than L-MSVM-RFE. When combining the kernel MSVM-RFE and mRMR criterion it results in fewer numbers of selected genes and the classification errors decrease sometimes. The results are improved in four gene expression datasets except for 11TUMOR.

Table 4: 11TUMOR

Method	Gene#	Error
L-VOSVM-RFE	179	5.37%
K-VOSVM-RFE	108	1.54%
K-VOSVM-RFE+mRMR	**53**	**1.54%**
linear CSSVM-RFE	161	3.28%
K-CSSVM-RFE	129	**1.45%**
K-CSSVM-RFE+mRMR	**58**	**1.67%**

Table 5: GCM

Method	Gene#	Error
L-VOSVM-RFE	178	20.21%
K-VOSVM-RFE	135	14.1%
K-VOSVM-RFE+mRMR	**69**	**14.1%**
linear CSSVM-RFE	159	17.91%
K-CSSVM-RFE	127	14.66%
K-CSSVM-RFE+mRMR	**71**	**13.90%**

Table 6: The genes selected by the K-MSVM-RFE+mRMR method.

Index	Accession	Description
10797	1914_at	Human cyclin A1 mRNA, complete cds
7135	4052_at	Human rearranged mRNA for glutamine synthase
8105	34840_at	we38g03.x1 Homo sapiens cDNA, 3' end
11366	1325_at	Human Smad1 mRNA, complete cds
11297	1389_at	Human common acute lymphoblastic leukemia antigen (CALLA) mRNA
9005	38017_at	Human MB-1 gene, complete cds
3277	38242_at	Homo sapiens B cell linker protein BLNK mRNA, alternatively spliced, complete cds

Now, taking MLL data as an instance we give some annotation of the selected genes. MLL data consists of three classes of leukemia. As we apply our gene selection method 40 times on different subsets of the data set (4-fold stratified cross-validation repeated 10 times), we actually obtained 40 different gene subsets. After computing the frequency of each gene appearing in all the 40 gene subsets, we can identify the important genes which have been most frequently selected. We give these selected genes by K-MSVM-RFE+mRMR method in Table 6 and plot the expression levels of six informative genes in three different class samples in Figure 2. We can see from the mean expression level of gene on each class sample, these informative genes have great different expression level in different class samples. So they are the useful genes for classification.

Some genes selected by our algorithms have been identified as tumor or tissue specific-genes. Such as human cyclin A1 was cloned as an A-type cyclin that is highly expressed in acute myeloid leukaemia (AML). It suggests that cyclin A1 may have a role in hematopoiesis. High levels of cyclin A1 expression are especially associated with certain leukemias blocked at the myeloblast and promyelocyte stages of differentiation [15, 16]. MB-1 is a sensitive and specific reagent for B-lineage blasts that will aid in the classification of B-cell precursor ALL and in the identification of biphenotypic leukemia presenting as AML [17]. BLNK also consistently identified as one of the most informative genes. It has been proposed that BLNK deficiency is a primary cause of B-lineage acute lymphoblastic leukemia (ALL) [18, 19]. cALLA, initially known as CD10, was identified as one of the earliest markers expressed by leukemic cells of the lymphoblastic lineage [20].

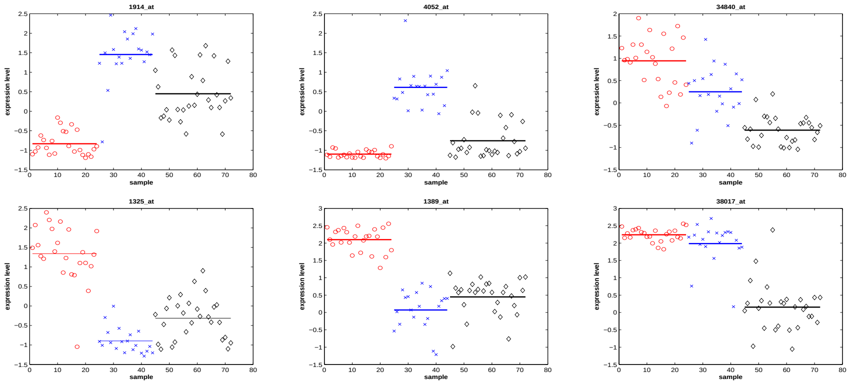

Figure 2: The expression levels of the first six genes in Table 6 on different class samples, –is the mean expression level of one class samples

5 Discussion

In this paper we have explored a new multi-class gene selection method that combines the kernel MSVM-RFE with the mRMR criteria. It can be observed from the numerical result that the introducing of kernel breaks the limitation of linear kernel and makes the gene selection with better performance firstly. After combining with mRMR criteria, the method selects least redundant genes with low classification error rate. The selected gene set can give better representation of whole dataset. It should be pointed out that it is time consuming to find the optimal parameter for classifies, due to the introducing of the kernel function. It is meaningful to construct a fast approach to select the optimal of parameters to reduce the time of gene selection.

References

[1] Guyon,I., Weston,J., Barhill,S. and Vapnik,V. (2002) Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning,* **46**, 389–422.

[2] Kai-Bo,D., Jagath,C., Rajapakse1,2, and Minh N. Nguyen1. (2007) One-Versus-One and One-Versus-All Multiclass SVM-RFE for Gene Selection in Cancer Classification. *Springer-Verlag Berlin Heidelberg,* **4447**, 47–56.

[3] Chai,H. and Domeniconi,C. (2004) An evaluation of gene selection methods for multi-class microarray data classification. *Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics,* 3–10.

[4] Xin,Z. and David,P. Tuck*. (2007) MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data. *Bioinformatics,* **23**, 1106–1114.

[5] Peng,H., Long,F., Ding,C. (2005) Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Patt. Anal. Machi. Intell,* **27**, 1226–1237.

[6] Vapnik,V. (1998) Statistical Learning Theory. *JohnWiley Sons.*

[7] Vapnik,V. (2000) The Nature of Statistical Learning Theory (2nd ed .) *New York:Springer*.

[8] Crammer,K. and Singer,Y. (2001) On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res., 2,* 265–292.

[9] Szedmak,S. and Shawe-Taylor,J. (2005) Multiclass Learning at One-class Complexity. Technical Report. *ISIS Group, Electronics and Computer Science.*

[10] Staunton,J.E., Slonim,D.K., Coller,H.A., Tamayo,P., Angelo, M.J., Park,J., Scherf,U., Lee,J.K., Reinhold,W.O., Weinstein,J.N., Mesirov,J.P., Lander,E.S. and Golub,T.R. (2001) Chemosensitivity prediction by transcriptional profiling. *Proceedings of the National Academy of Sciences,* **98** (19), 10787–10792.

[11] Armstrong,S.A., Staunton,J.E., Silverman,L.B., Pieters,R., den Boer,M.L., Minden,M.D., Sallan,S.E., Lander,E.S., Golub,T.R. and Korsmeyer,S.J. (2002) MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics,* **30** (1), 41–47.

[12] Su,A.I., Welsh,J.B., Sapinoso,L.M., Kern,S.G., Dimitrov,P., Lapp,H., Schultz,P.G., Powell,S.M., Moskaluk,C.A., Frierson, Henry F., J. and Hampton,G.M. (2001) Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Research,* **61** (20), 7388–7393.

[13] Pomeroy,S., Tamayo,P., Gaasenbeek,M., Sturla,L., Angelo,M., McLaughlin,M., Kim,J., Goumnerova,L., Black,P., Lau,C. (2003) Prediction of Central Nervous System Embryonal Tumor Outcome Based on Gene Expression. *Nature,* **415**, 436–442.

[14] Breiman,L. (2003) Classification and Regression Trees. *Wadsworth and Brooks, Monterey, CA.*

[15] Kramer,A., Hochhaus,A., Saussele,S., Reichert,A., Willer,A., Hehlmann,R. (1998) Cyclin A1 is predominantly expressed in hematological malignancies with myeloid differentiation. *Leukemia,* **12** (6), 893–8.

[16] Yang,R., Nakamaki,T., LÍźbbert,M., Said,J., Sakashita,A., Freyaldenhoven BS., Spira,S., Huynh,V., MÍźller,C., Koeffler,HP. (1999) Cyclin A1 expression in leukemia and normal hematopoietic cells. *Blood,* **93** (6), 2067–74.

[17] Mason,DY., Cordell,JL., Brown,MH., Borst,J., Jones,M., Pulford,K., Jaffe,E., Ralfkiaer,E., Dallenbach,F., Stein,H. (1995) CD79a: a novel marker for B-cell neoplasms in routinely processed tissue samples. *Blood,* **86** (4), 1453–9.

[18] Imai,C., Ross,ME., Reid,G., Coustan-Smith,E., Schultz,KR., Pui,CH., Downing,JR., Campana,D. (2004) Expression of the adaptor protein BLNK/SLP-65 in childhood acute lymphoblastic leukemia. *Leukemia,* **18** (5), 922–5.

[19] Jumaa,H., Bossaller,L., Portugal,K., Storch,B., Lotz,M., Flemming,A., Schrappe,M., Postila,V., Riikonen,P., Pelkonen,J., Niemeyer,CM., Reth,M. (2003) Deficiency of the adaptor SLP-65 in pre-B-cell acute lymphoblastic leukaemia. *Nature,* **423** (6938), 452–6.

[20] Bene,MC. and Faure,GC. (1997) CD10 in acute leukemias. GEIL (Groupe d'Etude Immunologique des LeucÍęmies). *Haematologica,* **82** (2), 205–10.