BIOINFORMATICS

Vol. 17 Suppl. 1 2001
Pages S157–S164

Feature selection for DNA methylation based cancer classification

Fabian Model, Péter Adorján, Alexander Olek and Christian Piepenbrock

Epigenomics AG, Kastanienallee 24, D-10435 Berlin, Germany

Received on February 6, 2001; revised and accepted on April 2, 2001

ABSTRACT

Molecular portraits, such as mRNA expression or DNA methylation patterns, have been shown to be strongly correlated with phenotypical parameters. These molecular patterns can be revealed routinely on a genomic scale. However, class prediction based on these patterns is an under-determined problem, due to the extreme high dimensionality of the data compared to the usually small number of available samples. This makes a reduction of the data dimensionality necessary. Here we demonstrate how phenotypic classes can be predicted by combining feature selection and discriminant analysis. By comparing several feature selection methods we show that the right dimension reduction strategy is of crucial importance for the classification performance. The techniques are demonstrated by methylation pattern based discrimination between acute lymphoblastic leukemia and acute myeloid leukemia.

Contact: Fabian.Model@epigenomics.com

INTRODUCTION

In recent years there has been a large interest in the analysis of mRNA expression by using microarrays (Lockhart & Winzeler, 2000). This technology allows to look at thousands of genes, see how they are expressed as proteins and gain insight into cellular processes. An important and scientifically interesting application of this technology is the classification of tissue types, especially the prediction of tumor classes (Golub *et al.*, 1999; Ben-Dor *et al.*, 2001; Weston *et al.*, 2001).

However, there are some practical problems with the large scale analysis of mRNA based microarrays. They are primarily impeded by the instability of mRNA (Emmert-Buck *et al.*, 2000). Also sample preparation is complicated by the fact that expression changes occur within minutes following certain triggers. The inability to resolve the individual contributions of such influences on an expression profile, and difficulties with quantifying the gradual nature of the occurring changes complicates data analysis.

An alternative approach is to look at DNA methylation (Adorján *et al.*, 2001). Methylation is a modification of cytosine, which occurs either with or without a methyl group attached. This methylation of cytosine can only appear together with guanine as CpG. The methylated CpG can be seen as a 5th base and is one of the major factors responsible for expression regulation (Robertson & Wolffe, 2000). Here we demonstrate that cancer classification based solely on DNA methylation analysis is possible and that results comparable to mRNA expression can be achieved.

In order to perform a methylation based prediction we use the well known support vector machine algorithm (Vapnik, 1998; Christianini & Shawe-Taylor, 2000). This algorithm has shown outstanding performance in several areas of application and has already been successfully used to classify mRNA expression data (Ben-Dor *et al.*, 2001; Weston *et al.*, 2001; Brown *et al.*, 2000; Gaasterland & Bekiranov, 2000). The major problem of all classification algorithms for methylation and expression data analysis alike is the high dimension of input space compared to the small number of available samples. Although the support vector machine is designed to overcome this problem it still suffers from these extreme conditions. Therefore feature selection is of crucial importance for good performance (Blum & Langley, 1997; Weston *et al.*, 2001; Ben-Dor *et al.*, 2001) and we give special consideration to it by comparing several methods on our methylation data.

The data set (Adorján *et al.*, 2001) consists of cell lines and primary tissue obtained from patients with acute lymphoblastic leukemia (ALL) or acute myeloid leukemia (AML). A total of 17 ALL and 8 AML samples were included. The methylation status of these samples was evaluated at 81 CpG dinucleotide positions located in CpG rich regions of the promoters, intronic and coding sequences of 11 genes. These were randomly selected from a panel of genes representing different pathways associated with tumor genesis. Two of the 11 selected genes are located on the X-chromosome.

The rest of the paper is organized as follows. In Section 2, we give a short description of the process

used for generating the methylation data. Especially we demonstrate how the process can be validated and calibrated. In Section 3, we give a short introduction to the support vector machine and describe our experimental setting. In Section 4, we address the problem of feature selection by introducing and comparing several methods. Finally we conclude in Section 5 with a discussion of the potential impact of methylation analysis and future directions.

MICROARRAY-BASED METHYLATION ANALYSIS

In order to allow sequence specific distinction of methylated from unmethylated states of CpG dinucleotides by hybridization analyses, total DNA from all samples was bisulphite treated converting all unmethylated cytosines to uracil whereas methylated cytosines were conserved (Frommer *et al.*, 1992). Regions of interest were then amplified by PCR using fluorescently labeled primers converting originally unmethylated CpG dinucleotides to TG and conserving originally methylated CpG sites. PCR primers were designed complementary to DNA segments containing no CpG dinucleotides. This allowed unbiased amplification of both methylated and unmethylated alleles in one reaction. All PCR products performed on an individual sample were mixed and hybridized to glass slides carrying for each CpG position a pair of immobilized oligonucleotides. Each of these detection oligonucleotides was designed to hybridize to the bisulphite converted sequence around one CpG site which was either originally unmethylated (TG) or methylated (CG). Hybridization conditions were selected to allow the detection of the single nucleotide differences between the TG and CG variants. Ratios for the two signals were calculated based on comparison of intensity of the fluorescent signals.

The sensitivity of the method for detection of methylation changes was determined using artificially up- and down methylated DNA fragments mixed at different ratios. For each of those mixtures, a series of experiments was conducted to define the range of CG/TG ratios that corresponds to varying degrees of methylation at each of the CpG sites tested. In Fig. 1a results for two CpG positions located in exon 14 of the human factor VIII gene are shown as examples. For the mixtures of 3:0, 2:1, 1:2 and 0:3 the degree of methylation of the individual CpG sites could safely be distinguished.

To verify the detection of methylation changes in the real data set two X-chromosomal genes were included in the gene set. Because one of the two X-chromosomes in females becomes inactivated by methylation we can expect a higher degree of methylation of X-chromosomal genes in females compared to males. In Fig. 1b CpGs are ranked according to the significance of the difference between male and female methylation levels. As expected, the X-chromosomal genes (ELK1, AR) show a significantly higher methylation for females. This clearly demonstrates that the method really detects changes in methylation.

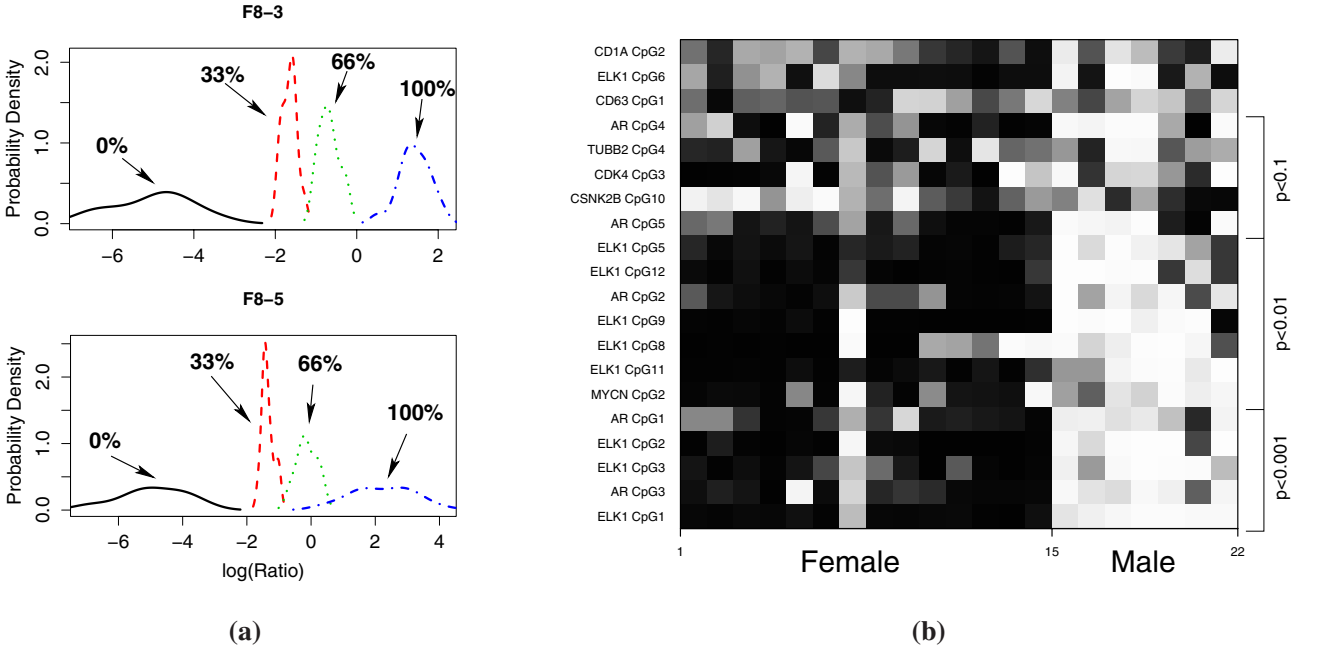
SUPPORT VECTOR MACHINES

In our case, the task of cancer classification consists of constructing a machine that can predict the leukemia subtype (ALL or AML) from a patients methylation pattern. For every patient sample this pattern is given as a vector of average[†] log $\frac{CG}{TG}$ ratios at 81 CpG positions. Based on a given set of training examples $X = \{\mathbf{x}^i : \mathbf{x}^i \in R^n\}$ with known diagnosis $Y = \{y^i : y^i \in \{ALL, AML\}\}$ a discriminant function $f : R^n \rightarrow \{ALL, AML\}$, where n is the number of CpGs, has to be learned. The number of misclassifications of f on the training set $\{X, Y\}$ is called training error and is usually minimised by the learning machine during the training phase. However, what is of practical interest is the capability to predict the class of previously unseen samples, the so called generalisation performance of the learning machine. This performance is usually estimated by the test error, which is the number of misclassifications on an independent test set $\{X', Y'\}$.

The major problem of training a learning machine with good generalisation performance is to find a discriminant function f which on the one hand is complex enough to capture the essential properties of the data distribution, but which on the other hand avoids over-fitting the data. The Support Vector Machine (SVM) tries to solve this problem by constructing a linear discriminant that separates the training data and maximises the distance to the nearest points of the training set. This maximum margin separating hyperplane minimises the ratio between the radius of the minimum enclosing sphere of the training set and the margin between hyperplane and training points. This corresponds to minimising the so called radius margin bound on the expected probability of a test error and promises good generalisation performance (Vapnik, 1998).

Of course there are more complex classification problems, where the dependence between class labels y^i and features $\mathbf{x}^i$ is not linear and the training set can not be separated by a hyperplane. In order to allow for non-linear discriminant functions the input space can be non-linearly mapped into a potentially higher dimensional feature space by a mapping function $\Phi : \mathbf{x}^i \mapsto \Phi(\mathbf{x}^i)$. Because the SVM algorithm in its dual formulation uses only the inner product between elements of the input space, the knowledge of the kernel function $k(\mathbf{x}^i, \mathbf{x}^j) = \langle \Phi(\mathbf{x}^i) \cdot \Phi(\mathbf{x}^j) \rangle$ is sufficient to train the SVM.

[†] Every hybridization experiment was at least 3 times repeated and the results averaged.



(a) **(b)**

Fig. 1. Validation of measurements. **a)** Quantification of methylation measurements for two CpG dinucleotides. A series of hybridizations was performed with mixtures of artificially up- and down-methylated DNA fragments of the factor VIII exon 14 gene. Down- and up-methylated DNA fragments were mixed at ratios: 0:3, 1:2, 2:1, 3:0, representing a methylation status of 100 %, 66 %, 33 % and 0 %, respectively. For the 4 kinds of compounds 59, 36, 40, 63 identical slides were made. The log-ratio of the CG and the TG detection oligomer hybridization intensity was calculated and then averaged for experimental subgroups each containing 3 identical experiments. The distribution function of the CG:TG ratios shows that measurement values of the different mixtures are well separated and therefore allow a high resolution detection of the methylation level of a single CpG. **b)** Gender separation. The 20 CpG sites with the most significant difference between female and male samples are shown. Only non cell line leukemia and healthy control samples were used. As expected the absolute majority of the significant CpG dinucleotides come from the two X-chromosome genes (ELK1, AR). High probability of methylation corresponds to black, uncertainty to grey and low probability to white. The labels on the left side of the plot are gene and CpG identifiers. The bottom to top ranking of the CpGs is according to the significance of the difference between the means of the two groups, estimated by a two sample t-test. Each row corresponds to a single CpG and each column to the methylation levels of one sample.

It is not necessary to explicitly know the mapping Φ and a non-linear SVM can be trained efficiently by computing only the kernel function. Here we will only use the linear kernel $k(\mathbf{x}^i, \mathbf{x}^j) = \langle \mathbf{x}^i \cdot \mathbf{x}^j \rangle$ and the quadratic kernel $k(\mathbf{x}^i, \mathbf{x}^j) = \left(\langle \mathbf{x}^i \cdot \mathbf{x}^j \rangle + 1\right)^2$.

In the next section we will compare SVMs trained on different feature sets. In order to evaluate the prediction performance of these SVMs we used a cross-validation method (Bishop, 1995). For each classification task, the samples were partitioned into 8 groups of approximately equal size. Then the SVM predicted the class for the test samples in one group after it had been trained using the 7 other groups. The number of misclassifications was counted over 8 runs of the SVM algorithm for all possible choices of the test group. To obtain a reliable estimate for the test error the number of misclassifications were averaged over 50 different partitionings of the samples into 8 groups.

FEATURE SELECTION

The simplest way for applying a SVM to our methylation data is to use every CpG position as a separate dimension, not making any assumption about the interdependence of CpG sites from the same gene. On the leukemia subclassification task the SVM with linear kernel trained on this 81 dimensional input space had an average test error of 16%. Using a quadratic kernel did not significantly improve the results (see Tab. 1). An obvious explanation for this relatively poor performance is that we have only 25 data points (even less in the training set) in a 81 dimensional space. Finding a separating hyperplane under these conditions is a heavily under-determined problem. And as it turns out, the SVM technique of maximising the margin is not sufficient to find the solution with optimal generalisation properties. It is necessary to reduce the dimensionality of the input space while retaining the relevant information for classification. This should be

Table 1. Performance of different feature selection methods. † The SVM was trained on all 81 features.

	Training Error 2 Features	Test Error 2 Features	Training Error 5 Features	Test Error 5 Features
Linear Kernel				
Fisher Criterion	0.01	0.05	0.00	0.03
Golub's Method	0.01	0.05	0.00	0.04
t-Test	0.05	0.13	0.00	0.08
Backward Elimination	0.02	0.17	0.00	0.05
PCA	0.13	0.21	0.05	0.28
No Feature Selection†	0.00	0.16		
Quadratic Kernel				
Fisher Criterion	0.00	0.06	0.00	0.03
Golub's Method	0.00	0.06	0.00	0.05
t-Test	0.04	0.14	0.00	0.07
Backward Elimination	0.00	0.12	0.00	0.05
PCA	0.10	0.30	0.00	0.31
Exhaustive Search	0.00	0.06	-	-
No Feature Selection†	0.00	0.15		

possible because it can be expected that only a minority of CpG positions has any connection with the two subtypes of leukemia.

Principle Component Analysis

The probably most popular method for dimension reduction is principle component analysis (PCA) (Bishop, 1995). For a given training set X, PCA constructs a set of orthogonal vectors (principle components) which correspond to the directions of maximum variance. The projection of X onto the first k principle components gives the 2-norm optimal representation of X in a k-dimensional orthogonal subspace. Because this projection does not explicitly use the class information Y, PCA is an unsupervised learning technique.

In order to reduce the dimension of the input space for the SVM we performed a PCA on the combined training and test set $\{X, X'\}$ and projected both sets on the first k principle components. This gives considerably better results than performing PCA only on the training set X and is justified by the fact that no label information is used. However, the generalisation results for $k = 2$ and $k = 5$, as shown in Tab. 1, were even worse than for the SVM without feature selection. The reason for this is that PCA does not necessarily extract features that are important for the discrimination between ALL and AML. It first picks the features with the highest variance, which are in this case discriminating between cell lines and primary patient tissue (see Fig. 2a), i.e. subgroups that are not relevant to the classification task. As is shown in Fig. 3, features carrying information about the leukemia subclasses appear only from the 9th principle component on. The generalisation performance including the 9th component is significantly better than for a SVM without feature selection. However, it seems clear that a supervised feature selection method, which takes the class labels of the training set into account, should be more reliable and give better generalisation.

Fisher Criterion and t-Test

A classical measure to asses the degree of separation between two classes is given by the Fisher criterion (Bishop, 1995). In our case it gives the discriminative power of the kth CpG as

$$J(k) = \frac{(\mu_k^{ALL} - \mu_k^{AML})^2}{\sigma_k^{ALL\,2} + \sigma_k^{AML\,2}},$$

where $\mu_k^{ALL/AML}$ is the mean and $\sigma_k^{ALL/AML}$ is the standard deviation of all x_k^i with $y_i = ALL/AML$. The Fisher criterion gives a high ranking for CpGs where the two classes are far apart compared to the within class variances. Fig. 2b shows the methylation profiles of the best 20 CpGs according to the Fisher criterion. The very similar criterion

$$G(k) = \frac{|\mu_k^{ALL} - \mu_k^{AML}|}{\sigma_k^{ALL} + \sigma_k^{AML}}$$

was used by Golub and coworkers for their ALL/AML classification based on mRNA expression data (Golub

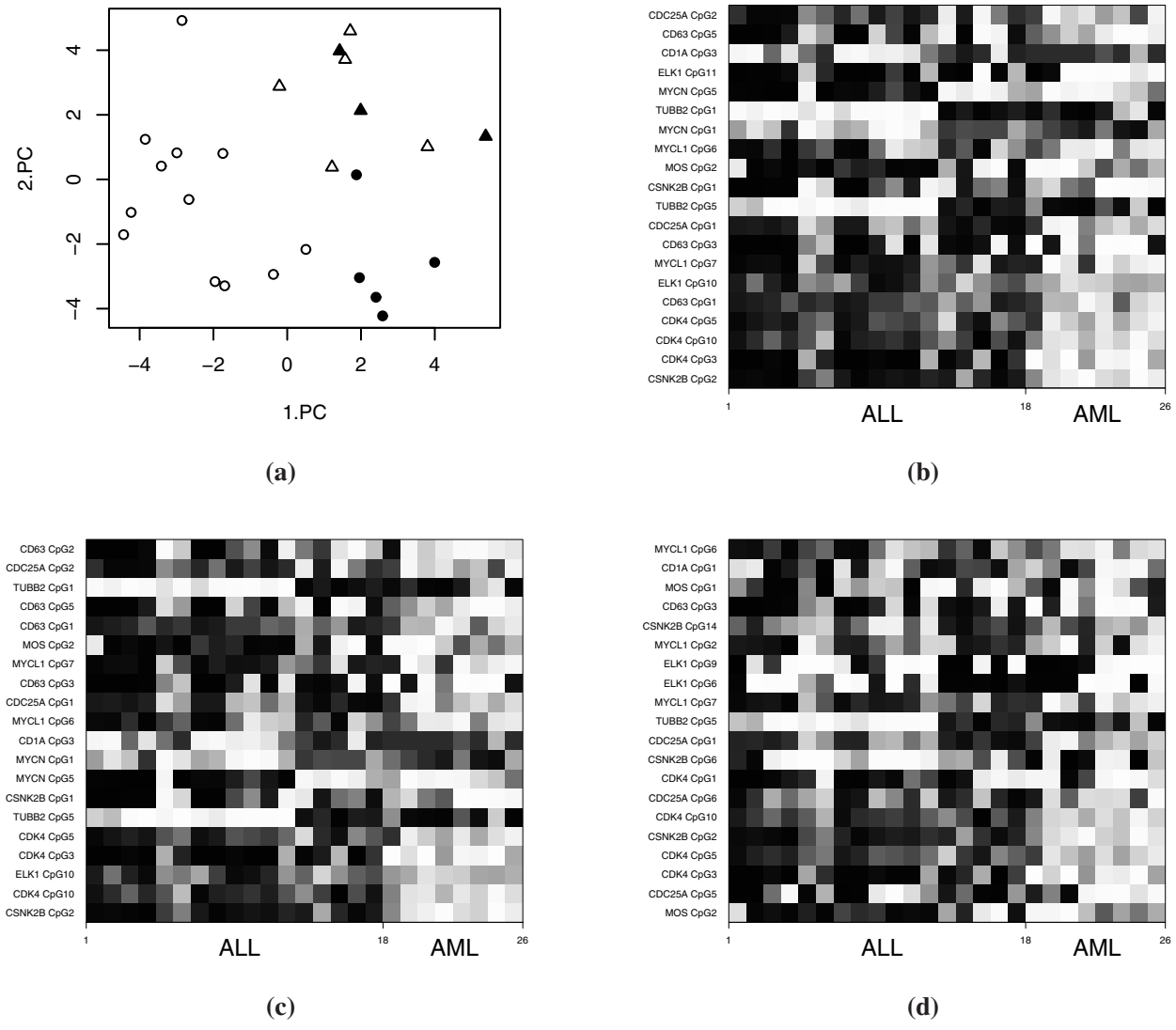


Fig. 2. Feature selection methods. **a)** Principle component analysis. The whole data set was projected onto its first 2 principle components. Circles represent cell lines, triangles primary patient tissue. Filled circles or triangles are AML, empty ones ALL samples. **b)** Fisher criterion. The 20 highest ranking CpG sites according to the Fisher criterion are shown. The highest ranking features are on the bottom of the plot. High probability of methylation corresponds to black, uncertainty to grey and low probability to white. **c)** Two sample t-test. **d)** Backward elimination.

et al., 1999). Its relation to the Fisher criterion is given by

$$G^2(k) = J(k)\left(1 + \frac{2\sigma_k^{ALL}\sigma_k^{AML}}{\sigma_k^{ALL\,2} + \sigma_k^{AML\,2}}\right)^{-1},$$

which shows the preference of Golub's ranking for features with different within class variances compared to the Fisher criterion.

Another approach to rank CpGs by their discriminative power is to use a test statistic for computing the significance of class differences. Here we assumed a normal distribution of the methylation levels of a CpG position within a class and used a two sample t-test to rank the CpGs according to the significance of the difference between the class means (Mendenhall & Sincich, 1995). Fig. 2c shows the ranking, which is very similar to the

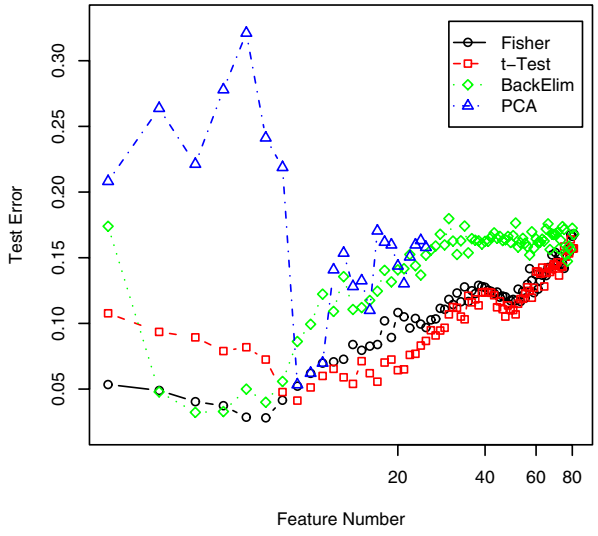


Fig. 3. Dimension dependence of feature selection performance. The plot shows the generalisation performance of a linear SVM with four different feature selection methods against the number of selected features. The x-axis is scaled logarithmically and gives the number of input features for the SVM, starting with two. The y-axis gives the achieved generalization performance. Note that the maximum number of principle components corresponds to the number of available samples. The performance of Golub's method was very similar to the Fisher criterion and is not shown.

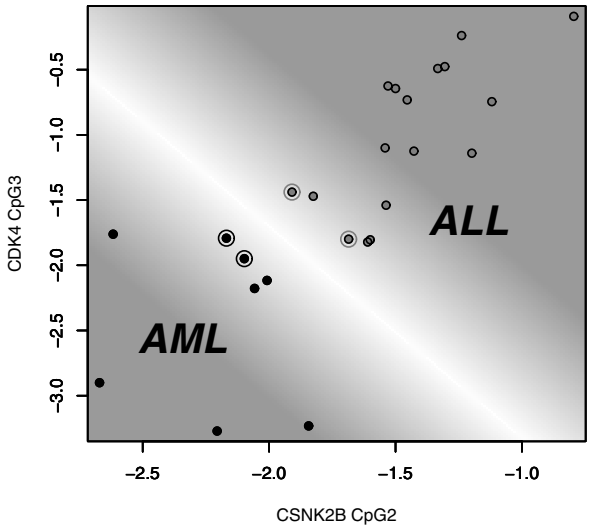


Fig. 4. Support Vector Machine on two best features of the Fisher criterion. The plot shows a SVM trained on the two highest ranking CpG sites according to the Fisher criterion with all ALL and AML samples used as training data. The black points are AML, the grey ones ALL samples. Circled points are the support vectors defining the white borderline between the areas of AML and ALL prediction. The grey value of the background corresponds to the prediction strength.

Fisher criterion because a large mean difference and a small within class variance are the important factors for both methods.

In order to improve classification performance we trained SVMs on the k highest ranking CpGs according to the Fisher criterion, Golub's method or t-test. Fig. 4 shows a trained SVM on the best two CpGs from the Fisher criterion. The test errors for $k = 2$ and $k = 5$ are given in Tab. 1. The results show a dramatic improvement of generalisation performance. Using the Fisher criterion for feature selection and $k = 5$ CpGs the test error was decreased to 3% compared to 16% for the SVM without feature selection. Fig. 3 shows the dependence of generalisation performance from the selected dimension k and indicates that especially the Fisher criterion gives dimension independent good generalisation for reasonable small k. The performance of Golub's ranking method was equal or slightly inferior to the Fisher criterion on our data set, whereas the t-test performance was considerably worse for small feature numbers.

Although the described CpG ranking methods give very good generalisation, they have some potential drawbacks. One problem is that they can only detect linear dependencies between features and class labels. A simple XOR or even OR combination of two CpGs would be completely missed. Another drawback is that redundant features are not removed. In our case there are usually several CpGs from the same gene which have a high likelihood of comethylation. This can result in a large set of high ranking features which carry essentially the same information. Although the good results seem to indicate that the described problems do not appear in our data set, they should be considered.

Backward Elimination

PCA, Fisher criterion and t-test construct or rank features independent of the learning machine that does the actual classification and are therefore called filter methods (Blum & Langley, 1997). Another approach is to use the learning machine itself for feature selection. These techniques are called wrapper methods and try to identify the features that are important for the generalisation capability of the machine. Here we propose to use the features that are important for achieving a low training error as a simple approximation. In the case of a SVM with linear kernel

these features are easily identified by looking at the normal vector **w** of the separating hyperplane. The smaller the angle between a feature basis vector and the normal vector the more important is the feature for the separation. Features orthogonal to the normal vector have obviously no influence on the discrimination at all. This means the feature ranking is simply given by the components of the normal vector as w_k^2. Of course this ranking is not very realistic because the SVM solution on the full feature set is far from optimal as we demonstrated in the last subsections. A simple heuristic is to assume that the feature with the smallest w_k^2 is really unimportant for the solution and can be safely removed from the feature set. Then the SVM can be retrained on the reduced feature set and the procedure is repeated until the feature set is empty. Such a successive feature removal is called backward elimination (Blum & Langley, 1997). The resulting CpG ranking on our data set is shown in Fig. 2d and differs considerably from the Fisher and t-test rankings. It seems backward elimination is able to remove redundant features. However, as shown in Tab. 1 and Fig. 3 the generalisation results are not better than for the Fisher criterion. Furthermore, backward elimination seems to be more dimension dependent and it is computationally more expensive. It follows that at least for this data set the simple Fisher criterion is the preferable feature selection technique.

Exhaustive Search

A canonical way to construct a wrapper method for feature selection is to evaluate the generalisation performance of the learning machine on every possible feature subset. Cross-validation on the training set can be used to estimate the generalisation of the machine on a given feature set. What makes this exhaustive search of the feature space practically useless is the enormous number of $\sum_{k=0}^{n} \binom{n}{k} = 2^n$ different feature combinations and there are numerous heuristics to search the feature space more efficiently (e.g. backward elimination) (Blum & Langley, 1997).

Here we only want to demonstrate that there are no higher order correlations between features and class labels in our data set. In order to do this we exhaustively searched the space of all two feature combinations. For every of the $\binom{81}{2} = 3240$ two CpG combinations we computed the leave-one-out cross-validation error of a SVM with quadratic kernel on the training set. From all CpG pairs with minimum leave-one-out error we selected the one with the smallest radius margin ratio. This pair was considered to be the optimal feature combination and was used to evaluate the generalisation performance of the SVM on the test set.

The average test error of the exhaustive search method was with 6% the same as the one of the Fisher criterion in the case of two features and a quadratic kernel. For five features the exhaustive computation is already infeasible. In the absolute majority of cross-validation runs the CpGs selected by exhaustive search and Fisher criterion were identical. In some cases suboptimal CpGs were chosen by the exhaustive search method. These results clearly demonstrate that there are no second order combinations of two features in our data set that are important for an ALL/AML discrimination. We expect that higher than second order combinations of more than two features can not be detected reliably with such a limited sample size. Therefore the Fisher criterion should be able to extract all classification relevant information from our data set.

CONCLUSIONS

To achieve reliable predictions on the basis of small training set sizes the selection of relevant features is necessary, even for advanced learning algorithms as the support vector machine. For classification tasks where the class information is directly correlated to single CpG dinucleotide markers the simple Fisher criterion is a powerful and efficient feature selection strategy. For more complex problems it will be necessary to derive feature selection algorithms that can remove or combine redundant features and handle higher order feature dependencies.

Taken together, our results clearly demonstrate that microarray based methylation analysis combined with supervised learning techniques can reliably predict known tumor classes. Classification results were comparable to mRNA expression data and our results suggest, that methylation analysis should be applied to other kinds of tissue. Well documented tissue samples with patient history can be obtained only as archived specimens. This strongly limits the amount and number of tissues available for expression analysis (Bowtell, 1999). The methylation approach has the potential to overcome this fundamental limitation: through the mere fact that the stable DNA is the object of study, extraction of material is possible form archived samples. This enables the examination of methylation patterns in large numbers of archived specimen with comprehensive clinical records and removes one of the major limitations for the discovery of complex biological processes by statistical means.

REFERENCES

Adorján, P., Distler, J., Lipscher, E., Model, F., Müller, J., Pelet, C., Braun, A., Florl, A., Gütig, D., Grabs, G., Howe, A., Kursar, M., Lesche, R., Leu, E., Lewin, A., Maier, S., Müller, V., Otto, T., Scholz, C., Schulz, W., Seifert, H., Schwope, I., Ziebarth, H., Berlin, K., Piepenbrock, C. & Olek, A. (2001). Tumour class prediction and discovery by microarray-based dna methylation analysis. Submitted.

Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. & Yakhini, Z. (2001). Tissue classification with gene expression

profiles. In *Proceedings of the Fifth Annual Conference on Computational Molecular Biology*. In press.

Bishop, C. (1995). Neural networks for pattern recognition. Oxford University Press, New York.

Blum, A. & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, **97**, 245–271.

Bowtell, D. (1999). Options available - from start to finish - for obtaining expression data by microarray. *Nature Genetics suppl.*, **21**, 25–32.

Brown, M., Grundy, W., Lin, D., Cristianini, N., Sugnet, C., Furey, T., Ares, M. & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA*, **97**, 262–267.

Christianini, N. & Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines. Cambridge University Press, Cambridge.

Emmert-Buck, M., Strausberg, R., Krizman, D., Bonaldo, M., Bonner, R., Bostwick, D., Brown, M., Buetow, K., Chuaqui, R., Cole, K., Duray, P., Englert, C., Gillespie, J., Greenhut, S., Grouse, L., Hillier, L., Katz, K., Klausner, R., Kuznetzov, V., Lash, A., Lennon, G., Linehan, W., Liotta, L., Marra, M., Munson, P., Ornstein, D., Prabhu, V., Prange, C., Schuler, G., Soares, M., Tolstoshev, C., Vocke, C. & Waterston, R. (2000). Molecular profiling of clinical tissue specimens: feasibility and applications. *Am. J. Pathol.*, **156**, 1109–1115.

Frommer, M., McDonald, L., Millar, D., Collis, C., Watt, F., Grigg, G., Molloy, P. & Paul, C. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proc. Natl. Acad. Sci. USA*, **89**, 1827–1831.

Gaasterland, T. & Bekiranov, S. (2000). Making the most of microarry data. *Nature Genetics*, **24**, 204–206.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. & Lander, E. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Lockhart, D. & Winzeler, E. (2000). Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.

Mendenhall, W. & Sincich, T. (1995). Statistics for engineering and the sciences. Prentice-Hall, New Jersey.

Robertson, K. & Wolffe, A. (2000). DNA methylation in health and disease. *Nature Reviews Genetics*, **1**, 11–19.

Vapnik, V. (1998). Statistical Learning Theory. Wiley, New York.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T. & Vapnik, V. (2001). Feature selection for SVMs. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press, Cambridge, MA. In press.