SVM applications: Microarrays

Nathalie Pochet, Frank De Smet, Johan Suykens, Bart De Moor
{nathalie.pochet, frank.desmet, johan.suykens, bart.demoor}@esat.kuleuven.ac.be

10th March 2003

Technological advances in molecular biology have led to the development of microarrays. These are capable of determining the expression levels of thousands of genes simultaneously. Oncology is one application area of this technology. Because the disordered expression of genes lies at the origin of the behavior of tumors, the measurement of it can be very helpful to predict or to model the clinical behavior of malignant processes. By these means the fundamental processes underlying carcinogenesis are involved in the clinical decision process. Figure 1 shows an image of some malignant lymphoblasts in the periferal blood of a patient suffering from acute lymphoblastic leukemia (ALL). These cells are typically selected for microarrays.

The development of the microarray has led to the generation and analysis of huge amounts of data. The extraction of clinically and biologically relevant information from these data requires specific procedures. This creates an interesting point of view for a relatively new science called bioinformatics. Typically, microarray data are represented by an expression matrix from which the rows and the columns respectively represent the gene expression profiles and the expression patterns of a patient. Figure 2 shows such an expression matrix. In the most recent years, several data sets have been made publicly available on the Internet.

Data sets generated by microarrays consist of a large number of gene expression levels for each patient and a relatively small number of patients (different classes of tumors). The large number of gene expression levels for each patient seems to be a problem for most methods. Consequently, often dimensionality reduction is applied to the data before they can be used. SVMs on the other hand seem to be capable of learning and generalizing these microarray data well despite of the **high dimensionality**.

In the future the amount of microarray data only will increase, which will cause problems for most of the methods. Most methods rely on linear functions to describe the relations in the data. Those methods are unable to discover the **non-linear relationships** in the microarray data. Using more complex kernel functions aims at a better understanding of these more complex data. Several researchers already have been experimenting with kernel

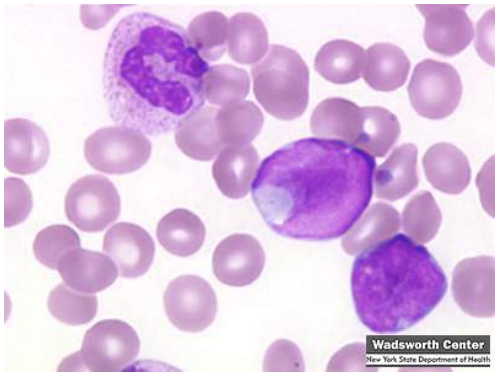

Figure 1: Some malignant lymphoblasts in the periferal blood of a patient suffering from acute lymphoblastic leukemia (ALL). These cells are typically selected for microarrays.

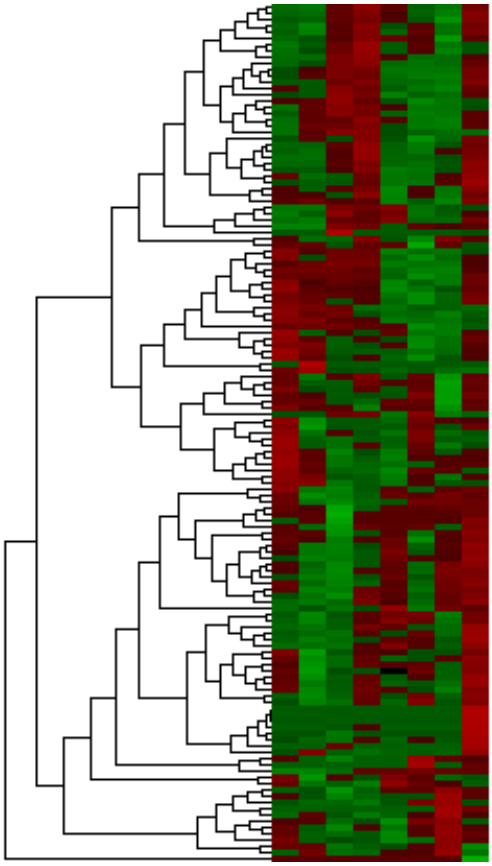

Figure 2: Microarray data are represented by an expression matrix from which the rows and the columns respectively represent the gene expression profiles and the expression patterns of a patient. (Moreau et al., 2002)

functions on microarray data. Choosing an optimal kernel function and tuning the kernel parameters seems to be not so obvious. Non-optimal use of kernel functions easily results in overfitting. In order to estimate the optimal kernel function and the kernel parameters, methods like cross-validation and bootstrapping can be used.

Microarray data are characterized by high dimensionality and complexity. A promising method to handle these data are SVMs, which are capable of using non-linear techniques expressed by kernel functions. Although SVMs have means to prevent overfitting, it is important to use them optimally. A first way is to tune the kernel function either by cross-validation or by taking the VC-dimension in relation to the upper bound on the generalization error. A second way exists of selecting the most relevant genes to act as input for the prediction model.

In the literature, three kinds of applications, arising from applying SVMs to microarray data, have already been pointed out. The first one is the performance of clinical predictions, the second one the performance of biological predictions, and the third one the discovery of relevant genes or groups of genes. These applications will be discussed now.

Case study 1: The performance of clinical predictions

The main goal here is to make predictions about the clinical information (e.g. diagnosis, prognosis, therapy response,...) of individual patients based on microarray measurements (possibly supplemented with other clinical data).

This can be realized by means of models that attempt to classify tissue samples (e.g. type of cancer) of patients based on selected features. The parameters of the model need to be determined using a collection of patients of who is known yet to which class they belong. Those are the patients for whom for example the stage determination, histopathological diagnosis, prognosis, therapy response,... are known yet. This collection of patients is called the training set and it is used to train the model. In practice, this amounts to determining the parameters of the model. The trained model can now be applied to make predictions about the patients whose classification is not known yet. Those are the patients of the test set.

When classifying patients, the data set is characterized by only **few samples** (< 100) having **large dimensions** ($5000 - 100000$).

Because of the small number of samples (patients), leave-one-out cross-validation is the preferred cross-validation method for estimating the generalization performance and for model selection. The complete SVM method executed by (Furey et al., 2000) can be described as follows. They begin by choosing a kernel, starting with the simple linear kernel, and tune the regularization parameter to achieve the best performance on leave-one-out cross-validation tests using the full dataset. The SVM tuning procedure is then

repeated with a specified number of the top-ranked features. In these cases, for each individual leave-one-out test, the features are ranked based on the method introduced by (Golub et al., 1999), using the scores from only the known samples, some number of the top features are extracted, and then these are used to train the SVM and classify the unknown sample. Examples that have been consistently misclassified in all tests are identified. These examples can then be investigated by a biologist, and if it is determined that the original label is incorrect, a correction is made, and the process is repeated. Alternatively, an example may be deemed an outlier that is very different from the rest, and is therefore removed.

It should be noted that it is very important that when feature selection is performed, the sample being tested must not be included in this process. Each individual leave-one-out test requires a new ranking of features using only those samples that are to be used for training. Inclusion of the test sample when doing feature selection can cause a leak of information, which invalidates the independence assumptions required to reasonably evaluate the methods performance. For this problem in particular, a lot of information can be leaked in this way.

Classification with SVMs points out that a linear kernel gives the best results. Using SVMs results in an almost perfect classification, but other algorithms perform well too (e.g. linear perceptron). The data sets currently available contain relatively few examples and thus do not allow one method to demonstrate superiority. SVMs perform well using a simple kernel, and (Furey et al., 2000) believe that as datasets containing more examples become available, the use of more complex kernels may become necessary and will allow SVMs to continue their good performance.

Making clinical predictions (diagnosis, prognosis, therapy response,...) for individual patients from microarray data seems to be possible, but larger-scale systematic experiments must be conducted. SVMs are expected to have good performances when data increase, but currently perform at the same level as other methods.

References

- S. Mukherjee, P. Tamayo, J.P. Mesirov, D. Slonim, A. Verri, T. Poggio. **Support vector machine classification of microarray data.** A.I. Memo 1677, MIT Artificial Intelligence Laboratory, 1998.

- T.S. Furey, N. Duffy, N. Cristianini, D. Bednarski, M. Schummer, D. Haussler. **Support vector machine classification and validation of cancer tissue samples using microarray expression data.** *Bioinformatics*, 16(10):906-914, 2000.

Case study 2: The performance of biological predictions

In this case, one aims at making predictions about the function of genes and the contribution of genes to the carcinogenesis.

SVMs, like all other supervised learning techniques, use a training set to specify in advance which data should cluster together. As applied to gene expression data, a SVM would begin with a set of genes that have a common function: for example in (Brown et al., 2000), genes coding for ribosomal proteins or genes coding for components of the proteasome. In addition, a separate set of genes that are known not to be members of the functional class is specified. These two sets of genes are combined to form a set of training examples in which the genes are labeled positively if they are in the functional class and are labeled negatively if they are known not to be in the functional class. Using this training set, a SVM would learn to discriminate between the members and non-members of a given functional class based on expression data. Having learned the expression features of the class, the SVM could recognize new genes as members or as non-members of the class based on their expression data. The SVM could also be reapplied to the training examples to identify outliers that may have previously been assigned to the incorrect class in the training set. Thus, a SVM would use the biological information in the investigators training set to determine what expression features are characteristic of a given functional group and use this information to decide whether any given gene is likely to be a member of the group.

When classifying gene expressions, one has to cope with an **unbalanced problem**. Each class contains few genes compared to the total number of genes (many negative examples).

SVMs with different kernel functions, namely linear, polynomial and Gaussian kernels, are tried out. Also a comparison with other methods like Parzen windows, Fishers linear discriminant and decision trees, is conducted. The results show that for all classes SVM with Gaussian kernels performs best. It is possible to improve classification by using only a carefully selected part of the genes. Note that (Brown et al., 2000) used three-fold cross-validation to measure the performance, this because of the large number of samples (genes).

SVMs with more complex kernel functions seem to perform better than other classical learning algorithms when classifying gene expressions.

References

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

Case study 3: The discovery of relevant genes or groups of genes

The objective here is to find genes potentially responsible for the classification of the tissue samples.

This can be realized by searching for relations between gene expressions and class labels. Not all gene expressions are correlated to the different diagnostic or biological classes. The intention here is to find gene expressions or groups of gene expressions that are correlated to the different diagnostic classes. This will allow to gain insight in the molecular biology underlying carcinogenesis. The selected genes could open up new perspectives for finding drug targets and for finding tumor markers.

When searching for relevant genes, one has to cope again with a data set that is characterized by only **few samples** (< 100) and **large dimensions** ($5000 - 100000$).

By using the score introduced by (Golub et al., 1999) for ranking and selecting the genes, there seems to be only few biological relevance. Another gene selection procedure is proposed by (Guyon et al., 2002): the SVM method of Recursive Feature Elimination (RFE). Genes are ranked based on their weight learned by a SVM. Genes are removed one by one (or by chunks), and a SVM is re-run at each iteration.

Experiments on two different cancer databases showed that taking into account mutual information between genes in the gene selection process impacts classification performance. The RFE method obtains significant improvements over the baseline method that makes implicit orthogonality assumptions. (Guyon et al., 2002) also verified the biological relevance of the genes found by SVMs. The top ranked genes found by SVMs all have a plausible relation to cancer. In contrast, other methods select genes that are correlated with the separation at hand but not relevant to cancer diagnosis. See figures 3 and 4.

Feature ranking methods do not dictate the optimum number of features to be selected. An auxiliary model selection criterion must be used for that purpose. The problem is particularly challenging because the leave-one-out error by itself is of little use since it is zero for a large number of gene subsets. Possible criteria that they have explored include the number of support vectors and a combination of the four metrics of classifier quality (error rate, rejection rate, extremal margin, and median margin) computed with the leave-one-out procedure. See figure 5. They have also explored with adding penalties for large numbers of features, using bounds on the expected error rate. Finding a good model selection criterion is an important avenue of experimental and theoretical research.

Again, it should be noted that the proper way to conduct leave-one-out cross-validation for feature selection is to avoid using a fixed set of features selected with the whole training data set, because this induces a bias in the results. Instead, one should withhold a pattern, select features, and assess the performance of the classifier with the selected features using the left out example. One then rotates over all the examples, recomputing the feature set

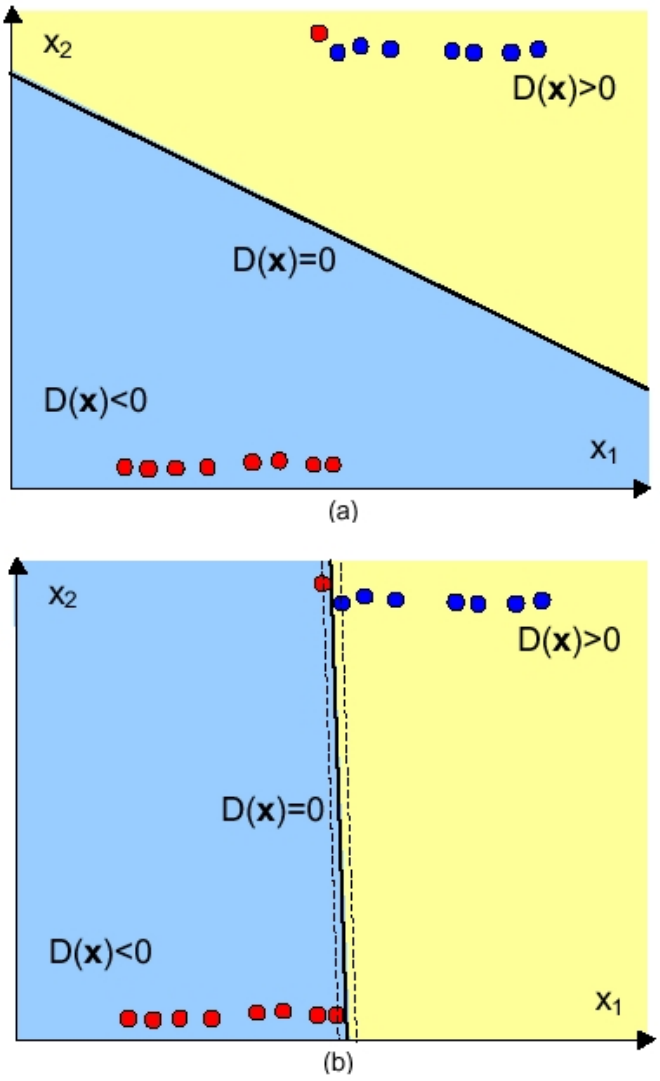

Figure 3: Feature selection and support vectors. This figure contrasts on a two dimensional classification example the feature selection strategy of "average case" type methods and that of SVMs. The red and blue dots represent examples of class (-) and (+) respectively. The decision boundary $D(x) = 0$ separates the plane into two half planes: if $D(x) < 0$ then x in class (-), and if $D(x) > 0$ then x in class (+). There is a simple geometric interpretation of the feature ranking criterion based on the magnitude of the weights: for slopes larger than 45 degrees, the preferred feature is x1, otherwise it is x2. The example was constructed to demonstrate the qualitative difference of the methods. Feature x2 separates almost perfectly all examples with a small variance, with the exception of one outlier. Feature x1 separates perfectly all examples but has a higher variance. (a) Baseline classifier (Golub, 1999). The preferred feature is x2. (b) SVM. The preferred feature is x1. (Guyon et al., 2002)

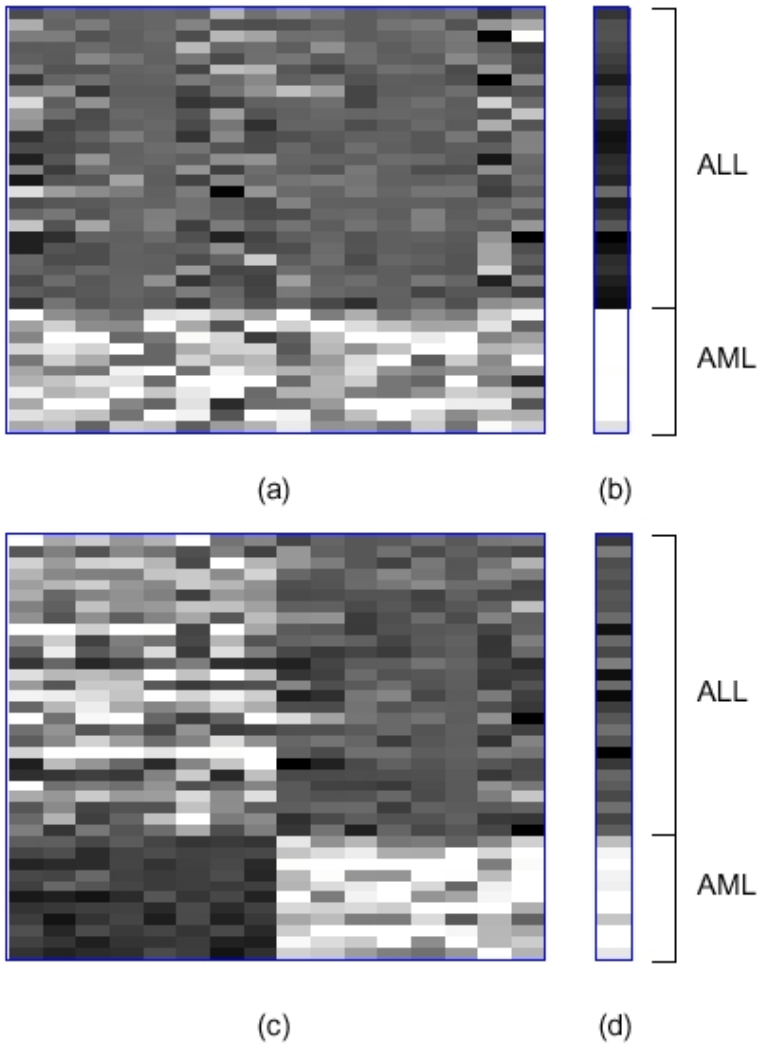

Figure 4: Best sets of 16 genes (Leukemia data). In matrices (a) and (c), the columns represent different genes and the lines different patients from the training set. The 27 top lines are ALL patients and the 11 bottom lines are AML patients. The gray shading indicates gene expression: the lighter the stronger. (a) SVM best 16 genes. Genes are ranked from left to right, the best one at the extreme left. All the genes selected are more AML correlated. (b) Weighted sum of the 16 SVM genes used to make the classification decision. A very clear ALL/AML separation is shown. (c) Baseline method (Golub, 1999) 16 genes. The method imposes that half of the genes are AML correlated and half are ALL correlated. The best genes are in the middle. (d) Weighted sum of the 16 baseline genes used to make the classification decision. The separation is still good, but not as contrasted as the SVM separation. (Guyon et al., 2002)

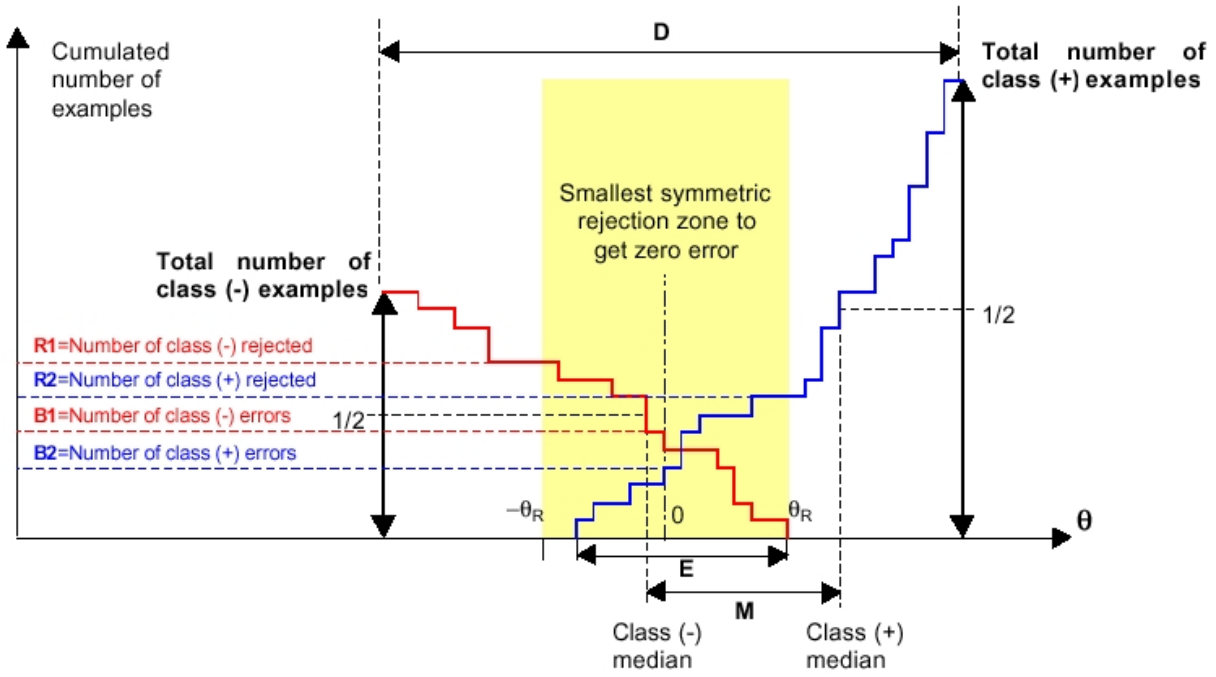

Figure 5: Metrics of classifier quality. The red and blue curves represent example distributions of two classes: class (-) and class (+). Red: Number of examples of class (-) whose decision function value is larger than or equal to θ. Blue: Number of examples of class (+) whose decision function value is smaller than or equal to θ. The number of errors B1 and B2 are the ordinates of $\theta = 0$. The number of rejected examples R1 and R2 are the ordinates of $-\theta_R$ and θ_R in the red and blue curves respectively. The decision function value of the rejected examples is smaller than θ_R in absolute value, which corresponds to examples of low classification confidence. The threshold θ_R is set such that all the remaining "accepted" examples are well classified. The extremal margin E is the difference between the smallest decision function value of class (+) examples and the largest decision function value of class (-) examples. On the example of the figure, E is negative. If the number of classification error is zero, E is positive. The median margin M is the difference in median decision function value of the class (+) density and the class (-) density. (Guyon et al., 2002)

and the classifier parameters each time. This way, the performance of a classifier using a given number of features can be assessed, not the predictive power of a given feature subset. This later problem is better addressed using an independent test set.

Gene selection and biological interpretation is still a research topic. The results for SVM-based extraction methods are encouraging.

References

- I. Guyon, J. Weston, S. Barnhill, V. Vapnik. Gene selection for cancer classification using support vector machines. Machine Learning, 46(1/3):389-422, Jan 2002.

Conclusions

SVMs have been tested on many bioinformatics problems in recent years. In many cases SVMs outperform other classification methods. However comparison is sometimes difficult because not all problems are stated as a clean machine learning problem. Handling multiclass is still not trivial (not only for SVMs). In today's examples, SVMs were only used as a replacement for Neural Networks or Fisher discriminant. See tomorrow for examples where SVMs provide more than that.

Combining different data sources

In order to further optimize clinical and biological predictions, it is possible to integrate heterogeneous data, like for example expert knowledge, anamnestical data, clinical data, histopathological data, image data (e.g. echos, scans),... This integration can be carried out in three different ways, depending on the time of integration. In an early phase, the data vectors of the microarray expression data and the external data can be concatenated into a single data vector (Vert et al., 2003). In an intermediate phase, it is possible to form a combined kernel by adding the microarray kernel and the external data kernel. In a late phase, two separate SVMs can be trained and the discriminant functions can then be added together (e.g. Committee networks) (Suykens et al., 2002).

Microarray data sets and pre-processing

Gene expression patterns are parallel measurements of the expression levels of thousands of genes simultaneously. These result in data vectors containing thousands of values. Microarrays consist of a reproducible pattern of thousands of DNA probes immobilized on a solid surface. Labeled cDNA, prepared from mRNA, is hybridized with the complementary

DNA probes spotted on the microarray. Hybridization can be measured by a laser scanner and quantitatively assessed. Two important types of microarrays are currently available: cDNA-microarrays and oligonucleotide microarrays. cDNA-microarrays (or spotted arrays) consist of ten thousands of known cDNAs mechanically deposited onto modified glass slides by contact or ink jet printing. Oligonucleotide microarrays (or DNA chips, Affymetrix) are produced by the synthesis of oligonucleotides on silicium chips. Both techniques have their own characteristics, which have to be taken into consideration when pre-processing the data. With cDNA microarrays, only relative expression levels (test/reference ratio) can be obtained. It is therefore a differential technique, which intrinsically normalizes for noise and background. An overview of the procedure that can be followed with cDNA microarrays is given in figure 6. With oligonucleotide microarrays absolute expression levels are obtained (no ratios). This technique needs additional mismatch control oligonucleotides (identical to the perfect match probes except for a single base-pair mismatch) to be added. These control probes allow estimation of cross-hybridization. More information about microarray data sets and pre-processing can be found in (Moreau et al., 2002).

Some publicly available microarray data sets

Leukemia data set:

- Bone marrow or peripheral blood samples are taken from 72 patients with either acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL). The data is split into a training set consisting of 38 samples of which 27 are ALL and 11 are AML, and the test set of 24 samples, 20 ALL and 14 AML. The dataset contains expression levels for 7129 human genes produced by Affymetrix high-density oligonucleotide microarrays. The scores in the dataset represent the intensity of gene expression after being re-scaled to make overall intensities for each chip equivalent. Following the methods in (Golub et al., 1999), normalization of these scores should be performed for each gene by subtracting the mean and dividing by the standard deviation of the expression values for that gene.

- 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. "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", Science 1999; 286: 531-537.

Breast cancer data set:

- A training data set containing 78 primary breast cancers was selected: 34 from patients who developed distant metastases within 5 years, 44 from patients who continued to be disease-free after a period of at least 5 years. All sporadic patients were

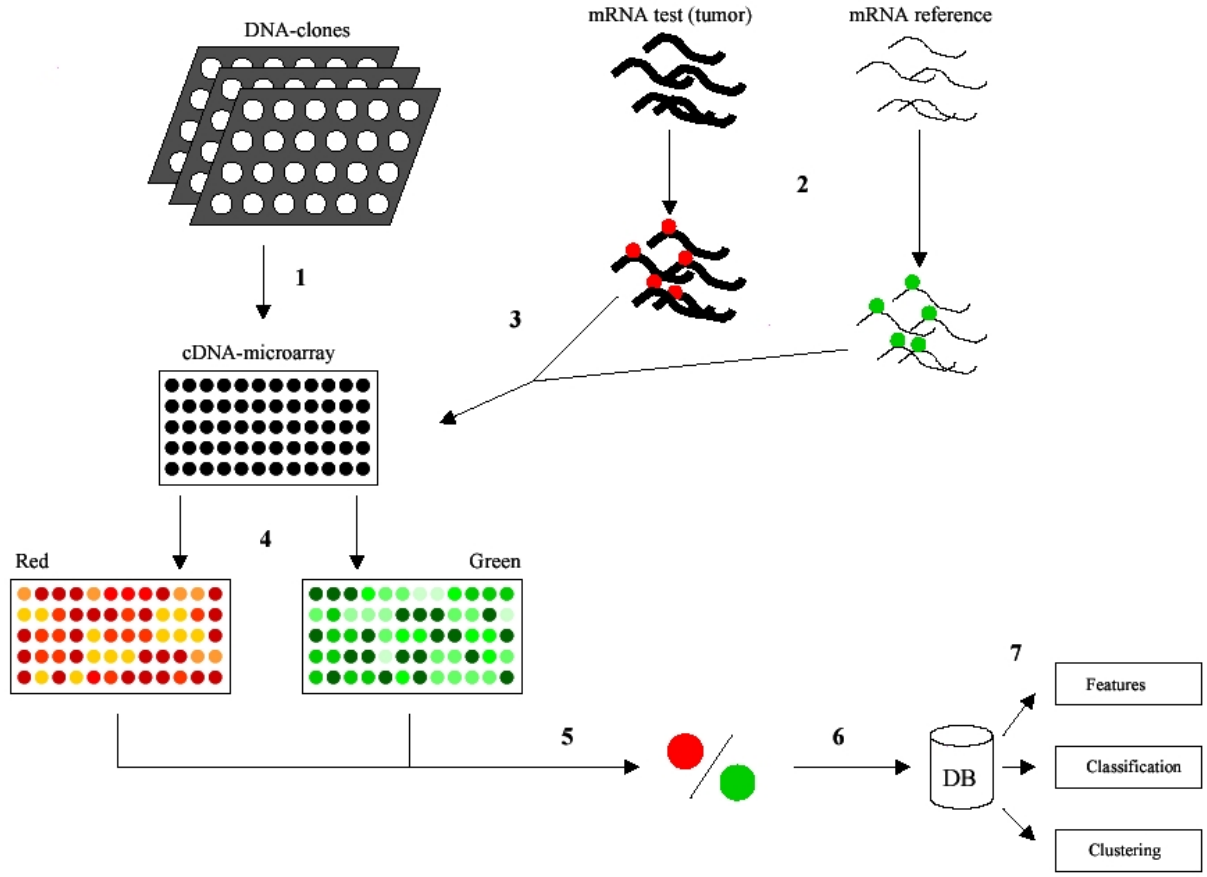

Figure 6: Schematic overview of an experiment with a cDNA-microarray. (1) Spotting of the pre-synthesized DNA-probes (derived from the genes to be studied) on the glass slide. These probes are the purified products from PCR-amplification of the associated DNA-clones. (2) Labeling (via reverse transcriptase) of the total mRNA of the test sample (tumor - red) and reference sample (green). (3) Pooling of the two samples and hybridization (4) Read-out of the red and green intensities separately (measure for the hybridization by the test and reference sample) in each probe. (5) Calculation of the relative expression levels (intensity in the red channel / intensity in the green channel). (6) Storage of results in a database. (7) Data mining. (De Smet et al., 2001)

lymph node negative, and under 55 years of age at diagnosis. From each patient, 5 μg total RNA was isolated from snap-frozen tumor material and used to derive complementary RNA (cRNA). A reference cRNA pool was made by pooling equal amounts of cRNA from each of the sporadic carcinomas. Two hybridizations were carried out for each tumor using a fluorescent dye reversal technique on microarrays containing 24481 human genes synthesized by inkjet technology. Fluorescence intensities of scanned images were quantified, normalized and corrected to yield the transcript abundance of a gene as an intensity ratio with respect to that of the signal of the reference pool. An additional independent set of primary tumors from 19 young, lymph-node-negative breast cancer patients was selected. This group consisted of 7 patients who remained metastasis free for at least 5 years, and 12 patients who developed distant metastases within 5 years.

- van 't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A. M., 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. "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer", Nature 2002; 415: 530-536.

Colon tumor data set:

- Using Affymetrix oligonucleotide arrays, expression levels for 40 tumor and 22 normal colon tissues (normal biopsies are from healthy parts of the colons of the same patients) were measured for 6500 human genes. Of these genes, the 2000 with the highest minimal intensity (highest confidence in the measured expression level) across the tissues were selected for classification purposes, and these were made publicly available. The scores in the dataset represent a gene intensity derived in a process described in (Alon et al., 1999). The data should not be processed further before performing classification.

- Alon, U., N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, PNAS 1999; 96(12): 6745-50.

References

F. De Smet, K. Marchal, D. Timmerman, B. De Moor, Y. Moreau. **Gebruik van microroosters in de klinische oncologie.** *Tijdschrift voor Geneeskunde*, vol. 57, no. 18, Sep. 2001, pp. 1225-1236.

Y. Moreau, K. Marchal, J. Mathys, **Computational biomedicine : a multidisciplinary crossroads**, Siemens Prize, FWO (Flanders, Belgium), 2002, 89 p.

Y. Moreau, F. De Smet, G. Thijs, K. Marchal, B. De Moor. **Functional bioinformatics of microarray data : from expression to regulation.** *Proceedings of the IEEE*, vol. 90, no. 11, Nov. 2002, pp. 1722-1743.

J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, **Least Squares Support Vector Machines**, World Scientific, Singapore, 2002 (ISBN 981-238-151-1).

J.-P. Vert and M. Kanehisa. **Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA**, To appear in *Advances in Neural Information Processing Systems 15*, MIT Press, Cambridge, MA, 2003.