**Seminar 28.9. – 2.10 2003 St. Gallen**
**Region Oesterreich – Schweiz (ROeS) of the**
**International Biometric Society**

# Analyses of DNA Methylation in Serum of Cancer Patients Using Least Squares Support Vector Machines

**Georg Göbel**
**Institute of Biostatistics and Documentation**
**University of Innsbruck**

**Abstract**

Changes in the status of DNA methylation are one of the most common molecular alterations in human neoplasia. As it is possible to detect these epigenetic alterations in patients´ bloodstream, we investigated whether aberrant DNA methylation in patients' pretherapeutic sera is of prognostic significance in breast cancer.We apply least square support vector machine (LS-SVM) classifiers to clinical data of 339 pretherapeutic serum samples of recently diagnosed primary breast cancer patients. Linear and RBF kernels are used within the model. The concept of the application of LS-SVM classifiers within the context of DNA methylation analysis will be presented.

## Clinical Background

Involvement of axillary lymph nodes and tumor size are the most important prognostic factors in breast cancer.(1-4) Although the presence or absence of metastatic involvement in the axillary lymph nodes is the most powerful prognostic factor available for patients with primary breast cancer, it is only an indirect measure reflecting the tumors' tendency to spread. In approximately one-third of women with breast cancer and negative lymph nodes the disease recurs, while about one-third of patients with positive lymph nodes are free of recurrence ten years after loco-regional therapy.(2;3) These data highlight the need for more sensitive and specific prognostic indicators, ideally reflecting the presence or absence of tumor-specific alterations in the bloodstream that may eventually even after years lead to metastasis. It is now widely accepted that adjuvant systemic therapy substantially improves disease-free and overall survival in both pre- and postmenopausal women up to the age of 70 years with lymph node-negative or lymph node-positive breast cancer.(2;3) It is also generally accepted that patients with poor prognostic features benefit the most from adjuvant therapy, whereas some patients with good prognostic features may be overtreated.(1;4;5) Moreover many other factors have been investigated for their potential to predict disease outcome, but in general they have only limited predictive value.(4) Recently, interesting prognostic parameters including gene-expression profiles,(6;7) cell cycle regulating proteins(8) and occult cytokeratin-positive metastatic cells in the bone marrow(9) have been added to the list of prognostic factors, but their prognostic relevance needs to be further evaluated.

## Methods

Support vector machines have been introduced for solving pattern recognition and function estimation problems by Vapnik 1995. According to the data is mapped to a higher

dimensional feature space and an optimal separating hyperplane is constructed in this space. Mercer's theorem enables to avoid an explicit formulation of this nonlinear mapping. The solution is written as a weighted sum of the data points. In the original formulation a quadratic programming problem is solved, which yields many zero weights. The data points corresponding to the non-zero weights are called support vectors. Kernel function parameters can be chosen such that a bound on the generalization error is minimized, expressed in terms of the VC dimension. There is a possibility of using polynomials, splines, radial basis function (RBF) networks or multilayer perceptrons as kernels. It is shown in Suykens and Vandewalle (14), that it is not absolutely necessary to apply Mercer's condition.

**Least-square Support Vector Machines**

In least squares support vector machines, equality constraints are used instead of inequality constraints as well as a least squares error term in order to obtain a linear set of equations in the dual space (Fig.1). However, to achieve a high level of performance, some parameters in the LS-SVM model must be tuned. These adjustable hyperparameters include: a regularization parameter which determines the tradeoff between minimizing the training errors and minimizing the model complexity; and a kernel parameter such as the width of the RBF kernel. One popular way to choose the hyperparameters is cross-validation.

Fig. 1
LS-SVM primal optimization problem and equality constraints

$$\min_{w,b,e} \mathcal{J}(w,e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{i=1}^{N} e_i^2 \quad \text{s.t.} \quad y_i \left[ w^T \varphi(x_i) + b \right] = 1 - e_i, \ \forall i$$

**Serum samples and DNA isolation**

The *dataset* consisted of pretherapeutic serum samples of recently diagnosed primary breast cancer patients (n=339; age range: 28.17 yrs to 91.00 yrs. (mean: 58.7 yrs.)). 61 patients died within 12 years. Clinico-pathological features are shown in Table 1. All patients underwent a core biopsy and were confirmed to have malign disease of the breast.

Table 1

| | | | |
|---|---|---|---|
| Demografic | No. of patients | | 339 |
| | Age | | 58.8 +/- 13.0 |
| | Postmenopausal | | 246 |
| Clinico-pathological F. | Therapy: | No adj. Th. | 57 |
| | | Chemo-Th. | 86 |
| | | Endocrine-Th. | 121 |
| | | Chemo-/ Endocrine Th. | 75 |
| | PT –Staging | T1 | 192 |
| | | T2 | 106 |
| | | T3 + T4 | 41 |
| | PN – Staging | N0 | 191 |
| | | N1 – N3 | 68 |
| | | > N3 | 80 |
| | Death | | 61 |
| Gene – Methylation (no. of patients with methylated Gene) | Gene 1 | | 54 |
| | Gene 2 | | 79 |
| | Gene 3 | | 31 |
| | Gene 4 | | 185 |
| | Gene 5 | | 70 |

The primary surgical procedure included breast-conserving lumpectomy or modified radical mastectomy and axillary lymph node dissection. 196 patients received systemic adjuvant treatment after surgery.

First, genomic DNA was isolated from serum samples. Thereafter, the sodium bisulfit-treated genomic DNA was analysed by means of MethyLight, a fluorescence-based, real-time PCR assay (10), which yielded in a quantitive PMR (percentage of fully methylated reference)

measurement.

Data-preprocessing resulted in 10 predictors (5 clinicopathological features, 5 Genes well-known to be frequently methylated in breast cancer and other malignancies) and 1 outcome variable (death). For SVM analysis we used normalised logarithms of the PMR values.

**Application of LS-SVMs to the serum data** (see [13])

We applied the LS-SVMs to our dataset and assessed the performance by receiver operator characteristic curve (ROC) analysis. All experiments were conducted using Matlab R12 and the LS-SVM 1.4 Toolbox (www.esat.kuleuven.ac.be/sista/lssvmlab).

The first step of the model fitting procedure is the construction of an LS-SVM classifier. At this stage it is a continous number, which can be positive or negative and is located around + 1 or – 1. At the second stage, the output probability is computed, indicating the posterior probability for a patient to survive. We used LS-SVM classifiers with linear and RBF kernels.

The crossvalidation was carried out as follows: The data was once permutated randomly, then it was divided into 10 (stratified) disjunct sets holding the same proportion of surviving patients. In the i-th iteration, the i-th set was used to estimate the performance ('validation set') of the model trained on the other 9 sets ('training set'). At last, the 10 different estimates of the performance were combined by the 'mean'. The assumption was made that the input data are distributed independently and identically over the input space.

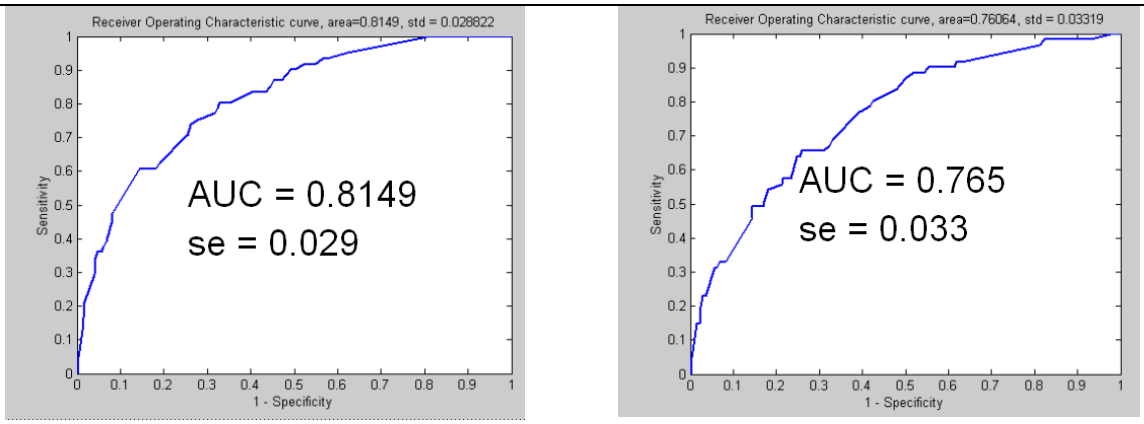
For the evaluation of the model we used ROC analysis, as it is independent of class distributions or error costs and has been widely used in the biomedical field.

## Results

Our experiments show that relations between DNA methylation and carcinogenesis are worth to being analysed using LS-SVM models. The application of the framework revealed acceptable results (Fig. 2). RBF Kernels performed slightly better than linear Kernels, which corresponds to several publications (12;14) and leads to the speculation, that DNA methylation problems are as non-linear as DNA expression. The fact, that serum samples had been gathered during more than twelve years, must be discussed in two ways. Despite the fact, that no differences in demografic or clinicopathological features over time could be found within the dataset, the criteria for the decision for adjuvant treatments in breast cancer (ER status etc.) have changed, especially for postmenopausal patients. Additionally the quality of frozen serum samples, which are older than 10 years, might be an issue.

Fig. 2

Evaluation of RBF (left) and the linear Kernel LS-SVM Model (right)




## Conclusion

Performing this analysis we gathered first experiences in applying LS-SVM classifiers to DNA methylation data. The LS-SVM Model may additionally be applicable to unsupervised

learning methods. Several issues like variable selection and transformation must be considered. The LS-SVM Toolbox offers a broad range of features and provides a stable framework for further analyses. It needs to be investigated, wether the model can assist clinicians in making correct diagnoses or supporting decisions for correct therapies.

**References**

1. Goldhirsch A, Glick JH, Gelber RD, Coates AS, Senn HJ. Meeting highlights: International Consensus Panel on the Treatment of Primary Breast Cancer. Seventh International Conference on Adjuvant Therapy of Primary Breast Cancer. J.Clin.Oncol. 2001;19(18):3817-27.

2. Early Breast Cancer Trialists' Collaborative Group. Polychemotherapy for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group. Lancet 1998;352(9132):930-42.

3. Early Breast Cancer Trialists' Collaborative Group. Tamoxifen for early breast cancer: an overview of the randomised trials. Early Breast Cancer Trialists' Collaborative Group. Lancet 1998;351(9114):1451-67.

4. Hayes DF, Isaacs C, Stearns V. Prognostic factors in breast cancer: current and new predictors of metastasis. J.Mammary.Gland.Biol.Neoplasia. 2001;6(4):375-92.

5. Hayes DF, Trock B, Harris AL. Assessing the clinical impact of prognostic factors: when is "statistically significant" clinically useful? Breast Cancer Res.Treat. 1998;52(1-3):305-19.

6. 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415(6871):530-6.

7. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW et al. A gene-expression signature as a predictor of survival in breast cancer. N.Engl.J.Med. 2002;347(25):1999-2009.

8. Keyomarsi K, Tucker SL, Buchholz TA, Callister M, Ding Y, Hortobagyi GN et al. Cyclin E and survival in patients with breast cancer. N.Engl.J.Med. 2002;347(20):1566-75.

9. Braun S, Pantel K, Muller P, Janni W, Hepp F, Kentenich CR et al. Cytokeratin-positive cells in the bone marrow and survival of patients with stage I, II, or III breast cancer. N.Engl.J.Med. 2000;342(8):525-33.

10. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nat.Rev.Genet. 2002;3(6):415-28.

11. Golub et. al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286 (1999) p531.

12. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D.Support vector machine classification and validation of cancer tissues samples using microarray expression data. Furey et. al. Bioinformatics 16/10 (2000) p961.

13. Lu C, Van Gestel T, Suykens JA, Van Huffel S, Vergote I, Timmerman D.. Preoperative prediction of malignancy of ovarian tumors using least squares support vector machines. Artificial Intelligence in Medicine 28 (2003) p281.

14. Suykens JA, Vandewalle J, De Moor B. Optimal control by least squares support vector machines. Neural Netw. 2001 Jan;14(1):23-35.