INTEGRATING MICROARRAY AND PROTEOMICS DATA TO PREDICT THE RESPONSE ON CETUXIMAB IN PATIENTS WITH RECTAL CANCER

ANNELEEN DAEMEN[1]*, OLIVIER GEVAERT[1], TIJL DE BIE[2], ANNELIES DEBUCQUOY[3], JEAN-PASCAL MACHIELS[4], BART DE MOOR[1] AND KARIN HAUSTERMANS[3]

[1] *Katholieke Universiteit Leuven, Department of Electrical Engineering (ESAT), SCD-SISTA (BIOI), Kasteelpark Arenberg 10 - bus 2446, B-3001 Leuven (Heverlee), Belgium*

[2] *University of Bristol, Department of Engineering Mathematics, Queen's Building, University Walk, Bristol, BS8 1TR, UK*

[3] *Katholieke Universiteit Leuven / University Hospital Gasthuisberg Leuven, Department of Radiation Oncology and Experimental Radiation, Herestraat 49, B-3000 Leuven, Belgium*

[4] *Université Catholique de Louvain, St Luc University Hospital, Department of Medical Oncology, Ave. Hippocrate 10, B-1200 Brussels, Belgium*

To investigate the combination of cetuximab, capecitabine and radiotherapy in the preoperative treatment of patients with rectal cancer, forty tumour samples were gathered before treatment (T_0), after one dose of cetuximab but before radiotherapy with capecitabine (T_1) and at moment of surgery (T_2). The tumour and plasma samples were subjected at all timepoints to Affymetrix microarray and Luminex proteomics analysis, respectively. At surgery, the Rectal Cancer Regression Grade (RCRG) was registered. We used a kernel-based method with Least Squares Support Vector Machines to predict RCRG based on the integration of microarray and proteomics data on T_0 and T_1. We demonstrated that combining multiple data sources improves the predictive power. The best model was based on 5 genes and 10 proteins at T_0 and T_1 and could predict the RCRG with an accuracy of 91.7%, sensitivity of 96.2% and specificity of 80%.

1. Introduction

A recent challenge for genomics is the integration of complementary views of the genome provided by various types of genome-wide data. It is likely

*To whom correspondence should be addressed: anneleen.daemen@esat.kuleuven.be

that these multiple views contain different, partly independent and complementary information. In the near future, the amount of available data will increase further (e.g. methylation, alternative splicing, metabolomics, etc). This makes data fusion an increasingly important topic in bioinformatics.

Kernel Methods and in particular Support Vector Machines (SVMs) for supervised classification are a powerful class of methods for pattern analysis, and in recent years have become a standard tool in data analysis, computational statistics, and machine learning applications.[1–2] Based on a strong theoretical framework, their rapid uptake in applications such as bioinformatics, chemoinformatics, and even computational linguistics, is due to their reliability, accuracy, computational efficiency, demonstrated in countless applications, as well as their capability to handle a very wide range of data types and to combine them (e.g. kernel methods have been used to analyze sequences, vectors, networks, phylogenetic trees, etc). Kernel methods work by mapping any kind of input items (be they sequences, numeric vectors, molecular structures, etc) into a high dimensional space. The embedding of the data into a vector space is performed by a mathematical object called a 'kernel function' that can efficiently compute the inner product between all pairs of data items in the embedding space, resulting into the so-called kernel matrix. Through these inner products, all data sets are represented by this real-valued square matrix, independent of the nature or complexity of the objects to be analyzed, which makes all types of data equally treatable and easily comparable.

Their ability to deal with complexly structured data made kernel methods ideally positioned for heterogeneous data integration. This was understood and demonstrated in 2002, when a crucial paper integrated amino-acid sequence information (and similarity statistics), expression data, protein-protein interaction data, and other types of genomic information to solve a single classification problem: the classification of transmembrane versus non transmembrane proteins.[3] Thanks to this integration of information a higher accuracy was achieved than what was possible based on any of the data sources separately. This and related approaches are now widely used in bioinformatics.[4–6]

Inspired by this idea we adapted this framework which is based on a convex optimization problem solvable with semi-definite programming (SDP). As supervised classification algorithm, we used Least Squares Support Vector Machines (LS-SVMs) instead of SVMs. LS-SVMs are easier and faster for high dimensional data because the quadratic programming problem is converted into a linear problem. Secondly, LS-SVMs are also more suitable

as they contain regularization which allows tackling the problem of overfitting. We have shown that regularization seems to be very important when applying classification methods on high dimensional data.[7]

The algorithm described in this paper will be adapted on data of patients with rectal cancer. To investigate the combination of cetuximab, capecitabine and radiotherapy in the preoperative treatment of patients with rectal cancer, microarray and proteomics data were gathered from forty rectal cancer patients at three timepoints during therapy. At surgery, different outcomes were registered but here we focus on the Rectal Cancer Regression Grade[8] (RCRG), a pathological staging system based on Wheeler for irradiated rectal cancer. It includes a measurement of tumour response after preoperative therapy. In this paper, patients were divided into two groups which we would like to distinguish: the positive group (RCRG pos) contained Wheeler 1 (good responsiveness; tumour is sterilized or only microscopic foci of adenocarcinoma remain); the negative group (RCRG neg) consisted of Wheeler 2 (moderate responsiveness; marked fibrosis but with still a macroscopic tumour) and Wheeler 3 (poor responsiveness; little or no fibrosis with abundant macroscopic tumour). We refer the readers to Ref. 9 for more details about the study and the patient characteristics.

In this paper, we would like to demonstrate that integrating multiple available data sources in an appropriate way using kernel methods increases the predictive power compared to models built only on one data set. The developed algorithm will be demonstrated on rectal cancer patient data. The goal is to predict at T_1 (= before the start of radiotherapy) the RCRG.

2. Data sources

Forty patients with rectal cancer (T3-T4 and/or N+) from seven Belgian centers were enrolled in a phase I/II study investigating the combination of cetuximab, capecitabine and radiotherapy in the preoperative treatment of patients with rectal cancer.[9] Tissue and plasma samples were gathered before treatment (T_0), after one dose of cetuximab but before radiotherapy with capecitabine (T_1) and at moment of surgery (T_2). At all these three timepoints, the frozen tissues were used for Affymetrix microarray analysis while the plasma samples were used for Luminex proteomics analysis. Because we had to exclude some patients, ultimately the data set contained 36 patients.

The samples were hybridized to Affymetrix human U133 2.0 plus gene

chip arrays. The resulting data was first preprocessed for each timepoint separately using RMA.[10] Secondly, the probe sets were mapped on Entrez Gene Ids by taking the median of all probe sets that matched on the same gene. Probe sets that matched on multiple genes were excluded and unknown probe sets were given an arbitrary Entrez Gene Id. This reduces the number of features from 54613 probe sets to 27650 genes. Next, one can imagine that the number of differentially expressed genes will be much lower than these 27650 genes. Therefore, a prefiltering without reference to phenotype can be used to reduce the number of genes. Taking into account the low signal-to-noise ratio of microarray data, we decided to filter out genes that show low variation across all samples. Only retaining the genes with a variance in the top 25% reduces the number of features at each timepoint to 6913 genes.

The proteomics data consist of 96 proteins, previously known to be involved in cancer, measured for all patients in a Luminex 100 instrument. Proteins that had absolute values above the detection limit in less than 20% of the samples were excluded for each timepoint separately. This results in the exclusion of six proteins at T_0, four at T_1 and six at T_2. The proteomics expression values of transforming growth factor alpha (TGFα), which had also too many values below the detection limit, were replaced by the results of ELISA tests performed at the Department of Experimental Oncology in Leuven. For the remaining proteins the missing values were replaced by half of the minimum detected for each protein over all samples, and values exceeding the upper limit were replaced by the upper limit value. Because most of the proteins had a positively skewed distribution, a log transformation (base 2) was performed.

In this paper, only the data sets at T_0 and T_1 were used because the goal of the models is to predict before start of chemoradiation the RCRG.

3. Methodology

3.1. *Kernel methods and LS-SVMs*

Kernel methods are a group of algorithms that do not depend on the nature of the data because they represent data entities through a set of pairwise comparisons called the kernel matrix. The size of this matrix is determined only by the number of data entities, whatever the nature or the complexity of these entities. For example a set of 100 patients each characterized by 6913 gene expression values is still represented by a 100×100 kernel matrix.[4] Similarly as 96 proteins characterized by their 3D structure are

also represented by a 100 × 100 kernel matrix. The kernel matrix can be geometrically expressed as a transformation of each data point x to a high dimensional feature space with the mapping function $\Phi(x)$. By defining a kernel function $k(x_k, x_l)$ as the inner product $\langle \Phi(x_k), \Phi(x_l) \rangle$ of two data points x_k and x_l, an explicit representation of $\Phi(x)$ in the feature space is not needed anymore. Any symmetric, positive semidefinite function is a valid kernel function, resulting in many possible kernels, e.g. linear, polynomial and diffusion kernels. They all correspond to a different transformation of the data, meaning that they extract a specific type of information from the data set. Therefore, the kernel representation can be applied to many different types of data and is not limited to vectorial or matrix form.

An example of a kernel algorithm for supervised classification is the Support Vector Machine (SVM) developed by Vapnik and others.[11] Contrary to most other classification methods and due to the way data is represented through kernels, SVMs can tackle high dimensional data (e.g. microarray data). The SVM forms a linear discriminant boundary in feature space with maximum distance between samples of the two considered classes. This corresponds to a non-linear discriminant function in the original input space. A modified version of SVM, the Least Squares Support Vector Machine (LS-SVM), was developed by Suykens *et al.*[12−13] On high dimensional data sets this modified version is much faster for classification because a linear system instead of a quadratic programming problem needs to be solved. The LS-SVM also contains regularization which tackles the problem of overfitting. In the next section we describe the use of LS-SVMs with a normalized linear kernel to predict the RCRG in rectal cancer patients based on the kernel integration of microarray and proteomics data at T_0 and T_1.

3.2. *Data fusion*

There exist three ways to learn simultaneously from multiple data sources using kernel methods: early, intermediate and late integration.[14] Figure 1 gives a global overview of these three methods in the case of 2 available data sets. In this paper, intermediate integration is chosen. In this way, kernel functions can be better adapted to each data set separately. And by adding the kernel matrices before training the LS-SVM, only one predicted outcome per patient is obtained which makes an extra decision function unnecessary.

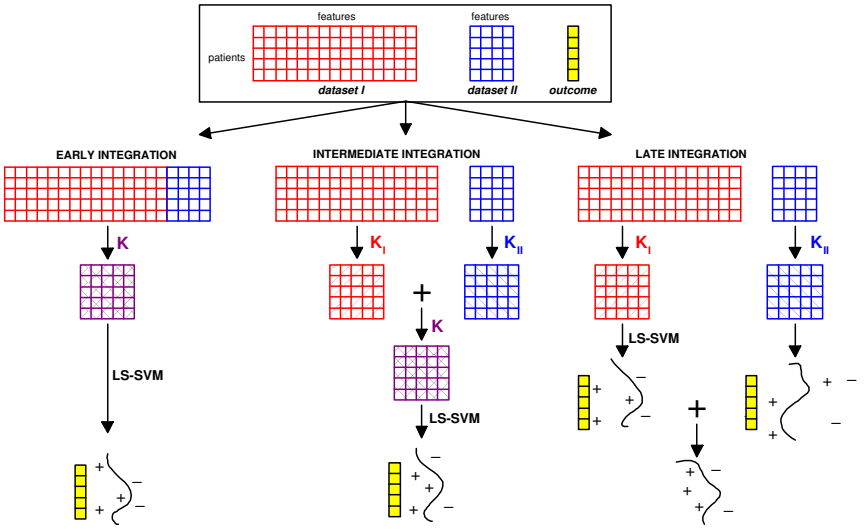

Figure 1. *Three methods to learn from multiple data sources.* In early integration, an LS-SVM is trained on the kernel matrix, computed from the concatenated data set. In intermediate integration, a kernel matrix is computed for both data sets and an LS-SVM is trained on the sum of the kernel matrices. In late integration, two LS-SVMs are trained separately for each data set. A decision function results in a single outcome for each patient.

3.3. *Model building*

In this paper, the normalized linear kernel function was used:

$$k(x_k, x_l) = k(x_k, x_l)/\sqrt{k(x_k, x_k)k(x_l, x_l)} \tag{1}$$

with $k(x_k, x) = x_k^T x$ instead of the linear kernel function $k(x_k, x_l) = x_k^T x_l$. With the normalized version, the values in the kernel matrix will be bounded because the data points are projected onto the unit sphere, while these elements can take very large values without normalization. Normalizing is thus required when combining multiple data sources to guarantee the same order of magnitude for the kernel matrices of the data sets.

There are four data sets that have to be combined: microarray data at T_0, at T_1 and proteomics data at T_0 and at T_1. Because each data set is represented by a kernel matrix, these data sources can be integrated in a straightforward way by adding the multiple kernel matrices according to intermediate integration explained previously. In this combination, each of the matrices is given a specific weight μ_i. The resulting kernel matrix

is given in Eq. 2. Positive semidefiniteness of the linear combination of kernel matrices is guaranteed when the weights μ_i are constrained to be non-negative.

$$K = \mu_1 K_1 + \mu_2 K_2 + \mu_3 K_3 + \mu_4 K_4. \quad (2)$$

The choices of the weights are important. Previous studies have shown that the optimization of the weights only leads to a better performance when some of the available data sets are redundant or contain much noise.[3] In our case we believe that the microarray and proteomics data sets are equally reliable based on our results of LS-SVMs on each data source separately (data not shown). Therefore to avoid optimizing the weights, they were chosen equally: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = 0.25$.

Due to the data set size, we chose a leave-one-out cross-validation (LOO-CV) strategy to estimate the generalization performance (see Fig. 2). Since both classes were unbalanced (26 RCRG pos and 10 RCRG neg), the minority class was resampled in each LOO iteration by randomly duplicating a sample from the minority class and adding uniform noise ([0,0.1]). This was repeated until the number of samples in the minority class was at least 70% of the majority class (chosen without optimization).

After choosing the weights fixed, three parameters are left that have to be optimized: the regularization parameter γ of the LS-SVM, the number of genes used from the microarray data sets both at T_0 and T_1 and the number of proteins used from the proteomics data sets. To accomplish this, a three-dimensional grid was defined as shown in Fig. 2 on which the parameters are optimized by maximizing a criterion on the training set. The possible values for γ on this grid range from 10^{-10} to 10^{10} on a logarithmic scale. The possible number of genes that were tested are 5, 10, 30, 50, 100, 300, 500, 1000, 3000 and all genes. The number of proteins used are 5, 10, 25, 50 and all proteins. Genes and proteins were selected by ranking these features using the Wilcoxon rank sum test. In each LOO-CV iteration, a model is built for each possible combination of parameters on the 3D-grid. Each model with the instantiated parameters is evaluated on the left out sample. This whole procedure is repeated for all samples in the set. The model with the highest accuracy is chosen. If multiple models with equal accuracy, the model with the highest sum of sensitivity and specificity is chosen.

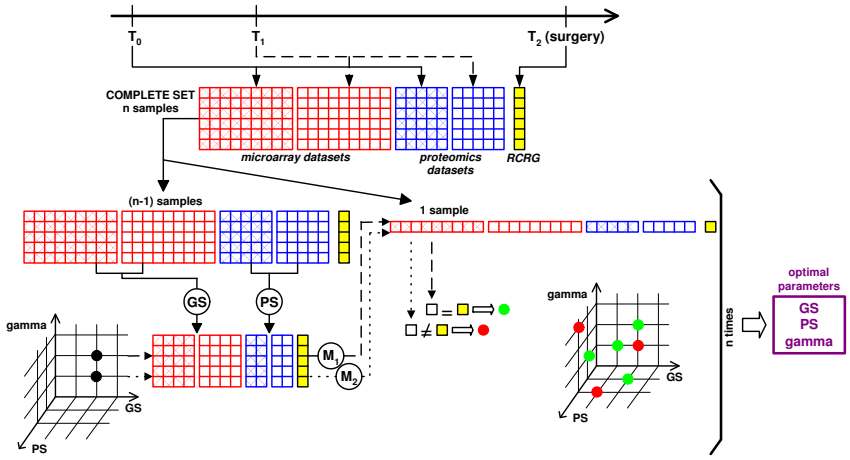

Figure 2. *Methodology for developing a classifier.* The available data contains microarray data and proteomics data both at T_0 and T_1. The regularization parameter γ and the number of genes (GS) and proteins (PS) are determined with a leave-one-out cross-validation strategy on the complete set. In each leave-one-out iteration, an LS-SVM model is trained on the most significant genes and proteins for all possible combinations of γ and the number of features. This gives a globally best parameter combination (γ,GS,PS).

4. Results

We evaluated our methodology as described in Sec. 3.3 on the rectal cancer data set to predict the Rectal Cancer Regression Grade. The model with the highest performance accuracy and an as high as possible sum of sensitivity and specificity was built on the five most significant genes and the ten most significant proteins at T_0 and T_1 according to the RCRG. From now on, we refer to this model as MPIM (Microarray and Proteomics Integration Model). To evaluate its performance, 6 other models were built on different combinations of data sources using the same model building strategy: MMT0 (Microarray Model at T_0: all microarray data at T_0), MMT1 (Microarray Model at T_1: all microarray data at T_1), MIM (Microarray Integration Model: microarray data at both timepoints), PMT0 (Proteomics Model at T_0: all proteomics data at T_0), PMT1 (Proteomics Model at T_1: all proteomics data at T_1) and PIM (Proteomics Integration Model: proteomics data at both timepoints).

Table 1 gives an overview of the performances of all these models. MPIM predicts the RCRG correctly in 33 of the 36 patients (=91.7%). Almost all

patients with RCRG positive are predicted correctly with a sensitivity of 96.2% and a positive predictive value of 0.926. Of the patients with RCRG negative, 80% are classified correctly. None of the other models performs better for one of the performance parameters shown in Table 1.

Table 1. Performance of MPIM compared to models based on different combinations of data sources.

Model	Nb genes	Nb proteins	TP	FP	FN	TN	Sens (in %)	Spec (in %)	PPV	NPV	Accuracy (in %)
MPIM	5	10	25	2	1	8	96.2	80	0.926	0.889	91.7 (33/36)
MMT0	1000	-	25	10	1	0	96.2	0	0.714	0	69.4 (25/36)
MMT1	3000	-	23	6	3	4	88.5	40	0.793	0.571	75 (27/36)
MIM	30	-	25	10	1	0	96.2	0	0.714	0	69.4 (25/36)
PMT0	-	all	21	4	5	6	80.8	60	0.84	0.545	75 (27/36)
PMT1	-	5	23	2	3	8	88.5	80	0.92	0.727	86.1 (31/36)
PIM	-	25	21	3	5	7	80.8	70	0.875	0.583	77.8 (28/36)

TP, true positive; FP, false positive; FN, false negative; TN, true negative; Sens, sensitivity; Spec, specificity; PPV, positive predictive value; NPV, negative predictive value; Acc, predictive accuracy.

The MPIM is built on 5 genes different at T_0 and T_1, 9 proteins different at T_0 and T_1 and 1 protein selected at both timepoints (ferritin).

Among the 10 genes, several were related to cancer. Bone morphogenetic protein 4 (BMP4) is involved in development, morphogenesis, cell proliferation and apoptosis. This protein, upregulated in colorectal tumours, seems to help initiate the metastasis of colorectal cancer without maintaining these metastases.[15] Integrin alpha V (ITGAV) is a receptor on cell surfaces for extracellular matrix proteins. Integrins play important roles in cell-cell and cell-matrix interactions during a.o. immune reactions, tumour growth and progression, and cell survival. ITGAV is related to many cancer types among which prostate and breast cancer for which it is important in the bone environment to the growth and pathogenesis of cancer bone metastases.[16]

Several of the proteins have known associations with rectal and colon cancer, such as ferritin, TGFα, MMP-2 and TNFα. Ferritin, the major intracellular iron storage protein, is an indicator for iron deficiency anemia. This disease is recognized as a presenting feature of right-sided colon cancer and increases in men significantly the risk of having colon cancer.[17] The transforming growth factor alpha (TGFα) is upregulated in some human cancers among which rectal cancer.[18] In colon cancer, it promotes depletion of tumour-associated macrophages and secretion of amphoterin.[19] TGFα is

closely related to epidermal growth factor EGF, one of the other proteins on which MPIM is built. EGF plays an important role in the regulation of cell growth, proliferation and differentiation. The matrix metalloproteinase-2 (MMP-2), known to be implicated in rectal and colon cancer invasion and metastasis, is associated with a reduced survival of these patients when being higher expressed in the malignant epithelium and in the surrounding stroma.[20] The tumour necrosis factor TNFα has important roles in immunity and cellular remodelling and influences apoptosis and cell survival. Dysregulation and especially overproduction of TNFα have been observed to occur in colorectal cancer.[21] Some of the other proteins such as IL-4 and IL-6 are important for the immune system whose function depends for a large part on interleukins. IL-4 is involved in the proliferation of B cells and the development of T cells and mast cells. It also has an important role in allergic response. IL-6 regulates the immune response, modulates normal and cancer cell growth, differentiation and cell survival.[22] It causes increased steady-state levels of TGFα mRNA in macrophage-like cells.[23]

5. Discussion

We presented a framework for the combination of multiple genome-wide data sources in disease management using a kernel-based approach (see Fig. 2). Each data set is represented by a kernel matrix, based on a normalized linear kernel function. These matrices are combined according to the intermediate integration method illustrated in Fig. 1. Afterwards, an LS-SVM is trained on the combined kernel matrix. In this paper, we evaluated the resulting algorithm on our data set consisting of microarray and proteomics data of rectal cancer patients to predict the Rectal Cancer Regression Grade after a combination therapy consisting of cetuximab, capecitabine and radiotherapy. The best model (MPIM) is based on 5 genes and 10 proteins at T_0 and at T_1 and can predict the RCRG with an accuracy of 91.7%, sensitivity of 96.2% and specificity of 80%. Table 1 shows that the performance parameters of MPIM are better than or equal to the values of the other models. This demonstrates that microarray and proteomics data are partly complementary and that the performance of our algorithm in which these various views on the genome are integrated improves the prediction of response to therapy upon LS-SVMs trained on a combination of less data sources. Many of the genes and proteins on which the MPIM is built are related to rectal cancer or cancer in general.

We were inspired by the idea of Lanckriet[3] and others[4–6] to integrate

multiple types of genomic information to be able to solve a single classification problem with a higher accuracy than possible based on any of the genomic information sources separately. In the framework of Lanckriet, the problem of optimal kernel combination is formulated as a convex optimization problem using SVMs and is solvable with semi-definite programming (SDP) techniques. However, LS-SVMs are easier and faster for high dimensional data because the problem is formulated as a linear problem instead of a quadratic programming problem and LS-SVMs contain regularization which tackles the problem of overfitting. Instead of applying this approach to protein function in yeast which requires the reformulation of the problem in 13 binary classification problems (equal to the number of different functional classes), we applied a modified version of this framework in the patient space where many of the prediction problems are already binary. To the author's knowledge, this is the first time that a kernel-based integration method has been applied on multiple high dimensional data sets in the patient domain for studying cancer. Our results show that using information from different levels in the central dogma improves the classification performance.

We already mentioned that kernel methods have a large scope due to their representation of the data. However, when the amount of available data will increase in the near future, the choice of the weights becomes more important, especially when applying the algorithm to problems where the reliability of the data sources differs much or is not known a priori. In this paper, we chose the weights equally. We cannot guarantee though that without optimizing the weights of the different data sources we get the most optimal model. However, this increases the computational burden significantly.

When more data sources will become available in the future, they can be easily added to this framework. Additionally, we are currently investigating ways to improve the optimization algorithm, especially for the choice of the weights. Next, we will also apply more advanced feature selection techniques. At this moment a simple statistical test is used but more advanced techniques could be applied. Finally, we will compare kernel methods with other integration frameworks (e.g. Bayesian techniques).[24]

Acknowledgments

AD is research assistant of the Fund for Scientific Research - Flanders (FWO-Vlaanderen). This work is partially supported by: **1.** Research

Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys. **2.** Flemish Government: FWO: PhD/postdoc grants, G.0499.04 (Statistics), G.0302.07 (SVM/Kernel). **3.** Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, 2007-2011). **4.** EU-RTD; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain.

References

1. N Cristianini and J Shawe-Taylor, *Cambridge University Press*, (2000).
2. J Shawe-Taylor and N Cristianini, *Cambridge University Press*, (2004).
3. G Lanckriet, T De Bie *et al.*, *Bioinformatics*, **20**(16), 2626 (2004).
4. B Schölkopf, K Tsuda and J-P Vert, *MIT Press*, (2004).
5. W Stafford Noble, *Nature Biotechnology*, **24**(12), 1565 (2006).
6. T De Bie, L-C Tranchevent *et al.*, *Accepted for Bioinformatics*, (2007).
7. N Pochet, F De Smet *et al.*, *Bioinformatics*, **20**(17), 3185 (2004).
8. J M D Wheeler, B F Warren *et al.*, *Dis Colon Rectum* **45**(8), 1051 (2002).
9. J-P Machiels, C Sempoux *et al.*, *Ann Oncol*, in press (2007).
10. R A Irizarry, B Hobbs *et al.*, *Biostatistics*, **4**, 249 (2003).
11. V Vapnik, *Wiley, New York* (1998).
12. J Suykens and J Vandewalle, *Neural Processing Letters*, **9**(3), 293 (1999).
13. J Suykens, T Van Gestel *et al.*, *World Scientific Publishing Co., Pte Ltd. Singapore* (2002).
14. P Pavlidis, J Weston *et al.*, *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology*, 242 (2001).
15. H Deng, R Makizumi *et al.*, *Exp Cell Res*, **313**, 1033 (2007).
16. J A Nemeth, M L Cher *et al.*, *Clin Exp Metastasis*, **20**, 413 (2003).
17. D Raje, H Mukhtar *et al.*, *Dis Colon Rectum*, **50**, 1 (2007).
18. T Shimizu, S Tanaka *et al.*, *Oncology*, **59**, 229 (2000).
19. T Sasahira, T Sasaki and H Kuniyasu, *J Exp Clin Cancer Res*, **24**(1), 69 (2005).
20. T-D Kim, K-S Song *et al.*, *BMC Cancer*, **6**, 211 (2006).
21. K Zins, D Abraham *et al.*, *Cancer Res*, **67**(3), 1038 (2007).
22. S O Lee, J Y Chun *et al.*, *The Prostate*, **67**, 764 (2007).
23. A L Hallbeck, T M Walz and A Wasteson, *Bioscience Reports*, **21**(3), 325 (2001).
24. O Gevaert, F De Smet *et al.*, Bioinformatics, **22**(14), e184 (2006).