Chapter 9
A Genome-Wide Computational Study of Copy Number Variations: an Example on Ovarian Cancer

Anneleen Daemen, Olivier Gevaert, Karin Leunen, Vanessa Vanspauwen, Geneviève Michils, Eric Legius, Ignace Vergote, and Bart De Moor

Abstract *Motivation:* Knowledge about the molecular mechanisms involved in sporadic and hereditary ovarian tumorigenesis is lacking. Due to the hypothesis that BRCA related ovarian cancer follows distinct pathways in their carcinogenesis, array comparative genomic hybridization (array CGH) was performed in 8 sporadic and 5 BRCA1 mutated ovarian cancer patients to identify copy number variations.
Results: Chromosomal regions characterizing each group of sporadic and BRCA1 related ovarian cancer were gathered using recurrent hidden Markov Models (HMM). The differential regions were reduced to a subset of features for classification by integrating different univariate feature selection methods. Least Squares Support Vector Machines (LS-SVM), a supervised classification method, resulted in a leave-one-out accuracy of 84.6%, sensitivity of 100% and specificity of 75%.
Conclusion: The combination of recurrent HMMs for the detection of copy number alterations with LS-SVM classifiers offers a novel methodological approach for classification based on copy number alterations. Additionally, this approach limits the chromosomal regions that are necessary to distinguish sporadic from hereditary ovarian cancer.

9.1 Introduction

In cancers, many gains and losses of chromosomes and chromosomal segments have been described. These aberrations defined as regions of increased or decreased DNA copy number can be detected at high resolution using an array comparative genomic hybridization (array CGH) technology. This technique measures variations in DNA copy number within the entire genome of a disease sample compared to a normal sample [1]. This makes array CGH ideally suitable for a genome-wide identification

Anneleen Daemen and Olivier Gevaert
Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven, Leuven, Belgium

and localization of genetic alterations involved in human diseases. An overview of algorithms for array CGH data analysis is given in [2]. Segmentation approaches identify chromosomal regions of adjacent clones with the same mean log ratio. Disadvantages of these methods are that the segments that are gained or lost need to be determined in a further analysis and that results become unsatisfactory with high noise levels in the data. Therefore, segmentation and identification should be performed simultaneously because these two tasks can improve each other's performance. A popular method for combining them is the hidden Markov Model (HMM) with states defined as loss, neutral, one-gain and multiple-gain. Recently, this traditional procedure has been extended to a recurrent HMM in which a class of samples instead of individual samples is modeled by sharing information on copy number variations across multiple samples [4]. Here, we present a method to identify copy number alterations with the recurrent HMM which goes beyond the exploratory phase by using these alterations as features in a supervised classification setting and by validating these features biologically.

Because the exclusion of redundant and non-discriminatory features might avoid overfitting and identifies a smaller set of features able to distinguish good from bad, feature selection should be performed. To get rid of some of the arbitrariness with which a univariate feature selection method is chosen, different univariate test statistics were combined to suppress the false positive error rate [5]. For classification, we used the class of kernel methods which is powerful for pattern analysis. In recent years, these methods have become a standard tool in data analysis, computational statistics, and machine learning applications [6], [7]. Their rapid uptake in bioinformatics [8] is due to their reliability, accuracy and computational efficiency, which has been demonstrated in countless applications [9]. More specifically, as supervised classification algorithm we made use of the Least Squares Support Vector Machine (LS-SVM) which is an extension of the standard SVM and has been developed in our research group by Suykens *et al.* (1999), (2002) [10]-[11]. On high dimensional data, the LS-SVM is easier and faster to solve because the quadratic programming problem of the SVM is reduced to a set of linear equations.

We applied our method on ovarian cancer which is the fourth most common cause of cancer death and ranks as the most frequent cause of death from gynaecological malignancies among women in western countries [12]. In a total of 5-10% of epithelial ovarian carcinomas, a family history of breast and ovarian cancer is noted with germline mutations in the tumour suppressor genes BRCA1 or BRCA2 in most of them. A mutation of the BRCA1 gene cumulates the risk for ovarian carcinoma with 26-85% while a BRCA2 mutation increases the cumulative risk with 10% [14]. The knowledge of different copy number variations between both sporadic and hereditary groups may help to better understand tumorigenesis of these cancers. When applied to larger study groups, this method could result in a better comprehension of the different clinical behaviour of both groups, probably necessitating different treatment strategies.

The outline of this chapter is as follows. In section 9.2, we describe the data set and the array CGH technology used for the analysis as well as the recurrent HMM, the classifier and the feature selection method applied. In addition, the workflow of

our proposed methodology is given in detail together with the functional annotation analysis to validate agreement of the selected chromosomal regions with biology. We determine the gene sets from the Molecular Signatures Database (MSigDB) [13] that are enriched in the identified regions of copy number alteration. We describe our results on ovarian cancer in Section 9.3 and conclude in Section 9.4.

9.2 Materials and Methods

9.2.1 Patients and Data

The data for this study were collected from patients treated for ovarian cancer at the University Hospital of Leuven, Belgium. A distinction could be made between patients with a sporadic tumour and carriers of a mutation in the tumour suppressor genes BRCA1 or BRCA2. Both genes are involved in DNA damage repair and transcriptional regulation [15]. All tumour samples were collected at the time of primary surgery. Only patients with similar clinical characteristics were retained: eight sporadic and five BRCA1 mutated ovarian cancer patients. One patient with BRCA2 was excluded and none of the patients out of the sporadic group had a positive family history of breast and/or ovarian cancer. Array comparative genomic hybridization was performed using a 1Mb array CGH platform, version CGH-SANGER 3K 7 developed by the Flanders Institute for Biotechnology (VIB), Department of Microarray Facility, Leuven, Belgium.

9.2.2 Array Comparative Genomic Hybridization

Array comparative genomic hybridization (array CGH) is a high-throughput technique for measuring DNA copy number variations (CNV) within the entire genome of a disease sample relative to a normal sample [1]. In an array CGH experiment, total genomic DNA from tumour and normal reference cell populations are isolated and subsequently labeled with different fluorescent dyes before being hybridized to several thousands of probes on a glass slide. This allows to calculate the log ratios of the fluorescence intensities of the tumour to that of the normal reference DNA. Because the reference cell population is normal, an increase or decrease in the log intensity ratio indicates a DNA copy number variation in the genome of the tumour cells such that negative log ratios correspond to deletions (losses), positive log ratios to gains or amplifications and zero log ratios to neutral regions in which no change occurred.

9.2.3 Recurrent HMM

As was stated in the introduction, we will use a recurrent hidden Markov Model (HMM) proposed by Shah *et al.* (2007) for the identification of extended chromosomal regions of altered copy numbers labeled as gain or loss [4]. The goal of this model is to construct features that distinguish the sporadic from the BRCA1 related group and subsequently to use them in a classifier (see Section 9.2.4). Because of the sensitivity of traditional HMMs to outliers being measurement noise, mislabeling and copy number polymorphisms in the normal human population, a robust HMM was first proposed by Shah *et al.* (2006) which handles outliers and integrates prior knowledge about copy number polymorphisms into the analysis [16]. To further reduce the influence of various sources of noise on the detection of recurrent copy number alterations, Shah *et al.* (2007) extended the robust HMM to a multiple sample version in which array CGH experiments from a cohort of individuals are used to borrow statistical strength across samples instead of modeling each sample individually [4]. This makes even copy number alterations in a small number of adjacent clones reliable when shared across many samples.

In this study, a recurrent HMM is constructed on a chromosomal basis separately for the group of sporadic and the group of BRCA1 mutated ovarian cancer. Both HMMs result in chromosomal regions with genetic alterations characterizing sporadic samples and samples with a BRCA1 mutation, respectively. A differential region is defined as a chromosomal region which is gained/lost in one group while not being gained/lost in the other group.

9.2.4 Kernel Methods and Least Squares Support Vector Machines

The differential regions that result from the recurrent HMM are used as features in a classifier for which we chose kernel methods. These 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 [17]. This 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. In this manuscript, a linear kernel function was used.

An example of a kernel algorithm for supervised classification is the Support Vector Machine (SVM) developed by Vapnik [18] and others. 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. This kernel method also contains regularization which allows tackling the problem of overfitting. It has been shown that regularization seems to be very important when applying classification methods on high dimensional data [9]. A modified version of SVM, the Least Squares Support Vector Machine (LS-SVM), was developed by Suykens *et al.* (1999), (2002) [10]-[11]. 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.

9.2.5 Feature Selection

The choice of a feature selection technique is a widely discussed topic [19]. Lai *et al.* (2006) found that univariate gene selection, computationally simple and fast for high dimensional data, leads to good and stable performances across many cancer types and yields in many cases consistently better results than multivariate approaches [20]. Therefore, we will use a univariate method. Because no comparison of univariate gene selection techniques has been made across a sufficiently wide range of benchmark data sets and due to the dependency of the best performing technique on the data set used, Yang *et al.* (2005) proposed a method in which some of the arbitrariness with which univariate methods are chosen for high dimensional data is vanished [5]. This technique, called DEDS (Differential Expression via Distance Synthesis) is based on the integration of different test statistics via a distance synthesis scheme because features highly ranked simultaneously by multiple measures are more likely to be differential expressed than features highly ranked by a single measure. The statistical tests which were combined are ordinary fold changes, ordinary t-statistics, SAM-statistics and moderated t-statistics. The performance of DEDS is favorably comparable with the best individual statistic which is in practice often unknown and which depends on the data set used. Additionally, DEDS is not adversely affected by the worst performing statistic and achieves robustness properties which are lacked by the individual statistics. DEDS is available as a BioConductor package in R.

9.2.6 Proposed Methodology

Due to the limited number of samples, a leave-one-out (LOO) cross-validation strategy is applied. The 4 different steps that have to be accomplished in each LOO iteration are shown in Figure 9.1. After leaving out one sample, a recurrent HMM (see Sect. 9.2.3) is constructed in step 1 for both groups of sporadic and BRCA1 mutated ovarian cancer to determine the chromosomal regions with genetic alterations that characterize each group. Combining these regions results in the chromosomal regions that are differential between the remaining n-1 sporadic and BRCA1 mutated

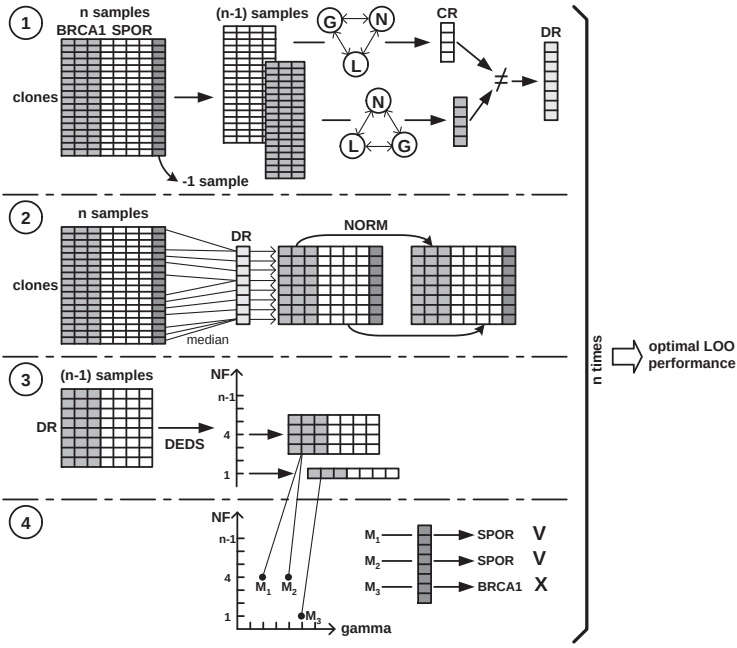

Fig. 9.1 Methodology consisting of 4 steps: step 1 - recurrent HMM; step 2 - conversion of clones to differential regions and normalization per sample; step 3 - feature selection using DEDS; step 4 - LS-SVM training and validation on left out sample (CR = Chromosomal Region; DR = Differential Region; NORM = Normalization; DEDS = Differential Expression via Distance Synthesis; NF = Number of Features)

samples. Because multiple clones can be located within each differential region, the clones need to be combined. This is done per sample in the second step by taking the median of the log ratios of the clones in each region. Afterwards, a standardization is performed per sample (i.e. meanshifting to 0 and autoscaling to 1) because the raw log ratios cannot be compared in absolute values between the samples. In step 3, DEDS determines which preprocessed log ratios, called features, best discriminate the n-1 samples (see Sect. 9.2.5). The number of included features is iteratively increased according to the obtained feature ranking without including more features than the number of samples on which the optimal number of features is determined [21]. This subset of features forms the input for classification in the last step (see Sect. 9.2.4). The LS-SVM contains a regularization parameter γ which, together with the number of features needs to be optimized. For all possible combinations of γ and number of features, an LS-SVM is built on the training set and validated on the left out sample. This is repeated n times such that each sample has been left out once. For the LS-SVM, a linear kernel function $k(x_k, x_l) = x_k^T x_l$ was chosen. An RBF kernel resulted in similar performances (data not shown).

9.2.7 Functional Annotation Analysis

To validate the selected chromosomal regions, gene set enrichment was performed as an indication for agreement with "known" biology. Two groups of gene sets as defined in the Molecular Signatures Database (MSigDB) were used: curated gene sets (i.e. sets of co-regulated genes from online pathway databases, publications in PubMed and knowledge of domain experts) and Gene Ontology (GO) gene sets (i.e. genes annotated by the same GO term) [13]. Using the HUGO gene nomenclature[1] [22], genes within the differential chromosomal regions were divided into 9 gene signatures, depending on the group (BRCA1 versus sporadic versus both) and CNV type (gain versus loss versus both). For each signature, the overlap was calculated between all gene sets and the signature and 5000 equally-sized signatures containing genes randomly selected from the genome. The corrected method of North *et al.* (2002) was used to calculate the empirical p-value for each gene set as $(r+1)/(n+1)$ with n the number of random signatures (i.e. 5000) and r the number of them with an equal or higher overlap with the gene set than obtained with the actual signature [23]. Only gene sets with r smaller than 10 (p-value < 0.002) were further investigated.

9.3 Results

Eight sporadic and five BRCA1 mutated ovarian cancer patients were included in this study and profiled using array CGH technology. Figure 9.2 gives an impression of array CGH data with which chromosomal regions that are different between 2 classes of samples can be identified. This figure shows an example of a recurrent amplification in BRCA1 patients which is not present in sporadic patients.

When applying the proposed methodology on this data set, CNVs in 11 chromosomal regions were sufficient to correctly classify 11 out of 13 samples. The LS-SVM had a LOO accuracy of 84.6%, a sensitivity of 100% (5/5) and a specificity of 75% (6/8).

Table 9.1 and Figure 9.3 show information on the 11 differential regions. Five regions are gained and 3 lost in BRCA1 mutated samples while the sporadic ovarian cancer patients are characterized by loss of 3 regions. A comparison of the 11 regions found in each of the 13 LOO iterations shows a limited variability in the selected regions. Table 9.1 also shows the number of LOO iterations in which each feature resulting from the complete data set is chosen which indicates stability of the 11 regions. The top 5 of features with the lowest p-value according to DEDS appeared in 8 to 11 of the 13 LOO iterations. Three less significantly features appeared in 4 LOO iterations.

Because we hypothesize that genes in the 11 chromosomal regions participate in processes that distinguish sporadic from hereditary ovarian cancer, a gene set

[1] http://www.genenames.org

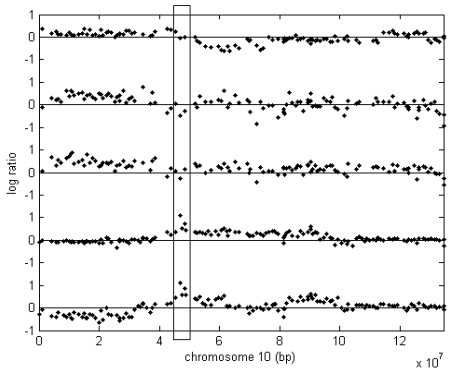

Fig. 9.2 Array CGH profile of chromosome 10 for 3 sporadic (top) and 2 BRCA1 mutated samples (bottom). The horizontal lines indicate the 0 log ratios for all samples. Both groups have a different profile within the first $3x10^7$ base pairs and an amplification indicated with the vertical box occurs within the BRCA1 mutated samples around $5x10^7$ base pairs.

Table 9.1 Chromosomal information on the 11 differential regions with the number of LOO iterations in which each of these regions was selected

Feature	Chromosome	Group	CNV type	Startbase	Stopbase	nb genes	nb LOO iter
1	13	BRCA1	loss	55423625	55550461	0	8
2	23	BRCA1	gain	3273880	7085387	5	11^{μ}
3	12	BRCA1	gain	101502349	101656438	0	9
4	4	BRCA1	gain	10384154	19905375	22	11^{ς}
5	4	sporadic	loss	4932958	8382645	24	10
6	3	BRCA1	gain	24167220	35751756	32	5
7	10	BRCA1	loss	4290650	17074128	66	7^{ς}
8	16	sporadic	loss	56587489	67418517	81	4
9	19	BRCA1	loss	12159479	13216789	39	4
10	6	BRCA1	gain	24267702	29367215	86	4^{ς}
11	16	sporadic	loss	70089429	75199166	36	6

$^{\mu}$ Approximate correlation with LOO: region 10-50% smaller in 2 LOO runs
$^{\varsigma}$ Approximate correlation with LOO: region 10-40% smaller in 1 LOO run

enrichment-based approach was followed (see Sect. 9.2.7). The most important gene sets enriched in the signatures are summarized below.

One of the components of the human SWI/SNF complex, regulating gene expression by remodeling nucleosomal structure in an ATP-dependent manner, is the gene BAF57 (a BRG1-associated factor). This gene mediates interaction with transcriptional activators or repressors and mutation of this gene has been found to be associated with a wide variety of tumours [24]. It is known that there is a direct interaction between BRG1- and hBRM-associated factors and the BRCA1 tumour suppressor protein. The human SWI/SNF complex affects cell growth and proliferation by interacting with tumour suppressor pathways and probably controlling them. Recent studies have shown the importance of complexes containing BAF57

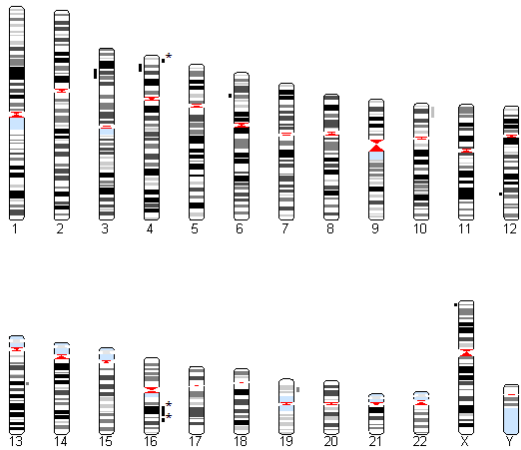

Fig. 9.3 BRCA1 - five gained regions (shown at the left of the chromosomes) and three lost regions (shown at the right in gray); sporadic - three lost regions (shown at the right indicated with the symbol *).

in transcriptional repression of tumour suppressor genes among which BRCA1. Wang and colleagues found 410 up-regulated and 469 down-regulated genes in cells with BAF57 re-expressed. Ten of the down-regulated genes (i.e. MED28, SUSD5, NCAPG, SLC4A7, MXRA5, MRS2, BST1, QDPR, LAP3, HS3ST1; p-value $< 2x10^{-4}$) were found in four of the five regions gained in the BRCA1 mutated samples.

Another gene set consisting of 96 genes down-regulated at any time point (1-24 hours) following treatment of mammary carcinoma cells with exogenous human growth hormone (hGH) [25] was significantly overrepresented in the regions gained in BRCA1 ovarian cancer with an overlap of 7 genes (GPLD1, HIST1H2BK, HS3ST1, SLC4A7, SLC17A1; p-value = $8x10^{-4}$).

Many HOX genes, a subset of the homeobox genes, were recently found to be aberrantly expressed in a variety of cancers among which breast, kidney and skin suggesting that these HOX genes contribute to the progression of tumours. The homeobox HOXA5 encodes a transcriptional factor with an important role in embryogenesis, hematopoiesis and tumorigenesis. In human, it has been shown that HOXA5 mRNA levels are markedly reduced or even lost in more than 60% of breast cancer cell lines and primary breast carcinoma cells. This suggests that HOXA5 may act as a tumour suppressor gene in breast cells which makes loss of expression of this gene an important step in tumorigenesis [26]. Six genes, normally up-regulated in HOXA5-induced cells (with HOXA5 being a positive regulator), were found to be lost in BRCA1 ovarian cancer (ZNF44, DCLRE1C, ZNF136, KIN, JUNB, IER2; p-value = $8x10^{-4}$).

Tumour necrosis factor alpha (TNFα) is a proinflammatory cytokine with important roles in control of immune and inflammatory responses as well as cell cycle proliferation and apoptosis [27]. Of the genes up-regulated in TNFα-induced HeLa cells, four were found in 2 regions lost in BRCA1 ovarian cancer (IER2, PRDX2, JUNB, GDI2; p-value = $1.4x10^{-3}$).

Three highly related Myb transcription factors (i.e. A-Myb, B-Myb and c-Myb) are expressed in vertebrates. The c-Myb gene, the proto-oncogene progenitor of the v-myb oncogene, is highly expressed in a.o. pancreatic, colon and breast tumours and his expression correlates with proliferation. A functional c-Myb protein is required for normal hematopoiesis. The A-Myb gene is expressed in a subset of the cells that expresses c-Myb [28]. Sporadic ovarian cancer is characterized by a loss of 9 genes activated by A-Myb or c-Myb genes (ATP6V0D1, MMP15, RRAD, S100P, E2F4, CTCF, PSMD7, CDH1, NFATC3; p-value = $6x10^{-4}$).

9.4 Conclusion

In this manuscript, a new methodology is proposed in which copy number variations resulting from array CGH are transformed into features for classification purpose. This general method which is independent of cancer site allows to find a small set of chromosomal regions for distinguishing two classes of patients and to biologically validating them. It can also result in clinically relevant models based on a limited set of features. As increasing amounts of array CGH data become available, there is a need for algorithms to identify recurrent gains and losses based on statistically sound methods and to use them for classification. A large number of approaches for the analysis of array CGH data have already been proposed recently, ranging from mixture models and HMMs to wavelets and genetic algorithms [2]. However, most cancer studies that gather array CGH data only apply methods for exploratory analysis. Often a fixed threshold is used for defining gains and losses making these studies less robust against systematic changes in the baseline copy number measurements between samples [29]. A HMM on the contrary is a probabilistic method that can handle the uncertainty in the data in a formal way compared to deterministic algorithms. This makes the HMM more robust against outliers such as measurement noise and wrong recordings of locations of clones. Moreover, we used a special variant of HMM able to capture recurrent copy number alterations by coupling the HMMs of individual samples. This makes weak copy number alterations but shared across many samples reliable features. In our setup we used this property by first modeling the copy number variations in the group of sporadic and BRCA1 mutated patients separately. Subsequently, the alterations that were different between these two groups were used as features in an LS-SVM for classification. In our opinion this is one step further compared to many other studies that only perform an exploratory analysis.

The stability of the regions selected in each of the LOO iterations strengthens our confidence that the chromosomal regions found with our methodology are robust.

Two of the regions lacking genes with an annotated HUGO symbol seem uninteresting at first sight. However, recent research findings on 1% of the genome indicated that 93% of the bases are transcribed, increasing the importance of non-protein-coding RNA [30]. The remaining 9 regions were validated biologically using a gene set enrichment-based approach. Keep in mind that, because the number of features is minimized, one can expect that biological validation using pathways may fail because not all genes belonging to a certain pathway may be needed in a classification setting. In our subset the genes BAF57 and HOXA5 seemed to be correlated with hereditary ovarian cancer, whereas loss of the v-myb oncogene seemed more characteristic for the sporadic group.

Acknowledgements

AD is research assistant of the Fund for Scientific Research - Flanders (FWO-Vlaanderen). BDM is a full professor at the Katholieke Universiteit Leuven, Belgium. This work is partially supported by: **1.** Research Council KUL: GOA AM-BioRICS, CoE EF/05/007 SymBioSys, PROMETA, several PhD/postdoc & fellow grants. **2.** Flemish Government: **a.** FWO: PhD/postdoc grants, projects G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0318.05 (subfunctionalization), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); **b.** IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame, SBO-MoKa, TBM-Endometriosis. **3.** Belgian Federal Science Policy Office: IUAP P6/25 (BioMaG-Net, Bioinformatics and Modeling: from Genomes to Networks, 2007-2011). **4.** EU-RTD: ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain, FP6-STREP Strokemap.

References

1. Pinkel, D., Albertson, D.G.: Array comparative genomic hybridization and its applications in cancer. Nature Genetics **37**(Suppl.) (2005) 11–17
2. Lai, W.R., Johnson, M.D. *et al.*: Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics **21**(19) (2005) 3763–3770
3. Guha, S., Li, Y. *et al.*: Bayesian hidden markov modeling of array CGH data. Harvard University Biostatistics Working Paper Series, Working paper 24 (October 2006)
4. Shah, S., Lam, W.L. *et al.*: Modeling recurrent DNA copy number alterations in array CGH data. Bioinformatics **23** (2007) i450–i458
5. Yang, Y.H., Xiao, Y. *et al.*: Identifying differentially expressed genes from microarray experiments via statistic synthesis. Bioinformatics **21**(7) (2005) 1084–1093
6. Cristianini, N., Shawe-Taylor, J.: An introduction to Support Vector Machines and other kernel-based learning methods. Cambridge University Press, Cambridge (2000)
7. Shawe-Taylor, J., Cristianini, N.: Kernel methods for pattern analysis. Cambridge University Press, Cambridge (2004)

8. Bhaskar, H., Hoyle, D.C. *et al.*: Machine learning in bioinformatics: A brief survey and recommendations for practitioners. Computers in Biology and Medicine **36**(10) (2006) 1104–1125
9. Pochet, N., De Smet, F. *et al.*: Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. Bioinformatics **20** (2004) 3185–3195
10. Suykens, J., Vandewalle, J.: Least Squares Support Vector Machine classifiers. Neural Processing Letters **9** (1999) 293–300
11. Suykens, J.A.K., Van Gestel, T. *et al.*: Least Squares Support Vector Machines. World Scientific, Singapore (2002)
12. Gajewski, W., Legare, R.D.: Ovarian cancer. Surgical Oncology Clinics of North America **7** (1998) 317–333
13. Subramanian, A., Tamayo, P. *et al.*: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences **102**(43) (2005) 15545–15550
14. Burke, W., Daly, M. *et al.*: Recommendations for follow-up care of individuals with an inherited predisposition to cancer. II. BRCA1 and BRCA2. Cancer Genetics Studies Consortium. Journal of the American Medical Association **277** (1997) 997–1003
15. Starita, L.M., Parvin, J.D.: The multiple nuclear functions of BRCA1: trancription, ubiquitination and DNA repair. Current Opinion in Cell Biology **15**(3) (2003) 345–350.
16. Shah, S., Xuan, X. *et al.*: Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics **22**(14) (2006) e431–e439
17. Schölkopf, B., Tsuda, K. *et al.*: Kernel methods in computational biology. MIT Press, United States (2004)
18. Vapnik V.: Statistical Learning Theory. Wiley, New York (1998)
19. Saeys, Y., Inza, I. *et al.*: A review of feature selection techniques in bioinformatics. Bioinformatics **23**(19) (2007) 2507–2517
20. Lai, C., Reinders, M.J.T. *et al.*: A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. BMC Bioinformatics **7** (2006) 235–244
21. Li, W., Yang, Y.: How many genes are needed for a discriminant microarray data analysis. In Methods of Microarray Data Analysis, eds = Lin, S.M. and Johnson, K.F., Kluwer Academic (2002) 137–150
22. Wain, H.M., Bruford, E.A. *et al.*: Guidelines for human gene nomenclature. Genomics **79**(4) (2002) 464–470
23. North, B.V., Curtis, D. *et al.*: A note on the calculation of empirical p values from Monte Carlo procedures. American Journal of Human Genetics **71** (2002) 439–441
24. Wang, L., Baiocchi, R.A. *et al.*: The BRG1- and hBRM-associated factor BAF57 induces apoptosis by stimulating expression of the cylindromatosis tumor suppressor gene. Molecular and Cellular Biology **25**(18) (2005) 7953–7965
25. Xu, X.Q., Emerald, S. *et al.*: Gene expression profiling to identify oncogenic determinants of autocrine human growth hormone in human mammary carcinoma. The Journal of Biological Chemistry **280**(25) (2005) 23987–24003
26. Chen, H., Rubin, E. *et al.*: Identification of transcriptional targets of HOXA5. The Journal of Biological Chemistry **280**(19) (2005) 19373–19380
27. Zhou, A., Scoggin, S. *et al.*: Identification of NF-κB-regulated genes induced by TNFα utilizing expression profiling and RNA interference. Oncogene **22** (2003) 2054–2064
28. Lei, W., Rushton, J.J. *et al.*: Positive and negative determinants of target gene specificity in Myb transcription factors. The Journal of Biological Chemistry **279**(28) (2004) 29519–29527
29. Klijn, C., Holstege, H. *et al.*: Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. Nucleic Acids Research **36**(2) e13
30. The ENCODE Project Consortium: Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature **447** 799–816