Classification of Sporadic and BRCA1 Ovarian Cancer Based on a Genome-Wide Study of Copy Number Variations

Anneleen Daemen[1,⋆], Olivier Gevaert[1], Karin Leunen[2], Vanessa Vanspauwen[3], Geneviève Michils[3], Eric Legius[3], Ignace Vergote[2], and Bart De Moor[1]

[1] Department of Electrical Engineering (ESAT)
Katholieke Universiteit Leuven, Leuven, Belgium
[2] Department of Obstetrics and Gynaecology, Division of Gynaecologic Oncology
Multidisciplinary Breast Centre, University Hospital Leuven, Leuven, Belgium
[3] Department of Human Genetics, University Hospital Leuven, Leuven, Belgium

Abstract. *Motivation:* Although studies have shown that genetic alterations are causally involved in numerous human diseases, still not much is known about the molecular mechanisms involved in sporadic and hereditary ovarian tumorigenesis.

Methods: Array comparative genomic hybridization (array CGH) was performed in 8 sporadic and 5 BRCA1 related ovarian cancer patients.

Results: Chromosomal regions characterizing each group of sporadic and BRCA1 related ovarian cancer were gathered using multiple sample hidden Markov Models (HMM). The differential regions were used as features for classification. 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 multiple sample HMMs for the detection of copy number alterations with LS-SVM classifiers offers an improved methodological approach for classification based on copy number alterations. Additionally, this approach limits the chromosomal regions necessary to distinguish sporadic from hereditary ovarian cancer.

1 Introduction

Many defects in human development leading to e.g. cancer and mental retardation are due to gains and losses of chromosomes and chromosomal segments. These aberrations defined as regions of increased or decreased DNA copy number can be detected 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 and

⋆ Corresponding author.

I. Lovrek, R.J. Howlett, and L.C. Jain (Eds.): KES 2008, Part II, LNAI 5178, pp. 165–172, 2008.

localization of genetic alterations involved in human diseases. An overview of algorithms for array CGH data analysis is given in [2]. Segmentation approaches identify adjacent clones with a same mean log ratio. These methods have as disadvantages that a further analysis is needed to determine the segments that are gained or lost and that results become unsatisfactory with high noise levels in the data. Therefore, segmentation and classification should be performed simultaneously because these two tasks can improve each other's performance. A popular method to combine them is the hidden Markov Model (HMM) with states defined as loss, neutral, one-gain and multiple-gain. Recently, this traditional procedure has been exploited to a multiple sample HMM in which a class of samples instead of individual samples is modeled by sharing information on copy number variations across multiple samples [3]. Here, we present a method to identify copy number alterations with the multiple sample HMM and that goes beyond the exploratory phase by using these alterations as features in a supervised classification setting.

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 [4]. Their rapid uptake in bioinformatics is due to their reliability, accuracy and computational efficiency, which has been demonstrated in countless applications [5]. More specifically, as supervised classification algorithm we made use of the Least Squares Support Vector Machine (LS-SVM) which is an extension of the more regular SVM and has been developed in our research group by Suykens et al [6]. On high dimensional data, the LS-SVM is easier and faster compared to the SVM.

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 [7]. 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. 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% [8].

The outline of this article is as follows. In section 2, we describe the data set and the array CGH technology used for the analysis as well as the multiple sample HMM, the classifier and the feature selection method applied. In addition, the workflow of our proposed methodology is given in detail. In Section 3, we describe our results on ovarian cancer and finally, conclusions and future research directions are given in Section 4.

2 Materials and Methods

2.1 Patients and Data

Data from patients treated for ovarian cancer at the University Hospital of Leuven, Belgium were collected for participation at this study. All tumour samples were collected at the time of primary surgery. Only patients with similar clinical

characteristics were retained: eight sporadic and five BRCA1 related 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.

2.2 Array Comparative Genomic Hybridization

Array comparative genomic hybridization (array CGH) is a high-throughput technique for measuring variations in DNA copy number 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, different fluorescently labeled and 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.

2.3 Multiple Sample HMM

As was stated in the introduction, we will use a multiple sample hidden Markov Model (HMM) proposed by Shah et al [3] for the identification of chromosomal aberrations and to detect extended chromosomal regions of altered copy numbers labeled as gain or loss. 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 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 [9] which handles outliers and integrates prior knowledge about copy number polymorphisms into the analysis. To further reduce the influence of various sources of noise on the detection of recurrent copy number alterations, Shah et al 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 [3]. This makes even copy number alterations in a small number of adjacent clones reliable when shared across many samples.

In this study, a multiple sample HMM is constructed on a chromosomal basis separately for the group of sporadic and the group of BRCA1 related ovarian cancer. Both HMMs result in chromosomal regions with genetic alterations characterizing sporadic and BRCA1 related samples, 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.

2.4 Kernel Methods and Least Squares Support Vector Machines

The differential regions we just constructed 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 [10]. 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 [11] 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. We have shown that regularization seems to be very important when applying classification methods on high dimensional data [5]. A modified version of SVM, the Least Squares Support Vector Machine (LS-SVM), was developed by Suykens et al [6]. 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.

2.5 Feature Selection

Because it has been shown in [13] that univariate gene selection methods lead to good and stable performances across many cancer types and yield in many cases consistently better results than multivariate approaches, we used the method DEDS (Differential Expression via Distance Synthesis) [14]. This technique 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. DEDS is available as a BioConductor package in R.

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 1. After leaving out one sample, a multiple

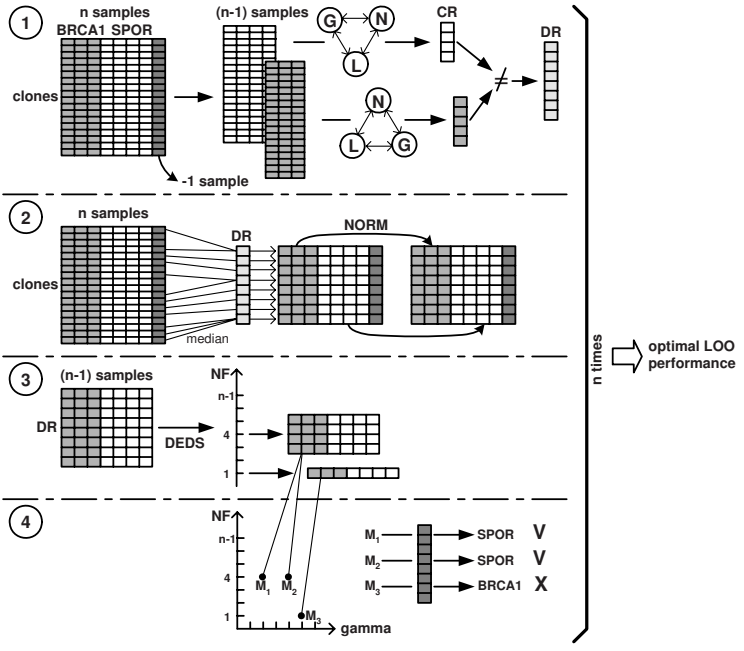

Fig. 1. Methodology consisting of 4 steps: step 1 - multiple sample HMM; step 2 - conversion of clones to differential regions and normalization per sample; step 3 - feature selection using DEDS; step 4 - LS-SVM training and validation on left out sample (CR = Chromosomal Region; DR = Differential Region;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

sample HMM (see Sect. 2.3) is constructed in step 1 for both groups of sporadic and BRCA1 related 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 related 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. 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 [15]. This subset of features forms the input for classification in the last step (see Sect. 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).

3 Results

Our data set contains 8 sporadic and 5 BRCA1 related ovarian cancer patients. The array CGH data of chromosome 10 is shown in Figure 2 for 3 sporadic and 2 BRCA1 related samples. Both groups have a different profile within the first $3x10^7$ base pairs and an amplification occurs within the BRCA1 related samples around $5x10^7$ base pairs. When applying the proposed methodology on this data set, 11 out of 13 samples could be classified correctly using measured copy number changes in only 11 differential regions. The LS-SVM had a LOO accuracy of 84.6%, a sensitivity of 100% (5/5) and a specificity of 75% (6/8).

A comparison of the 11 differential regions found in each of the 13 LOO iterations shows a limited variability in the selected regions. Table 1 shows the number of LOO iterations in which the same features were chosen as the ones most differentially between all 13 samples. 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. These results strengthen our confidence that the chromosomal regions found with our methodology are robust and we hypothesize that genes in these regions participate in processes that distinguish sporadic from hereditary ovarian cancer.

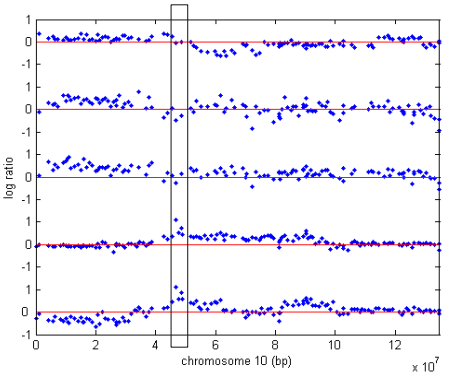

Fig. 2. Array CGH profile of chromosome 10 for 3 sporadic (top) and 2 BRCA1 related samples (bottom). The horizontal lines indicate the 0 log ratios for all samples. The vertical box indicates the amplification for the 2 BRCA1 related samples.

Table 1. Number of LOO iterations in which each of the 11 chromosomal regions was selected

Feature	1	2	3	4	5	6	7	8	9	10	11
Nb LOO iterations	8	11	9	11	10	5	7	4	4	4	6

4 Conclusion and Future Work

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 applicable to all types of cancer allows to find a small set of chromosomal regions for distinguishing two classes of patients and may further improve biological validation. It can also result in clinical relevant models for a simpler prediction based on a limited set of features. As increasing amounts of array CGH data become available, there is a need for algorithms to identify gains and losses statistically, rather than merely detect trends in the data. A large number of approaches for the analysis of array CGH data has already been proposed recently, ranging from mixture models and HMMs to wavelets and genetic algorithms [2]. However, most studies of cancer with gathered array CGH data apply less sophisticated methods for an exploratory analysis. Such studies apply a fixed threshold for defining gains and losses. A HMM on the contrary is a more intelligent way to detect copy number alterations in the genome of each sample by exploiting the spatial correlation between clones within an aberrated region. This makes the HMM also more robust against outliers such as measurement noise and wrongly recordings of locations of clones. Secondly, a robust HMM improves the reliability of the found chromosomal regions by taking into account copy number polymorphisms occurring in the normal human population. Thirdly, the multiple sample HMM improves the ability of detecting aberrations common for one group by borrowing strength across samples instead of modeling each sample individually. This makes also copy number alterations in a small number of adjacent clones reliable when shared across many samples and may prevent the loss of these possibly important biological features. Subsequently, the aberrations that are different between the group of sporadic and BRCA1 related samples are considered as features characterizing these samples. Finally, classification is performed to determine a small set of chromosomal regions that can distinguish sporadic from BRCA1 related ovarian cancer.

In the near future, an extensive study of the 11 differential regions may result in an increased knowledge on genes and pathways involved in sporadic versus hereditary ovarian cancer. Furthermore, we will analyze new patients with an in-house developed array CGH technology with a higher resolution to strengthen our hypotheses and to refine the found regions of genetic alterations possibly involved in ovarian cancer.

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 AMBioRICS, CoE EF/05/007 SymBioSys, PROMETA. **2.** Flemish Government: **a.** FWO projects G.0241.04, G.0499.04, G.0318.05, G.0302.07; **b.** IWT:

GBOU-McKnow-E, GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame, SBO-MoKa, TBM-Endometriosis. **3.** Belgian Federal Science Policy Office: IUAP P6/25. **4.** EU-RTD: ERNSI, 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. Nat. Genet. 37(Suppl.), 11–17 (2005)
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), 3763–3770 (2005)
3. Shah, S., Lam, W.L., et al.: Modeling recurrent DNA copy number alterations in array CGH data. Bioinformatics 23, i450–i458 (2007)
4. Shawe-Taylor, J., Cristianini, N.: Kernel methods for pattern analysis. Cambridge University Press, Cambridge (2004)
5. Pochet, N., De Smet, F., et al.: Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. Bioinformatics 20, 3185–3195 (2004)
6. Suykens, J.A.K., Van Gestel, T., et al.: Least Squares Support Vector Machines. World Scientific, Singapore (2002)
7. Gajewski, W., Legare, R.D.: Ovarian cancer. Surg. Oncol. Clin. N. Am. 7, 317–333 (1998)
8. 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. J. Am. Med. Assoc. 277, 997–1003 (1997)
9. Shah, S., Xuan, X., et al.: Integrating copy number polymorphisms into array CGH analysis using a robust HMM. Bioinformatics 22(14), e431–e439 (2006)
10. Schölkopf, B., Tsuda, K., et al.: Kernel methods in computational biology. MIT Press, United States (2004)
11. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
12. Saeys, Y., Inza, I., et al.: A review of feature selection techniques in bioinformatics. Bioinformatics 23(19), 2507–2517 (2007)
13. 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, 235–244 (2006)
14. Yang, Y.H., Xiao, Y., et al.: Identifying differentially expressed genes from microarray experiments via statistic synthesis. Bioinformatics 21(7), 1084–1093 (2005)
15. Li, W., Yang, Y.: How many genes are needed for a discriminant microarray data analysis. In: Lin, S.M., Johnson, K.F. (eds.) Methods of Microarray Data Analysis, pp. 137–150. Kluwer Academic, Dordrecht (2002)