SUPERVISED CLASSIFICATION OF ARRAY CGH DATA WITH HMM-BASED FEATURE SELECTION

ANNELEEN DAEMEN[1]*, OLIVIER GEVAERT[1], KARIN LEUNEN[2], ERIC LEGIUS[3], IGNACE VERGOTE[2] AND BART DE MOOR[1]

[1] *Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven*
SCD-SISTA (BIOI), Kasteelpark Arenberg 10 - bus 2446,
B-3001 Leuven (Heverlee), Belgium
[2] *Department of Obstetrics and Gynaecology, Division of Gynaecologic Oncology*
Multidisciplinary Breast Centre, University Hospital Leuven,
Herestraat 49 - bus 7003 , B-3000 Leuven, Belgium
[3] *Department of Human Genetics, O&N I, University Hospital Leuven,*
Herestraat 49 - bus 603, B-3000 Leuven, Belgium

Motivation: For different tumour types, extended knowledge about the molecular mechanisms involved in tumorigenesis is lacking. Looking for copy number variations (CNV) by Comparative Genomic Hybridization (CGH) can help however to determine key elements in this tumorigenesis. As genome-wide array CGH gives the opportunity to evaluate CNV at high resolution, this leads to huge amount of data, necessitating adequate mathematical methods to carefully select and interpret these data.
Results: Two groups of patients differing in cancer subtype were defined in two publicly available array CGH data sets as well as in our own data set on ovarian cancer. Chromosomal regions characterizing each group of patients 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. Weighted Least Squares Support Vector Machines (LS-SVM), a supervised classification method which takes unbalancedness of data sets into account, resulted in leave-one-out or 10-fold cross-validation accuracies ranging from 88 to 95.5%.
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 classify patients according to cancer subtype.

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

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 (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 localization of genetic alterations involved in human diseases. An overview of algorithms for array CGH data analysis is given by Lai *et al.*[2] Segmentation approaches divide each single-sample signal into regions of constant copy number, called segments. Subsequent classification labels the segments as gain or loss. A popular method combining both tasks 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 an individual sample 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 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.

For classification, we started from the class of kernel methods which is powerful for pattern analysis[4]. Their rapid uptake in bioinformatics is due to their reliability, accuracy, and computational efficiency[5]. More specifically, we made use of the weighted Least Squares Support Vector Machine (LS-SVM), taking the unbalancedness of data sets into account[6−7].

We applied our method on two publicly available data sets of lung cancer and oral squamous cell carcinoma, and subsequently on our data set of ovarian cancer. The knowledge of different copy number variations between specific groups of patients 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 article is as follows. In section 2, we describe the data sets, the applied methods for segmentation, feature selection, and classification, the workflow of our proposed methodology, as well as the functional annotation analysis for biological validation. We describe our results on the different data sets in Section 3 and conclude in Section 4.

2. Materials and Methods

2.1. *Data Sets*

In the study of Snijders *et al.*[8] a genome-wide analysis of copy number aberrations was carried out in 89 patients with oral squamous cell carcinoma, using the HumArray2.0 genome scanning array with 2464 clones from the University of California San Francisco. Missing copy number values were imputed unsupervisedly using the k-nearest neighbours method with k set to 15[9]. The final data set contained 2056 unique clones. Patients were subdivided according to *TP53* mutation with 59 of them being wildtype and 16 having a mutation for the *TP53* gene. The mutation status was unknown for the remaining 14 patients.

Garnis and colleagues[10] measured genome copy number profiles for 28 commonly used non-small cell lung carcinoma (NSCLC) cell lines across the whole genome using the submegabase resolution tiling array, consisting of 32433 BAC clones. 29781 unique clones remained after preprocessing. Twenty-two of the cell lines belonged to one of the 2 main subgroups of NSCLC: adenocarcinoma (13) and squamous cell carcinoma (9).

Our data were collected from patients treated for ovarian cancer at the University Hospital of Leuven, Belgium. Eight patients had a sporadic tumour without a positive family history for breast and/or ovarian cancer, while five patients were carrier of a mutation in the tumour suppressor gene *BRCA1*, involved in DNA damage repair and transcriptional regulation[11]. Array comparative genomic hybridization was performed using a 1Mb array CGH platform, version CGH-SANGER 3K 7, containing 3593 clones and 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 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 calculating 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. *Recurrent HMM*

As was stated in the introduction, we will use a recurrent hidden Markov Model (HMM) proposed by Shah *et al.*[3] for the identification of extended chromosomal regions of altered copy numbers, labeled as gain or loss. The goal of this model is to construct features that distinguish two groups of patients and subsequently to use them in a classifier (see Section 2.4). Due to sensitivity of traditional HMMs to outliers being measurement noise, mislabeling, and copy number polymorphisms within the normal human population, a robust HMM was proposed by Shah *et al.*[12] and extended 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 reduces the influence of various sources of noise on the detection of recurrent copy number alterations and 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 for each group of patients separately, resulting in regions with the probabilities of genetic alterations across these patients. A clone was labeled as gain or loss when its probability of occurring was more than 70%. An important parameter of the recurrent HMM that needs to be set is the recurrent variable ϵ. This variable represents the probability with which samples will get reflected in the recurrent CNV profile - accounting for sample-specific random effects - and is inversely proportional to the sparseness of the profile. As a trade-off between sparseness and false positive rate, ϵ was set to 0.8, although ϵ=0.7 is already sufficient for small data sets with less heterogeneity within each group of samples (i.e. our data set with 8 and 5 samples, respectively). A differential region was 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 Weighted 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 by representing data entities through a set of pairwise comparisons called the kernel matrix[13]. This matrix can be geometrically expressed as a transformation of each data point x to a high dimensional feature space with the mapping function $\Phi(x)$. An explicit representation of $\Phi(x)$ is not needed when 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.

A kernel algorithm for supervised classification which can tackle high dimensional data and contains regularization, is the Support Vector Machine (SVM) developed by Vapnik[14] and others. Given a training set $\{x_k, y_k\}_{k=1}^{N}$ of N samples with feature vectors $x_k \in \mathbb{R}^n$ and output labels $y_k \in \{-1, +1\}$, the SVM forms a linear discriminant boundary $y(x) = \text{sign}[w^T \Phi(x) + b]$ in the 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.*[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.

In many two-class problems, data sets are skewed in favour of one class such that the contribution of false negative and false positive errors to the performance assessment criterion are not balanced. We therefore used a weighted LS-SVM in which a different weight ζ_k is given to positive and negative samples in order to account for the unbalancedness in the data set[7]. The constrained optimization problem for the weighted LS-SVM has the following form:

$$\min_{w,b,e} J_P(w, e) = \frac{1}{2} w^T w + \gamma \frac{1}{2} \sum_{k=1}^{N} \zeta_k e_k^2$$

with

$$\zeta_k = \begin{cases} \frac{N}{2N_P} \text{ if } y_k = +1 \\ \frac{N}{2N_N} \text{ if } y_k = -1 \end{cases}$$

and N_P and N_N representing the number of positive and negative samples, respectively. Variables e_k represent the error variables, tolerating misclassifications in case of overlapping distributions, and γ the regularization parameter which allows tackling the problem of overfitting.

2.5. *Feature Selection*

Because it has been shown by Lai and colleagues[15] 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)[16]. 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 combined are ordinary fold changes, ordinary t-statistics, SAM-statistics and moderated t-statistics (i.e. B-statistics[17]). DEDS is available as a BioConductor package in R.

2.6. *Proposed Methodology*

For the data set of Snijders, a 10-fold cross-validation (CV) strategy was applied, while a leave-one-out (LOO) cross-validation strategy was chosen for the other, rather small data sets. The 4 different steps that have to be accomplished in each CV iteration are shown in Figure 1. After leaving out m samples (i.e. N/10 for 10-fold CV; 1 for LOO), a recurrent HMM (see Sect. 2.3) is constructed for both groups in step 1 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-m samples from both groups. 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-m 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[18]. This subset of features forms the input for classification in the last step (see Sect. 2.4). For all possible combinations of γ and number of features, an LS-SVM is built on the training set and validated on the m left out samples. This is repeated until each sample has been left out once. The parameter combination is chosen corresponding to the highest AUC (area under the

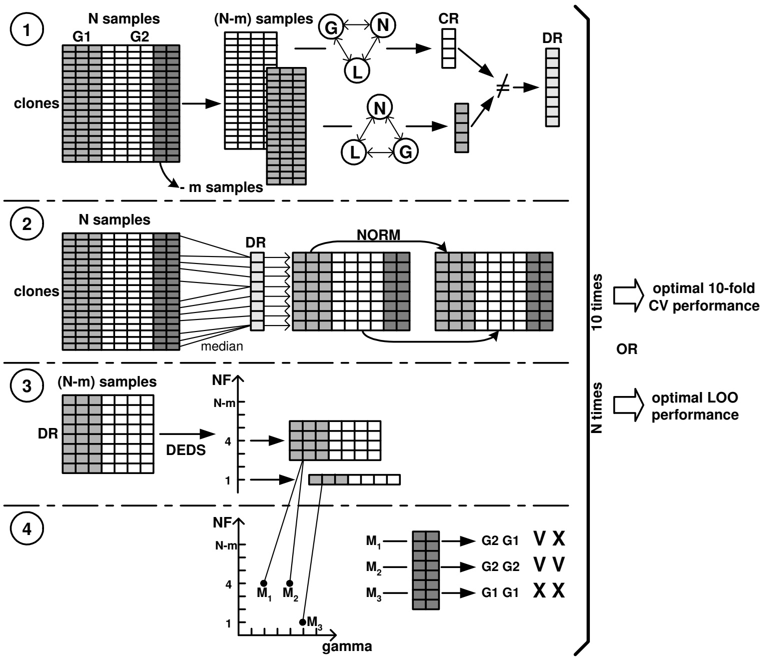

Figure 1. The methodology, applied in a 10-fold or LOO CV setting, consists of 4 steps: step 1 - recurrent HMMs applied on both groups of samples, resulting in differential regions; step 2 - conversion of clones to differential regions, and normalization (performed for each sample separately); step 3 - feature selection using DEDS; step 4 - LS-SVM training on 2-dimensional grid for regularization parameter γ and number of features, and validation on m left out samples (G1 = group 1; G2 = group 2; CR = Chromosomal Region; DR = Differential Region; NORM = Normalization; DEDS = Differential Expression via Distance Synthesis; NF = Number of Features)

ROC curve). When multiple such combinations, the combination with the lowest balanced error rate and an as high as possible sum of sensitivity and specificity is considered as optimal. For the LS-SVM, a linear kernel function $k(x_k, x_l) = x_k^T x_l$ was chosen. A kernel function that accounts for the correlation between neighbouring clones[19] did not lead to a better performance because the HMM already takes the properties of CGH data into account.

2.7. *Functional Annotation Analysis*

To validate the selected chromosomal regions, gene set enrichment was performed as an indication for agreement with “known” biology. Curated gene sets as defined in the Molecular Signatures Database (MSigDB) (i.e. sets of co-regulated genes from online pathway databases, publications in PubMed and knowledge of domain experts) were used[20]. Using the HUGO gene nomenclature[21], genes within the differential chromosomal regions were divided into 9 gene signatures, depending on the group (G1 versus G2 versus both) and CNV type (gain versus loss versus both). For each signature, the overlap was calculated between all gene sets and the signature as well as 5000 equally-sized signatures containing genes randomly selected from the genome. The corrected method of North *et al.*[22] 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. Only gene sets with r smaller than 10 (p-value < 0.002) were further investigated.

3. Results

We applied our method on two publicly available array CGH data sets, as well as on our own data set of ovarian cancer for the prediction of cancer subtype.

The data set of Snijders was used to distinguish patients being wildtype for the *TP53* gene from those having a mutation for *TP53*. 55 of the 59 wildtype patients and 11 of the 16 *TP53*-mutated patients could be classified correctly based on 10 differential regions, none of them containing *TP53* located on chromosome 17 (see Table 1). The wildtype group was characterized by 3 gained regions on chromosome arms 2p, 11q, and 18q, and 1 lost region on 12p. The mutated group was distinguishable with 1 gained region on 10p and 5 lost regions on chromosome arms 8p, 11q, and 18q. Five of these regions were selected in 8 to 10 of the 10 CV iterations while the remaining 5 regions appeared in 1 to 5 CV iterations. Loss of chromosome arms 8p, 11q, and 18q in *TP53*-mutated samples was mentioned in the original manuscript as well.

When applying our methodology on the data set of Garnis to distinguish adenocarcinomas from squamous cell carcinomas, 21 out of the 22 cell lines could be classified correctly based on 8 regions on chromosome arms 1q, 2p, 3q, 5p, 7q, 11p, 13q, and 20p, all lost in adenocarcinoma (see Table 1). Five regions were selected in all LOO iterations, the regions on 7q and 11p in 19 and 15 LOO iterations, respectively while the lowest ranked region according to DEDS appeared in 9 of the 22 LOO iterations. The

Table 1. Performances of the three considered data sets

Data set	Nb regions	Accuracy$^\mu$	Sensitivity	Specificity	AUC$^\mu$
Snijders[8]	10	88 (66/75)	93.2 (55/59)	68.8 (11/16)	0.840
Garnis[10]	8	95.5 (21/22)	92.3 (12/13)	100 (9/9)	0.983
own data	11	92.3 (12/13)	100 (5/5)	87.5 (7/8)	0.875

$^\mu$Represents 10-fold CV performance (Snijders) or LOO performance (own data, Garnis)

chromosome arms 2p and 13q were also found by Garnis and colleagues to be frequently deleted in adenocarcinoma while being amplified in the cell lines of the squamous cell carcinomas.

Eight sporadic and five *BRCA1* mutated ovarian cancer patients were included and profiled using array CGH technology. When applying the proposed methodology on this data set, CNVs in 11 chromosomal regions were sufficient to correctly classify 12 out of 13 samples (see Table 1). Five regions on chromosome arms 3p, 4p, 6p, 12q, and Xp were gained and 3 regions on 10p, 13q, and 19q were lost in *BRCA1* mutated samples. The sporadic ovarian cancer patients were characterized by loss of 3 regions on 4p and 16. The top 5 of features with the lowest p-value according to DEDS appeared in 8 to 11 of the 13 LOO iterations while the lower ranked regions were selected in 4 to 7 LOO iterations.

Because we hypothesize that genes in the chromosomal regions participate in processes that distinguish subtypes of cancer, a gene set enrichment-based approach was followed (see Sect. 2.7). This analysis highlighted enriched gene sets containing genes that may act as an oncogene or tumour suppressor gene or for which is shown that they repress tumour suppressor genes. The gene *LAMA3* (alpha-3 subunit of laminin 5) was proposed by Snijders and colleagues as one of the candidate driver genes in the development of oral squamous cell carcinoma and appears to be important for classification of *TP53* status as well. It may play an active role in cell adhesion, migration, and proliferation of the carcinoma cells[23]. Furthermore, the protein cortactin, contributing to tumour progression and metastasis formation[24], was gained in the wildtype samples, while the *TP53*-mutated samples were characterized by loss of the Smad transcription factors *Smad2*, *Smad4*, and *Smad7*. Members of the *TGFβ/Smad* signaling pathway are potent tumour suppressor genes[25] and the inactivation of Smad factors is correlated with loss of responsiveness to *TGFβ*-mediated signal transduction[26].

Important genes lost in the adenocarcinoma cell lines were *CUTL1* which activates a transcriptional program regulating genes involved in cell motility, invasion, and extracellular matrix composition[27], and *FGF-10*, a key regulator of lung branching morphogenesis[28]. In our data set the genes *BAF57* (important in transcriptional repression of tumour suppressor genes among which *BRCA1*[29]) and *HOXA5* (suggested to act as a tumour suppressor gene in breast cells[30]) seemed to be correlated with hereditary ovarian cancer, whereas loss of the *v-myb* oncogene seemed more characteristic for the sporadic group.

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 identifying a small set of class-specific aberrations that can distinguish patients, and 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[31]. 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 each group of 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.

A comparison of the regions found in each of the CV or LOO iterations showed a limited variability in the selected regions. This strengthens our confidence that the chromosomal regions found with our methodology are robust. 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[32]. For each data set the remaining 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.

It was not self-evident to find appropriate data sets because current studies utilize array CGH, not for classification purpose but for characterizing a specific set of patients. Clinical or histopathological variables that can be used to define two classifiable subgroups of patients are lacking in most of these studies. We applied our method on two extra data sets for the prediction of disease specific survival and distant recurrence, even though the main purpose of the original manuscripts was not survival[33−34] (results not shown). Although previous studies on CGH have shown evidence for an association of amplification with poor prognosis[35−36], we were not able to classify the patients in these two data sets properly according to survival using array CGH data. Further research is required on the appropriateness of array CGH data in survival-related studies.

The proposed method performed well in distinguishing adenocarcinoma from squamous cell carcinoma. This can partly be caused by the use of cell lines instead of clinical specimens, although a number of studies have characterized clinical cases of NSCLC with comparable CNV profiles as found by Garnis *et al.* in cell lines[10]. This good performance may also be due to the use of a high resolution tiling array. The use of such arrays yields potential for utilizing CNV profiles in classification for which a method has been presented here.

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 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, TAD-BioScope-IT, Silicos, SBO-BioFrame, SBO-MoKa, TBM-Endometriosis, TBM IOTA3. **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. D Pinkel, D G Albertson, *Nature Genetics*, **37**(Suppl.), 11 (2005).
2. W R Lai, M D Johnson *et al.*, *Bioinformatics*, **21**(19), 3763 (2005).
3. S Shah, W L Lam *et al.*, *Bioinformatics*, **23**, i450 (2007).
4. J Shawe-Taylor, N Cristianini, *Cambridge University Press*, (2004).
5. H Bhaskar, D C Hoyle *et al.*, *Comput Biol Med*, **36**(10), 1104 (2006).
6. J A K Suykens, T Van Gestel *et al.*, *World Scientific*, (2002).
7. G C Cawley, *Proceedings of the International Joint Conference on Neural Networks*, 1661 (2006).
8. A M Snijders, B L Schmidt *et al.*, *Oncogene*, **24**, 4232 (2005).
9. O Troyanskaya, M Cantor *et al.*, *Bioinformatics*, **17**(6), 520 (2001).
10. C Garnis, W W Lockwood *et al.*, *Int J Cancer*, **118**, 1556 (2006).
11. L M Starita, J D Parvin, *Current Opinion in Cell Biology*, **15**(3), 345 (2003).
12. S Shah, X Xuan *et al.*, *Bioinformatics*, **22**(14), e431 (2006).
13. B Schölkopf, K Tsuda *et al*, *MIT Press*, (2004).
14. V Vapnik, *Wiley*, (1998).
15. C Lai, M J T Reinders *et al.*, *BMC Bioinformatics*, **7**, 235 (2006).
16. Y H Yang, Y Xiao *et al.*, *Bioinformatics*, **21**(7), 1084 (2005).
17. I Lönnstedt and T P Speed, *Statist Sinica*, **12**, 31 (2001).
18. W Li, Y Yang, *Kluwer Academic*, 137 (2002).
19. J Liu, S Ranka *et al.*, *Bioinformatics*, **24**, i86 (2008).
20. A Subramanian, P Tamayo *et al.*, *Proc Natl Acad Scie*, **102**(43), 15545 (2005).
21. H M Wain, E A Bruford *et al.*, *Genomics*, **79**(4), 464 (2002).
22. B V North, D Curtis *et al.*, *Am J Hum Genet*, **71**, 439 (2002).
23. J Lohi, *Int J Cancer*, **94**, 763 (2001).
24. L Buday, J Downward, *Biochim Biophys Acta*, **1775**(2), 263 (2007).
25. M M Reinholz, M W An *et al.*, *Breast Cancer Res Treat*, **86**(1), 75 (2004).
26. A Kretschmer, K Moepert *et al.*, *Oncogene*, **22**(43), 6748 (2003).
27. P Michl, A R Ramjaun *et al.*, *Cancer Cell*, **7**(6), 521 (2005).
28. W Y Park, B Miranda *et al.*, *Dev Biol*, **201**(2), 125 (1998).
29. L Wang, R A Baiocchi *et al.*, *Mol Cell Biol*, **25**(18), 7953 (2005).
30. H Chen, E Rubin *et al.*, *J Biol Chem*, **280**(19), 19373 (2005).
31. C Klijn, H Holstege *et al.*, *Nucleic Acids Research*, **36**(2), e13 (2008).
32. The ENCODE Project Consortium, *Nature*, **447**, 799 (2007).
33. J Fridlyand, A M Snijders *et al.*, *BMC Cancer*, **6**, 96 (2006).
34. K Chin, S De Vries *et al.*, *Cancer Cell*, **10**, 529 (2006).
35. K Al-Kuraya, P Schraml *et al.*, *Cancer Research*, **64**, 8534 (2004).
36. H Blegen, J S Will *et al.*, *Anal Cell Pathol*, **25**, 103 (2003).