Cross-Platform Analysis with Binarized Gene Expression Data

Salih Tuna* and Mahesan Niranjan

School of Electronics and Computer Science,
ISIS Research Group,
University of Southampton, UK
{st07r,mn}@ecs.soton.ac.uk

Abstract. With widespread use of microarray technology as a potential diagnostics tool, the comparison of results obtained from the use of different platforms is of interest. When inference methods are designed using data collected using a particular platform, they are unlikely to work directly on measurements taken from a different type of array. We report on this cross-platform transfer problem, and show that working with transcriptome representations at binary numerical precision, similar to the gene expression bar code method, helps circumvent the variability across platforms in several cancer classification tasks. We compare our approach with a recent machine learning method specifically designed for shifting distributions, *i.e.,* problems in which the training and testing data are not drawn from identical probability distributions, and show superior performance in three of the four problems in which we could directly compare.

Keywords: Cross-platform analysis, binary gene expression, classification.

1 Introduction

The ability to observe the expression levels, or relative mRNA abundances, of thousands of genes in a given biological sample makes microarray technology a widely used tool in experimental biology. The potential of the technology as a diagnostic tool, producing a high dimensional feature vector upon which statistical pattern classification techniques such as Support Vector Machines (SVM) can be trained and applied, has received significant attention over the last decade [1]. Datasets from complex diseases including different types of cancer and diabetes have been analyzed in this manner, subsets of genes that are useful in discriminating the population with a disease from normal population have been identified for further validation.

A particular issue in such studies is variability at the biological and technical levels. Reproducibility of microarray results across different biological samples taken from the same tissue is reported to be very poor [2], while reproducibility

* Corresponding author.

V. Kadirkamanathan et al. (Eds.): PRIB 2009, LNBI 5780, pp. 439–449, 2009.

across technical replicates of amplified isolated mRNA is generally good [3]. Reasons for this have to do with the fact that mRNA is taken from a population of cells, each of which carrying a very small number of copies of each species. Except in experimental settings where the cells are artificially synchronized, this observation is largely true, leading to large biological variability. Similarly, variations in results across different laboratories and across platforms have been noted [4,5]. Much research in microarray studies is aimed at developing analytical techniques that are robust to systematic measurement variations.

In our past work [6], motivated by the observation that high numerical precision with which gene expression levels are reported in archives is incompatible with large biological variability, we showed that the quality of inference drawn from microarray studies is often not affected by progressive quantization of the expression levels. We established this in a number of different inference problems: classification, cluster analysis, detection of periodically expressed genes and the analysis of developmental time-course data. Building on this, we further showed that with a binary representation of the transcriptome, i.e., retaining only the information whether a gene is expressed or not, one could often achieve superior results by proper choice of distance metrics. Specifically, we used the Tanimoto similarity [7], borrowed from the chemoinformatics literature, and were able to explain some of the improvements obtained by a systematic variation in the probe level uncertainties of Affymetrix gene arrays [8]. We also established that in such reduced numerical precision representations, variability of inference arising from algorithmic choice in the pipeline of various pre-processing stages can be significantly reduced.

Binary representation of transcriptome has been shown to be effective in dealing with variation between laboratories by Zilliox and Irizzary [9], in their bar code method. The bar code is simply a binary representation of microarray outputs, but is computed over a very large collection of hybridizations of a particular type of array. In [9], the authors studied Affymetrix HGU133A Human array and using their barcodes and a simple nearest distance to template classifier demonstrated impressive results of tissue specificity of cancer populations. A particular limitation of the approach is distance-to-template classification, because it is known in statistical pattern recognition that such a classifier is optimal only for equal variance isotropic class conditional densities [10]. For gene expression data, this is a poor assumption because genes regulated by common transcription factors and those acting on signal transduction common pathways are often co-expressed. Complex diseases are often realized as disruptions in pathways or regulation, thus correlated expression should be very common in such datasets. While on the data used in [9] good results are obtained, it is not too difficult to find counter examples in which the performance of the bar code method is poor (see section 2.3). Similarly, Shmulevich and Zhang [11] also note the advantage of working with binary transcriptome data.

Warnat *et al.* [12] and Gretton *et al.* [13] offer novel algorithmic approaches for dealing with cross-platform variations. In their formulation training data for a cancer vs non-cancer SVM classifier is assumed to come from a particular

microarray platform and the unseen test data is assumed to come from a different platform. As one would expect, with no adjustment to the data, test set performance is very poor. In [12], Warnat *et al.* offer two solutions to improving on this: the use of median rank scores and quantile discretizations. The former approach uses ranks of genes as features in computing similarity metrics while the latter quantizes data into eight bins, the ranges of which are set to equalize bin occupancy. The second method is similar in spirit to the method we advocate in that ours is to quantize down to binary levels. In [13], Gretton *et al.* develop an approach aimed at the more generic problem of test set distributions being different from training set distributions. A weighting scheme known as kernel mean matching (KMM) is developed and microarray cross-platform inference is used as a test problem to evaluate their algorithm.

Binarizing continuous valued data as a means of improving the performance of specific classifiers have been reported in the machine learning literature in the past [14]. Such work, however, is not generic and is merely a statement about accidental improvements over weak baseline classifiers (naive Bayes, decision trees etc.). Our results are specific to transcriptome data and build on observed properties of the measurement environment. Further, our comparisons are against classifier with high performance (i.e., SVM).

In this paper we show that a binary representation of the transcriptome, when combined with a suitable similarity metric and cast in a kernel classifier setting, can yield performance that is competitive, and often superior, to methods developed in the literature to address this problem. This, and other examples of high performance from binary representations we have reported previously, arise largely from the fact that often the useful information relating to gene expression is simply if it is transcribed or not, rather than in the actual cellular concentration of the transcripts. Even if the information is in transcript abundances, as noted earlier, heterogeneity within a population of cells makes the measurement unreliable. In this context, quantization of the data has a noise rejection property which our method takes advantage of.

2 Methods

2.1 Quantization

Quantization of microarray has been studied in the literature, for example [15,16,17]. Among possible methods, we choose the quantization method of Zhou *et al.* [15] where mixture of Gaussians are used for the different states of gene expression values. Our justification for choosing [15]'s method is that it is relatively more principled than other approaches for quantization. Arbitrary thresholds set by other researchers are not necessarily transferable across different platforms or experiments due to variabilities induced by image processing and normalization, while the method in [15] depends on the underlying probability density of the expression levels and hence the idea is portable to any situation. We focused on binary representation of these measurements. Gene expression values are quantized by fitting Gaussian mixture model to the expression values:

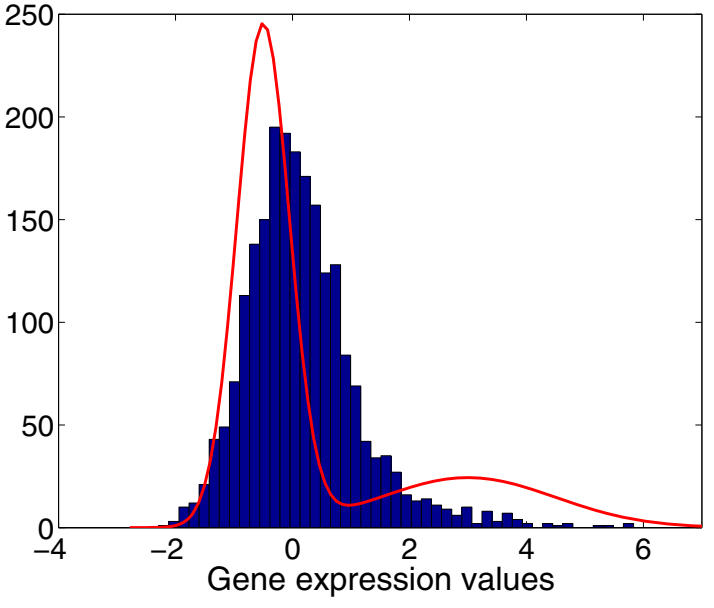

Fig. 1. Histogram of expression levels taken from [25] and a two component Gaussian mixture model of the distribution. The quantization threshold is a function of the means and standard deviations of the two mixture components (Eqn. 2).

$$p(\mathbf{x}) = \sum_{k=1}^{M} \lambda_k \, \mathsf{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k) \tag{1}$$

where $p(\mathbf{x})$ is the probability density of gene expression measurement, M, the number of mixture components, and $\mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ is a Gaussian density of mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$. Fitting such a model is by standard maximum likelihood techniques, and we used the gmm function in NETLAB software (http://www.ncrg.aston.ac.uk) for this purpose. We used two component mixtures, corresponding to $M = 2$ in the above equation. Fig. 1 shows an example of gene expression values fitted to two center GMM.

After learning parameters of the model, threshold Th is chosen as:

$$Th = \frac{\mu_1 + \sigma_1 + \mu_2 - \sigma_2}{2} \tag{2}$$

to achieve binary quantization.

2.2 Tanimoto Kernel

Tanimoto coefficient (T) [7], between two binary vectors of gene expressions, is defined as:

$$T = \frac{c}{a + b - c} \tag{3}$$

where a is the number of expressed points for the first gene, b is the number of expressed points for the second gene and c is the number of common expressed points in two genes. Tanimoto similarity ranges from 0 (no points in common) to 1 (exact match) and is the rate of the number of common bits on to the total number of bits on two vectors. It focuses on the number of common bits that are on.

Following the definition of Tanimoto similarity, Tanimoto kernel is defined as [18,19]:

$$K_{Tan}(\mathbf{x}, \mathbf{z}) = \frac{\mathbf{x}^T\mathbf{z}}{\mathbf{x}^T\mathbf{x} + \mathbf{z}^T\mathbf{z} - \mathbf{x}^T\mathbf{z}} \tag{4}$$

where $a = \mathbf{x}^T\mathbf{x}$, $b = \mathbf{z}^T\mathbf{z}$ and $c = \mathbf{x}^T\mathbf{z}$. It follows from the work of Swamidass *et al.* [18] and Trotter [19] that this similarity metric is useful as a valid kernel, i.e., kernel computations in the space of the given binary vectors map onto inner products in a higher dimensional space so that SVM type optimizations for large margin class boundaries is possible. We incorporated this kernel into the `MATLAB` SVM implementation of Steve Gunn [20] (`http://www.isis.ecs.soton.ac.uk/isystems/kernel/`).

2.3 Bar Code vs. SVM

Since the bar code method of Zilliox and Irizarry [9] is the closest in literature to our work, we give a quick overview and evaluation of its performance. The binary representation for a class of data (tumor in a particular tissue) is derived for a particular array, Affymetrix HGU133A Human array, by scanning through a large collection of expression levels archived in microarray repositories. Predictions on test data are made by computing nearest Euclidean distance to pre-computed bar codes. As we note in the introduction, we should be skeptical about high performance from a distance-to-template classifier as such an approach is only Bayes' optimal under isotropic equal variance assumptions. To verify this, we first established that the bar code approach cannot compete with SVM. We used the `R` code which is made available by the authors' at their web page: `http://rafalab.jhsph.edu/barcode/` and used three datasets downloaded from ArrayExpress (`http://www.ebi.ac.uk/arrayexpress/`) and Gene Expression Omnibus (GEO) (`http://www.ncbi.nlm.nih.gov/geo/`). Two of these were used in [9] and the other was not. Prediction accuracies for these three, comparing the bar code method to Tanimoto-SVM, are shown in Table 1. We note that training and testing on the same database, as we have done with Tanimoto-SVM, achieves consistently better prediction accuracies than the bar code method. But in fairness to the bar code method we remark that their intention is to make predictions on a new dataset based on accumulated historic knowledge, rather than repeat the training/testing process all over again. On this point, while there is impressive performance reported on the datasets

Table 1. Comparison of Tanimoto-SVM with [9]' s bar code

Dataset	Data type	Method	Accuracy
E-GEOD-10072	Binary	Bar code	0.50
Lung	Binary	Tanimoto-SVM	0.89 ± 0.03
Lung tumor vs. normal	Binary	Tanimoto-SVM	0.99 ± 0.03
GSE2665	Binary	Bar code	0.95
lymph node/tonsil	Binary	Tanimoto-SVM	0.99 ± 0.02
lymph node vs. tonsil	Binary	Tanimoto-SVM	1.0 ± 0.0
GSE2603	Binary	Bar code	0.90
Breast Tumor	Binary	Tanimoto-SVM	0.99 ± 0.01
Breast Tumor vs. normal	Binary	Tanimoto-SVM	0.99 ± 0.01

Zilliox and Irizarry [9] worked on, the method can fail badly too, as in the case of the lung cancer prediction task E-GEOD-10072 shown in Table 1. On Table 1 'Lung' corresponds to classifying lung vs. breast and lymph node/tonsil which is a similar approach to bar code. 'lung tumor vs. normal' corresponds classifying tumor vs. normal in lung only. The same terminology applies to the other two problems as well.

Part of the success of the Tanimoto kernel in the microarray setting comes from a systematic variability at the probe level of Affymetrix arrays. We have noted [8] that in a given experiment, the average probe level uncertainty computed amongst expressed genes systematically reduces with the number of expressed genes; i.e., the larger the number of expressed genes lower the uncertainty of measurements. Amongst 50 experiments we looked at there was only one experiment for which this observation did not hold. This variability has a direct bearing when using Tanimoto similarity. For two pairs of expression profiles which differ by the same Hamming (or Euclidean) distance, Tanimoto similarity will be higher for the pair that has a greater number of expressed genes (thereby placing a higher emphasis on experiments with lower probe level uncertainties). Other authors have also exploited probe level uncertainties in principal component analysis [22,23] and cluster analysis [24].

3 Experiments

3.1 Datasets

To demonstrate how binary representations help in cross-platform inference, we carried out experiments on breast and prostate cancer datasets. These datasets are the same as those used in [12] and [13] and were given to us by the authors in processed format (i.e., we worked with the expression levels rather than with the raw data at the CEL file or image levels). These data come from spotted cDNA and Affymetrix platforms, and details of the four datasets are summarized in Tables 2 and 3. Warnat *et al.* [12] preprocessed all the data and found the

Table 2. Details of breast cancer studies

Study	Breast cancer			
	Platform	No of common genes	Samples	Target variable
West *et al.* [25]	Affymetrix	2166	49	ER-status: 25(+), 24(-)
Gruvberger *et al.* [26]	cDNA	2166	58	ER-status: 28(+), 30(-)

Table 3. Details of prostate cancer studies

Study	Prostate cancer			
	Platform	No of common genes	Samples	Target variable
Welsh *et al.* [27]	Affymetrix	4344	33	9 normal, 24 tumor
Dhanasekaran *et al.* [28]	cDNA	4344	53	19 normal, 34 tumor

subset of common genes by means of the Unigene database (`http://www.ncbi.nlm.nih.gov/unigene`).

3.2 SVM Classification

In implementing SVM classifiers, we first ensured that our implementation achieves the same results as reported in [12]. Table 6, "cont-not normalized" column confirms that our implementation achieves the same results reported previously. Then, following the suggestion in [13], we normalized each array to have a mean of zero and standard deviation one, and trained and tested our SVM implementations. This normalization has a significant impact on the results ("cont-normalized", in Table 6). As used in these papers we used linear kernel SVMs with a setting of $C = 1000$ for the margin parameter, and confirmed previously quoted results are reproducible. We then quantized the data and applied Tanimoto kernel SVM. Note that this kernel has no tuning parameters. We implemented quantization on an array by array basis. In previous work we have experimented with different ways of quantization (array by array, gene by gene and a global method), and noted only small differences between these over a range of quantization thresholds [6].

3.3 Results

Tables 4 and 5 show the difference in classification between continuous and binary representations on the two cancer classification problems. Accuracies are shown for 25 random partitions of the data into training and testing sets, along with standard deviations quantifying the uncertainty in this process. We see that in three out of the four cases, binarization, and the use of Tanimoto kernel, offers significant improvements, and performs no worse than continuous data in the fourth. In Warnat *et al.* [12], results are averaged over 10 cross validation runs, but the paper does not report the variation across results.

Table 6 presents results of training SVMs with one type of data and testing the performance on data from a different platform. In this cross-platform

Table 4. Breast cancer results for cross-platform analysis. Data is randomly partitioned into training and testing for 25 times.

Dataset	Data type	Method	Accuracy
Gruvberger *et al.*	Cont.	Linear-SVM	0.80±0.07
Gruvberger *et al.*	Binary	Tanimoto-SVM	0.82±0.08
West *et al.*	Cont.	Linear-SVM	0.76±0.15
West *et al.*	Binary	Tanimoto-SVM	0.79±0.11

Table 5. Prostate cancer results for cross-platform analysis. Data is randomly partitioned into training and testing for 25 times.

Dataset	Data type	Method	Accuracy
Dhanasekaran *et al.*	Cont.	Linear-SVM	0.89 ± 0.06
Dhanasekaran *et al.*	Binary	Tanimoto-SVM	0.89 ± 0.05
Welsh *et al.*	Cont.	Linear-SVM	0.92 ± 0.06
Welsh *et al.*	Binary	Tanimoto-SVM	0.96 ± 0.06

Table 6. Cross-platform results. Array-by-array quantization. The notation "Gruvberger → West" indicates that we train on Gruvberger' s data and test on West' s data.

Dataset	Data type	Accuracy
Gruvberger → West	Cont.(not normalized)	0.49
Gruvberger → West	Cont.(normalized)	0.94
Gruvberger → West	Binary	0.96
West → Gruvberger	Cont.(not normalized)	0.52
West → Gruvberger	Cont.(normalized)	0.93
West → Gruvberger	Binary	0.90
Dhanasekaran → Welsh	Cont.(not normalized)	0.27
Dhanasekaran → Welsh	Cont.(normalized)	1
Dhanasekaran → Welsh	Binary	1
Welsh → Dhanasekaran	Cont.(not normalized)	0.64
Welsh → Dhanasekaran	Cont.(normalized)	0.93
Welsh → Dhanasekaran	Binary	1

comparison, normalization as a first step has a big impact. Further improvement is obtained by our binarized Tanimoto approach. While in one of the four experiments this approach gives poor performance, it proves useful in the other three. In Table 7 we give a comparison with other previously published results on the same datasets, namely the median rank and quantile discretization of [12] and the kernel mean matching approach of [13]. While the number of experiments is small, we note that the binarized Tanimoto method we advance has merit in terms of its performance in a cross-platform setting.

Table 7. Comparison of our approach to the published results in literature. Accuracies obtained by SVM are compared.

Study	Train → Test	Method			
		MRS	QD	KMM	Binary
Breast cancer	Gruvberger → West	0.63	0.86	0.94	**0.96**
	West → Gruvberger	**0.95**	0.92	**0.95**	0.90
Prostate cancer	Dhana → Welsh	0.88	0.97	0.91	**1**
	Welsh → Dhana	0.89	0.91	0.83	**1**

Note KMM is a sample re-weighting process designed to match the test set distribution to the training set (in feature space means) by a quadratic programming formulations. With microarray data, imposing such a shift is an artificial construct, whereas our results show that similar, if not superior, performance is achievable simply by choosing an appropriate data representation.

4 Conclusion

In this paper we show that a binary representation of gene expression profiles, combined with a kernel similarity metric that is appropriate for such data, has the potential to address the important problem in microarray based phenotype classifications of cross-platform inference. While the experimental work is on a very small number of datasets, which were the only ones available to us at this time from previous studies, we believe this advantage comes from using a data representation that respects properties of the measurement environment. This approach is not only limited to cross-platform analysis but can also be successfully applied in Affymetrix vs. Affymetrix (e.g. see results in Table 1) where we show that data from one Affymetrix platform can be robustly transferred to another. Our current work is on extending the study to a larger collection of datasets, the difficulty in doing this being the matching of the gene identities.

Acknowledgments. We are grateful to Arthur Gretton and Karsten Borgward for providing the datasets used in this study.

References

1. Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares Jr., M., Haussler, D.: Knowledge-based analysis of microarray gene expression data by using support vector machines. PNAS 97(1), 262–267 (2000)
2. Tomayko, M.M., Anderson, S.M., Brayton, C.E., Sadanand, S., Steinel, N.C., Behrens, T.W., Shlomchik, M.J.: Systematic Comparison of Gene Expression between Murine Memory and Naive B Cells Demonstrates That Memory B Cells Have Unique Signaling Capabilities. J. Immunol. 181(1), 27 (2008)
3. MAQC consortium, The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat. Biotechnol. 24, 1151–1161 (2006)

4. Draghici, S., Khatri, P., Eklund, A.C., Szallasi, Z.: Reliability and reproducibility issues in DNA microarray measurements. Trends Genet. 22, 101–109 (2006)
5. Kuo, W.P., Jenssen, T.K., Butte, A.J., Ohno-Machado, L., Kohane, I.S.: Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics 18(3), 405–412 (2002)
6. Tuna, S., Niranjan, M.: Inference from low precision transcriptome data representation. Journal of Signal Processing Systems (April 22, 2009), doi:10.1007/s11265-009-0363-2
7. Tanimoto, T.T.: IBM Internal Report, An elementary mathematical theory of classification and prediction (1958)
8. Tuna, S., Niranjan, M.: Classification with binary gene expressions. Journal of Biomedical Sciences and Engineering (in press, 2009)
9. Zilliox, M.J., Irizarry, R.A.: A gene expression bar code for microarray data. Nat. Met. 4(11), 911–913 (2007)
10. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons, USA (2001)
11. Shmulevich, I., Zhang, W.: Binary analysis and optimization-based normalization of gene expression data. Bioinformatics 18(4), 555–565 (2002)
12. Warnat, P., Eils, R., Brors, B.: Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes. BMC Bioinformatics 6, 265 (2005)
13. Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., Scholkopf, B.: Covariate shift by kernel mean matching. In: Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D. (eds.) Dataset shift in machine learning, pp. 131–160. Springer/The MIT Press, London (2009)
14. Dougherty, J., Kohavi, R., Sahami, M.: Supervised and Unsupervised Discretization of Continuous Features. In: International Conference on Machine Learning, pp. 194–202 (1995)
15. Zhou, X., Wang, X., Dougherty, E.R.: Binarization of microarray data on the basis of a mixture model. Mol. Cancer Ther. 2(7), 679–684 (2003)
16. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using Bayesian networks to analyze expression data. J. Comput. Biol. 7(3-4), 601–620 (2000)
17. Brazma, A., Jonassen, I., Vilo, J., Ukkonen, E.: Predicting Gene Regulatory Elements in Silico on a Genomic Scale. Genome Res. 8(11), 1202–1215 (1998)
18. Swamidass, S.J., Chen, J., Bruand, J., Phung, P., Ralaivola, L., Baldi, P.: Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. Bioinformatics 21(suppl. 1), i359–i368 (2005)
19. Trotter, M.W.B.: Support vector machines for drug discovery. Ph.D. thesis, University College London, UK (2006)
20. Gunn, S.R.: Support vector machines for classification and regression, Technical Report, University of Southampton (1997), `http://www.isis.ecs.soton.ac.uk/isystems/kernel/`
21. Milo, M., Fazeli, A., Niranjan, M., Lawrence, N.D.: A probabilistic model for the extraction of expression levels from oligonucleotide arrays. Biochem. Soc. Trans. 31(Pt 6), 1510–1512 (2003)
22. Rattray, M., Liu, X., Sanguinetti, G., Milo, M., Lawrence, N.D.: Propagating uncertainty in microarray data analysis. Brief Bioinform. 7(1), 37–47 (2006)
23. Sanguinetti, G., Milo, M., Rattray, M., Lawrence, N.D.: Accounting for probe-level noise in principal component analysis of microarray data. Bioinformatics 21(19), 3748–3754 (2005)

24. Liu, X., Lin, K., Andersen, B., Rattray, M.: Including probe-level uncertainty in model-based gene expression clustering. BMC Bioinformatics 8(1), 98 (2007)
25. West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson Jr., J.A., Marks, J.R., Nevins, J.R.: Predicting the clinical status of human breast cancer by using gene expression profiles. PNAS 98(20), 11462–11467 (2001)
26. Gruvberger, S., Ringnér, M., Chen, Y., Panavally, S., Saal, L.H., Borg, A., Ferno, M., Peterson, C., Meltzer, P.S.: Estrogen Receptor Status in Breast Cancer Is Associated with Remarkably Distinct Gene Expression Patterns. Cancer Res. 61(16), 5979–5984 (2001)
27. Welsh, J.B., Sapinoso, L.M., Su, A.I., Kern, S.G., Wang-Rodriguez, J., Moskaluk, C.A., Frierson, H.F., Hampton, G.M.: Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. Cancer Res. 61(16), 5974–5978 (2001)
28. Dhanasekaran, S.M., Barrette, T.R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K.J., Rubin, M.A., Chinnaiyan, A.M.: Delineation of prognostic biomarkers in prostate cancer. Nature 412(6849), 822–826 (2001)