

---

## SVD Analysis of Gene Expression Data

Krzysztof Simek<sup>1</sup> and Michał Jarzab<sup>2</sup>

<sup>1</sup> Institute of Automatic Control, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland; krzysztof.simek@polsl.pl

<sup>2</sup> Maria Skłodowska-Curie Memorial Cancer Center and Institute of Oncology, Gliwice Branch, Wybrzeże Armii Krajowej 15, 44-100 Gliwice, Poland; mjarzab@io.gliwice.pl

**Summary.** The analysis of gene expression profiles of cells and tissues, performed by DNA microarray technology, strongly relies on proper bioinformatical methods of data analysis. Due to the large number of analyzed variables (genes) and the usually low number of cases (arrays) in one experiment, limited by the high cost of the technology, the biological reasoning is difficult without previous analysis, leading to a reduction of the problem dimensionality. A wide variety of methods have been developed; the most useful, from a biological point of view, are methods of supervised gene selection with estimation of false discovery rate. However, supervised gene selection is not always satisfying for the user of microarray technology, as the complexity of biological systems analyzed by microarrays rarely can be explained by one variable. Among unsupervised methods of analysis, hierarchical clustering and principal component analysis (PCA) have gained wide biological application. In our opinion, singular value decomposition (SVD) analysis, which is similar to PCA, has additional advantages that are very essential for the interpretation of the biological data. In this chapter we shall present how to apply SVD to unsupervised analysis of transcriptome data obtained by oligonucleotide microarrays. These results have been derived from several experiments, carried out at the DNA oligonucleotide microarray Laboratory at the Institute of Oncology, Gliwice, and are currently analyzed from a biological point of view.

**Key words:** Singular value decomposition, gene expression data, gene selection, hierarchical clustering.

### 32.1 Unsupervised Methods in Analysis of Gene Expression Data

Unsupervised analysis should be the first step in all microarray experiments. It reveals the intrinsic structure of the data, which helps to verify whether the assumptions of the microarray experiment are held and whether the major observed variability is related to the experimental variables or to confounding factors. Moreover, it allows for rapid detection of outliers, thus it is a valuable method of quality control. Last, but not least, unsupervised analysis is invaluable in genome-wide experiments, in which we aim to classify the samples based on their gene expression profiles and in this way to gain biological knowledge about subgroups of these samples. These three ap-

proaches are discussed later in this chapter. Various methods of unsupervised analysis have been developed. Initially, they were based on known statistical methods of clustering. At the moment, a widely accepted method is hierarchical clustering with various distance definitions and metrics. Other algorithms, which were used mainly to cluster genes and were not efficient in clustering of a limited number of microarray samples, are less widespread and are used only for specific occasions. Hierarchical clustering is a very useful method of analysis and visualization (by dendrograms) of differences/similarities between samples or genes. It allows one to calculate the distance between samples and divides the whole group into a chosen number of clusters. Moreover, being similar to methods used in taxonomy, dendrograms are easily understood by biologists. Dendrogram classification of genes and samples accompanied by a heatmap plot is probably one of the most informative methods of microarray data visualization.

However, when used in unsupervised analysis, hierarchical clustering has serious drawbacks. To be a fully unsupervised method, it should be used on all microarray genes or on genes filtered without the use of any variables in question. Clustering by all genes is a method which allows one to detect large differences in gene expression profile (e.g., it can separate expression profiles of different tissues), but when used on less variable experimental data, it will rather depend on technical parameters of arrays. This way it is more suitable for visualization of technical differences, but it has no place in routine use to answer biological questions. This is partly caused by the large number of genes on some arrays (more than 50K transcripts on routinely used oligonucleotide microarrays), from which only a small proportion is expected to be expressed. Filtering of genes by different variance measures may be useful in reducing the number of genes for clustering, but taking into account that some coordinated biological changes show a rather low amplitude, this strategy does not seem to be optimal. Moreover, in routine hierarchical clustering of normalized data every gene has equal weight, and thus some (e.g., not expressed) genes may bias the final result. At present there is no widely accepted method to select just the important transcripts in a gene expression profile. The use of singular value decomposition (SVD) for analysis of microarray data has been detailed in [1]. A profound description of SVD and a comprehensive survey of its applications in gene expression data analysis are given in [2] and references therein. SVD is a standard method of linear algebra and it may be easily performed on large matrices without significant computational cost. The most important feature of SVD which predisposes it to be used for the analysis of microarray data is that the characteristic modes obtained from decomposition of a gene expression matrix of various samples usually have a meaningful biological interpretation. In a homogeneous biological system, like in *in vitro* cell culture, the majority of genes are coordinately regulated by a limited number of signals, thus they exhibit similar expression profiles. Obviously, the complexity of this regulation is large and many genes are affected by numerous transcription factors, but it has been shown in yeast that common gene expression patterns in a cell cycle may be easily detected by SVD [3]. In experiments on cells from cultures the diversity of gene expression depends on the complexity of each cell transcriptome and the differences in the cellular cycle stage. However, microarray experiments are very often performed on even more complex biological systems, i.e.,

tissue samples. In tissues, the inherent feature is the heterogeneity of cells within the sample and until the microdissection techniques (allowing one to dissect the tissue into cells of various morphology) become widespread, the expression profile of the tissue depends on the transcriptional changes in cells and the differences in cellular content of the tissue. As we will show later, this second factor very often prevails over the whole gene expression profile and the analysis of characteristic modes then helps to interpret their biological meaning. However, the shape of the modes rarely provides us with direct conclusions, and thus we extend the method to select the genes correlated to each characteristic mode. Such an unsupervised gene selection method, when used together with hierarchical clustering, is a strong and powerful tool for biological analysis of microarray data. We briefly describe the principles of our method and show some examples of its application.

## 32.2 Singular Value Decomposition to Select Major Variability Genes in Transcriptome Data

### 32.2.1 Definition of SVD

The SVD of any  $n \times m$  matrix  $A$  (gene expression matrix) has the form [4]

$$A = USV^T, \quad (32.1)$$

where  $U$  is an  $n \times m$  orthonormal matrix, whose columns are called the left singular vectors of  $A$  (gene coefficient vectors), and  $V$  is an  $m \times m$  orthonormal matrix, whose columns are called the right singular vectors of  $A$  (expression level vectors).  $S$  is a diagonal matrix  $S = \text{diag}(s_1, s_2, \dots, s_m)$ . The diagonal elements of matrix  $S$  are, as a convention, listed in a descending order,  $s_1 \geq s_2 \geq \dots \geq s_m \geq 0$ , and are called the singular values of  $A$ .

### 32.2.2 Characteristic Modes

Let us denote the rows of matrix  $SV^T$  by  $X_i, i = 1, \dots, m$ . The orthogonal vectors  $X_i = s_i v_i^T$  are called the *gene characteristic modes* associated with gene expression matrix  $A$ . In an analogous way, the rows of matrix  $SU^T$ , can be defined as *array characteristic modes*. The properties of both types of modes are similar so we present them only for the gene modes.

The profile of the  $j$ th gene, included in the row  $A_j$  of matrix  $A$ , can be obtained as linear combinations of the characteristic modes. The coefficients of the combination are the corresponding entries of matrix  $U$ ,

$$A_j = \sum_{i=1}^m U_{ji} X_i. \quad (32.2)$$

The contribution of modes to the gene pattern decreases from the higher order to the lower order modes. Usually not all characteristic modes are needed to reconstruct gene expression patterns with a reasonable accuracy. We may use a truncated expression,

$$A_j = \sum_{i=1}^l U_{ji} X_i, \quad l < m. \quad (32.3)$$

There are several heuristic methods to estimate the number  $l$  of the most significant characteristic modes [5]. One of the simplest is to retain just enough modes to capture a large percentage of the overall expression. Usually values of 70–90% are proposed. The other procedure is to exclude characteristic modes such that the fraction of expression  $p_i$  they capture is less than  $(70/m)\%$ . Another method is the examination of scree plots for  $s_i^2$  or  $\log s_i^2$ . Using this method one can usually find a natural border between significant and insignificant singular values (called the elbow). The singular values which represent the magnitudes of the corresponding modes can be used as measures of the relative significance of each characteristic mode in terms of the fraction of overall expression that it captures:

$$p_i = \frac{s_i^2}{\sum_{j=1}^m s_j^2}, \quad i = 1, \dots, tgm. \quad (32.4)$$

A similar index, defining the contribution of the  $i$ th mode to the pattern of the  $k$ th gene, can be defined in the form

$$c_k^i = \frac{(U_{ki} s_i)^2}{\sum_{j=1}^m (U_{kj} s_j)^2}. \quad (32.5)$$

### 32.2.3 Gene Selection Using SVD

In the clustering literature, SVD is sometimes applied to extract the cluster structure in the data and reduce its dimensionality prior to clustering. Since characteristic modes are uncorrelated and ordered, the first few most significant ones, which contain most of the variations in the data, are usually used. Namely, characteristic mode coefficients (gene coefficient vectors), instead of original variables, are used for clustering. Our approach differs from that known from the literature. We apply SVD as a preprocessing step before cluster analysis of gene expression data. As a result, a small set of original genes is selected and then applied to cluster samples using one of the standard algorithms.

**Algorithm of gene selection.** The gene selection algorithm inspects gene coefficient vectors (columns of matrix  $U$ ) corresponding to the set of the most significant characteristic modes. Each coefficient is compared to the threshold value [6], whose meaning is similar to a  $3\sigma$  statistical significance cutoff, equal to  $Wn^{-1/2}$ , where  $n$  is the number of genes and  $W$  is a weight factor whose recommended value is greater than 3. If the magnitude of the element is greater than the threshold value, the corresponding

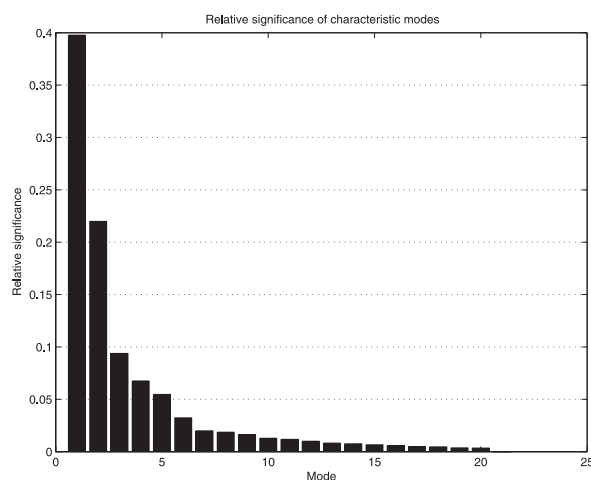
gene is selected to the clustering set. In practice we choose genes with sufficiently big coefficients for the most important characteristic modes, or in other words, genes for which values of index (32.5) for the most important modes are big enough. In the result we obtain a set of genes with patterns “similar” to the dominant modes.

### 32.3 Applications of Unsupervised Singular Value Decomposition Method to Microarray Data

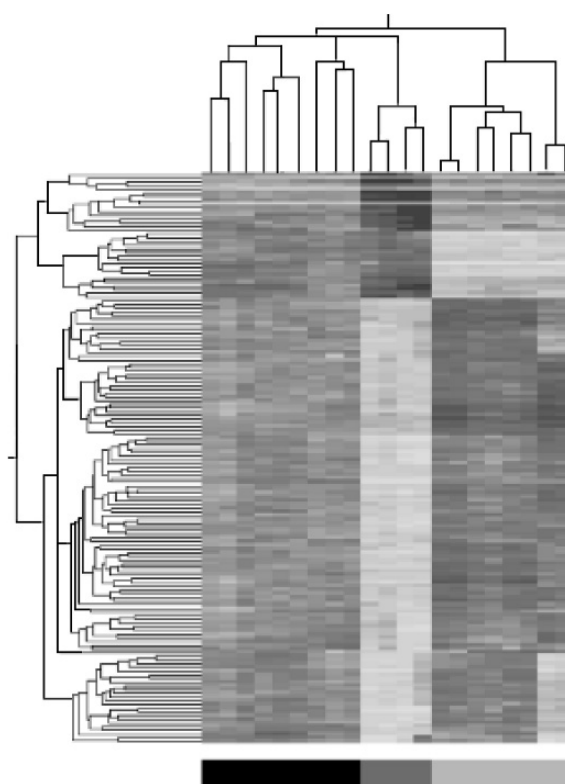
We applied the SVD algorithm to four datasets of gene expression profiles related to cancer. The first data set was obtained from a cell culture experiment with murine melanoma cells exposed to hypoxic conditions (low oxygen tension) [8]. In this experiment oligonucleotide microarrays were used (Affymetrix MG-U74). Hypoxia was obtained by three different methods: chemical mimicry with cesium chloride, a Billups–Rothenberg chamber and an incubator with regulated  $O_2$  tension, set to low  $O_2$  level. Three other experiments were carried out with clinical material: tumor samples and corresponding normal tissues analyzed by gene expression profiling. In the first experiment we compared gene expression of papillary thyroid cancer to normal thyroid tissue from the same individual. The analyzed dataset consisted of 16 tumor and 16 normal samples hybridized to HG-U133A oligonucleotide microarrays and preprocessed by MAS5 algorithm [7]. In the second experiment we compared two different histological types of thyroid cancer. We performed SVD on a set of 57 thyroid neoplasms: 38 of papillary histology and 19 follicular adenomas or carcinomas [9]. We also compared gene expression profiles of laryngeal cancer and corresponding normal tissue and analyzed by HG-U133 Plus 2 microarray 17 samples of laryngeal cancer (9 microarrays) and normal tissue (8 samples) [10].

#### 32.3.1 The Influence of Hypoxia on Gene Expression Profile

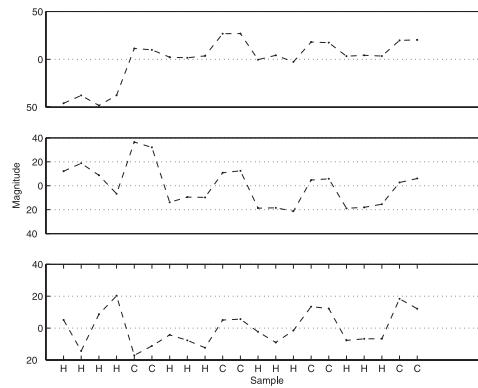
Hypoxia is an intrinsic feature of many malignant tumors and is highly related to their resistance to various treatment modalities (chemotherapy, radiation therapy). Thus, a sound understanding of the molecular mechanisms underlying hypoxia is crucial to the development of new molecular methods of therapy. In this experiment we obtained 13 microarrays from hypoxic cells and 8 from control samples. We performed SVD of the whole dataset and obtained 20 gene characteristic modes with the first mode accounting for 39.7% of variance and the first 5 modes describing 83% of the data variability (Fig. 32.1). This result was highly satisfying (more than 80% of the variability explained by only 5 expression patterns), but simultaneously showed that confounding factors influence the effect of hypoxia. We selected 154 genes correlated to the first mode profile. Hierarchical clustering of all samples based on these 154 genes (Fig. 32.2) ideally divided control and hypoxic samples which confirmed that the major source of variability in the analyzed dataset was the hypoxic-control difference. This was the proof that genes selected on the basis of this comparison are not strongly affected by technical factors. An important fact in understanding the SVD method was that the first mode profile itself was less clearly differentiating hypoxic and control



**Fig. 32.1.** Relative significance of gene characteristic modes in hypoxia experiment.



**Fig. 32.2.** Hierarchical clustering of first mode genes in hypoxia study. Real hypoxia (black) and hypoxia mimicry (dark grey) is distinctly different in expression profile from control samples (light grey).

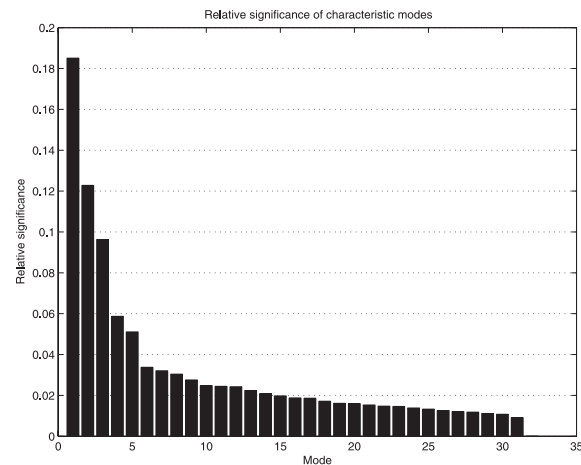


**Fig. 32.3.** First three gene characteristic modes in hypoxia study. Difference between hypoxic (H) and control (C) samples can be shown in all three modes, but it is influenced by other sources of variability (CoCl<sub>2</sub>-treated specimens, first four from the left, are clearly different from other hypoxic samples).

samples. The shape of this pattern was influenced not only by the hypoxic-normal difference, but also by other factors (Fig. 32.3). Thus, direct analysis of mode patterns was not fully justified and we preferred to use gene selection. This approach is very powerful, not only because it allows one to properly classify samples, but because it also gives a number of genes related to each expression pattern. The gene content of this list can facilitate the biological reasoning, as shown later. In summary, SVD analysis allowed us to confirm that the control-treated difference was the major source of variability in the performed experiment, detected the important changes between two hypoxic conditions and thus influenced the further supervised analysis [8].

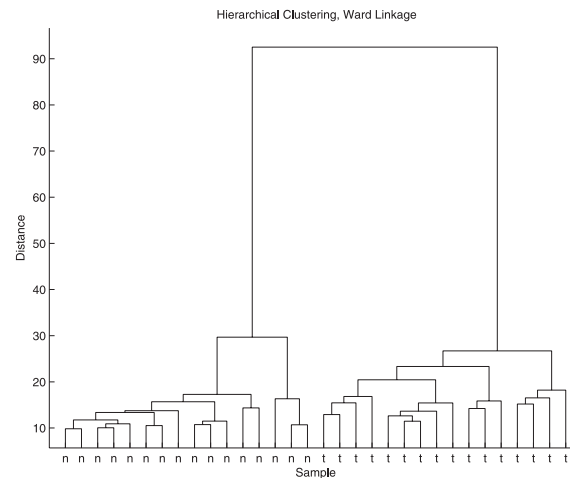
### 32.3.2 Gene Expression Profile of Papillary Thyroid Cancer

The next application of the SVD method is the analysis of the gene expression profile in papillary thyroid carcinoma. This malignant tumor is the most common cancer of the thyroid gland [7]. The first three characteristic modes in this dataset, which were considered significant, accounted for 40.4% of the variance in this dataset (Fig. 32.4). The much lower percentage of variability described by SVD analysis in comparison to the data described above is characteristic for the study of clinical specimens, where the samples differ not only by one variable (like hypoxia in the previous example), but by a whole set of features related to patients and the disease. Nevertheless, the most important factor of variability, revealed by the first mode, was the difference between normal thyroid and papillary thyroid cancer—all samples, clustered by 310 genes correlated to the first mode, were ideally separated into tumors and normals (Fig. 32.5). This was the proof of proper quality control in sample selection, and it confirmed that the tumor-normal difference in this type of cancer is large enough to be detected by unsupervised methods. Even more interesting were the results of clustering by the second and third modes: both these patterns were unexpected, and before the study we did



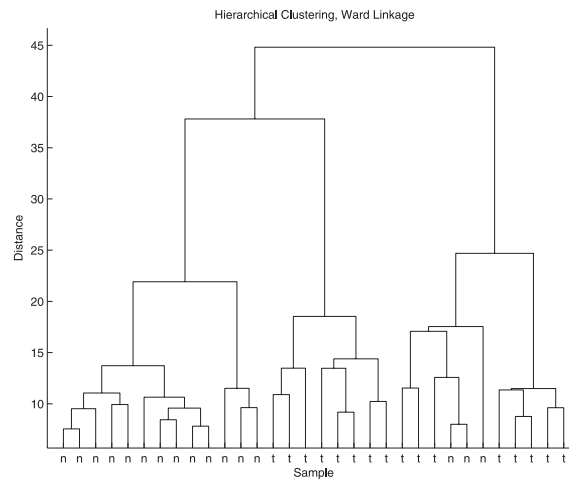
**Fig. 32.4.** Relative significance of characteristic modes in thyroid cancer study.

not have any knowledge as to what factors (except tumor-normal difference) might influence the expression profile. The clustering based on second mode expression profile revealed two groups of samples, both containing two subgroups of tumors and normals (Fig. 32.6). We could not attribute this subdivision to any clinical factor. The clustering based on the third mode genes did not show any attributable pattern, only some paired samples (tumor-normal) were co-clustering together. When we analyzed the content of the selected gene lists, we revealed that for the second and the third modes the significant proportion of transcripts was immune-related genes (40.3% in the second



**Fig. 32.5.** Hierarchical clustering of thyroid tumors (t) and normal thyroid tissues (n) based on genes selected by the first mode.





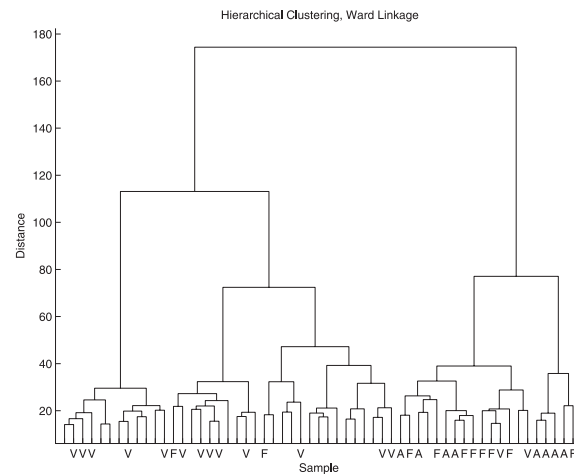
**Fig. 32.6.** Hierarchical clustering of thyroid tumors (t) and normal thyroid tissues (n) based on genes selected by the 2nd mode. The analysis reveals two distinct subgroups, each containing both tumor and normal tissues.

mode, 39.2% in the third mode). This suggested that the expression profile of papillary thyroid cancer is strongly influenced by expression of immune-related genes, the majority of them probably expressed in infiltrating leukocytes. However, the origin of these transcripts has still to be confirmed.

### 32.3.3 Distinguishing Between Histological Subtypes of Differentiated Thyroid Cancer

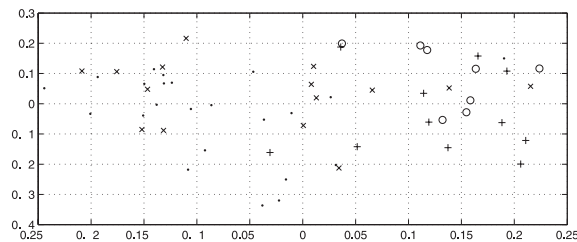
Papillary thyroid cancer is the most common malignant tumor of the thyroid, but there exist other histological types of thyroid neoplasms. The diagnosis of this tumor by microscopic analysis is related to numerous problems. The most difficult is the differentiation between two entities: follicular carcinoma and follicular adenoma, the latter one being a benign disease and not demanding the intensive treatment applied to both follicular and papillary cancers.

The situation is complicated further by the presence of a follicular variant of papillary thyroid cancer. The first four modes explained more than 45% of the variability (Fig. 32.7). The genes selected on the basis of mode correlation coefficients were very interesting from a biological point of view, and hierarchical clustering revealed distinct differences between tumors of both histological subtypes. However, the same conclusions were obtained by the analysis of array characteristic modes. Two-dimensional analysis of the first two modes (Fig. 32.8) revealed that tumors cluster not according to the difference between benign and malignant (which was expected from the difference in their clinical behavior), but according to the morphological distinction between papillary and follicular features. All follicular adenomas and carcinomas except one had positive values of the first mode coefficient, while all classic papillary tumors except

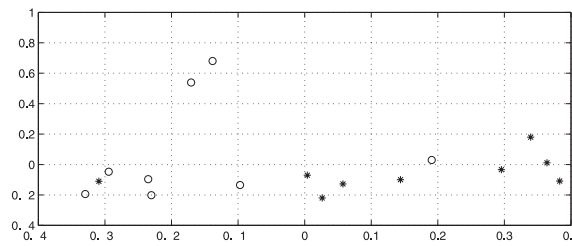


**Fig. 32.7.** Hierarchical clustering of thyroid cancer specimens of different histopathological subtype. Follicular cancers (F) and adenomas (A) are clustering together, while papillary cancers (not labelled) are within the other branch of dendrogram. Follicular variant of papillary cancer (V) samples are found within both subgroups.

two had negative coefficients. Benign and malignant follicular tumors could not be distinguished based on the gene expression profile. A very interesting category was the follicular variant papillary tumors: they were interdispersed between follicular tumors and papillary cancers, some of them with negative and some with positive coefficients. It is now a matter of debate and thorough analysis, whether this group is heterogeneous in biological nature or whether the histopathological criteria used to classify them are not adequate.



**Fig. 32.8.** Analysis of array characteristic modes shows distinction between follicular tumors (adenomas marked by circles, carcinomas by plus signs) and papillary tumors (papillary cancer marked by dots, follicular variant of papillary cancer denoted by x marks).



**Fig. 32.9.** Analysis of array characteristic modes in laryngeal cancer helps to detect outliers in gene expression profile: two mislabelled samples classifying in the inappropriate class of samples and two normal samples with distinctly different second mode coefficient values.

### 32.3.4 Detecting Outliers in Gene Expression Profile

SVD is a very powerful unsupervised method for detecting outliers in gene expression profiling experiments. A good example is our study of the gene expression profile in laryngeal cancer.

By analysis of array characteristic modes we found that two samples are probably mislabelled (one tumor clustered with normal tissues and one normal tissue with tumors), and two normal samples were distinctly different from all other samples within the second mode (Fig. 32.9). Using supervised methods of gene selection on the whole dataset we were unable to detect any genes significantly differentiating tumors and normal tissues. After exclusion of detected outliers, we determined a number of genes to have biological meaning, and we are further validating their significance.

## 32.4 Conclusions

Singular value decomposition is a reliable mathematical tool for revealing the main sources of variability in analyzed microarray datasets. When followed by a gene selection procedure based on *gene characteristic mode* coefficients, it is also a robust technique to provide biological interpretation of observed variability. Calculation and analysis of the *array characteristic modes* allow easy detection of outlier samples in microarray data.

## Acknowledgments

This work was supported in part by the Polish Committee of Scientific Research, grant 4T11F 018 24 in 2005, and grant 3P05B 112 25, 2004–2005.

## References

1. Alter, O., Brown P.O., Botstein, D.: Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*, **97**, 10101–10106 (2000).

2. Wall, M.E., Rechtsteiner, A., Rocha, L.M.: Singular value decomposition and principal component analysis. In: Berrar, D.P., Dubitzky, W., Granzow, M. (eds.) *A Practical Approach to Microarray Data Analysis*. Kluwer Academic Publishers, Boston Dordrecht London (2003).
3. Holter, N.S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J.R., Fedoroff, N.F.: Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc. Natl. Acad. Sci. USA*, **97**, 8409–8414 (2000).
4. Golub, G.H., van Loan, C.F.: *Matrix Computations*. Johns Hopkins University Press, Baltimore (1996).
5. Simek, K., Kimmel, M.: A note on estimation of dynamics of multiple gene expression based on singular value decomposition. *Mathematical Biosciences*, **182**, 183–199 (2003).
6. Simek, K., Fajarewicz, K., Swierniak, A., Kimmel, M., Jarzab, B., Wiench, M., Rzeszowska, J.: Using SVD and SVM methods for selection, classification, clustering and modeling of DNA microarray data, *Artificial Intelligence. Engineering Application of AI*, **17**, 417–427 (2004).
7. Jarzab, B., Wiench, M., Fajarewicz, K., Simek, K., Jarzab, M., Oczko-Wojciechowska, M., Wloch, J., Czarniecka A., Chmielik, E., Lange, D., Pawlaczek, A., Szpak, S., Gubala, E., Swierniak, A.: Gene expression profile of papillary thyroid cancer: sources of variability and diagnostic implications. *Cancer Research*, **65**, 1587–1597 (2005).
8. Olbryt, M., Jarzab, M., Jazowiecka-Rakus, J., Simek, K., Sochanik, A.: Gene expression profile in B16 (F10) murine melanoma cells in vitro under hypoxic conditions, paper submitted to *Cancer Research*.
9. Jarzab, B.: personal communication.
10. Markowski, J.: personal communication.