Analytical Cellular Pathology 24 (2002) 167–179
IOS Press

Chromosomal regions in prostatic carcinomas studied by comparative genomic hybridization, hierarchical cluster analysis and self-organizing feature maps

Torsten Mattfeldt [a,*], Hubertus Wolter [a], Danilo Trijic [a], Hans-Werner Gottfried [b] and Hans A. Kestler [c]
[a] *Department of Pathology, University of Ulm, Ulm, Germany*
[b] *Department of Urology, University of Ulm, Ulm, Germany*
[c] *Department of Neuroinformatics, University of Ulm, Ulm, Germany*

Abstract. Comparative genomic hybridization (CGH) is an established genetic method which enables a genome-wide survey of chromosomal imbalances. For each chromosome region, one obtains the information whether there is a loss or gain of genetic material, or whether there is no change at that place. Therefore, large amounts of data quickly accumulate which must be put into a logical order. Cluster analysis can be used to assign individual cases (samples) to different clusters of cases, which are similar and where each cluster may be related to a different tumour biology. Another approach consists in a clustering of chromosomal regions by rewriting the original data matrix, where the cases are written as rows and the chromosomal regions as columns, in a transposed form. In this paper we applied hierarchical cluster analysis as well as two implementations of self-organizing feature maps as classical and neuronal tools for cluster analysis of CGH data from prostatic carcinomas to such transposed data sets. Self-organizing maps are artificial neural networks with the capability to form clusters on the basis of an unsupervised learning rule. We studied a group of 48 cases of incidental carcinomas, a tumour category which has not been evaluated by CGH before. In addition we studied a group of 50 cases of pT2N0-tumours and a group of 20 pT3N0-carcinomas. The results show in all case groups three clusters of chromosomal regions, which are (i) normal or minimally affected by losses and gains, (ii) regions with many losses and few gains and (iii) regions with many gains and few losses. Moreover, for the pT2N0- and pT3N0-groups, it could be shown that the regions 6q, 8p and 13q lay all on the same cluster (associated with losses), and that the regions 9q and 20q belonged to the same cluster (associated with gains). For the incidental cancers such clear correlations could not be demonstrated.

Keywords: Artificial neural networks, bioinformatics, chromosome aberrations, cluster analysis, comparative genomic hybridization, multivariate analysis, prostatic cancer, self-organizing maps

1. Introduction

Modern molecular biological methods may produce large amounts of data which are difficult to survey. This statement applies particularly to gene array techniques, where the expression of thousands of genes may be measured. Here the problem may arise to find clusters of genes which behave in a similar manner [36,40]. To a smaller extent, analogous problems are also found during evaluation of comparative genomic hybridization (CGH) data [30].

CGH is a method which allows screening of the whole genome for gains and losses of the genetic material. Genomic DNA of tumor tissue as well as the DNA of normal tissue are isolated, differentially stained and hybridized to normal metaphase chromosomes. When the tumor DNA is stained green and the reference DNA is stained red, for example, this leads to a green stain at locations with a gain of tumor DNA, whereas a red stain is obtained at losses of tumor DNA because here the normal DNA dominates. The results are quantitated by digital image analysis. This leads to a series of ratio profiles for the 24 chromosomes (22 auto-

*Corresponding author: Prof. Dr. T. Mattfeldt, Department of Pathology, Oberer Eselsberg M23, D-89081 Ulm, Germany. Fax: ++49 731 58738; E-mail: torsten.mattfeldt@medizin.uni-ulm.de.

somes and 2 sex chromosomes). For convenience during this paper each chromosome arm was taken as a unit. Since the short arm of the acrocentric chromosomes and the sex chromosomes are uninformative, and 5 regions showed no alteration in any group, 34 chromosome arms as chromosomal regions were taken into account during this analysis. For each chromosome arm one of the alternatives 'unchanged', 'loss' or 'gain' (or equivalently 1, 0 or 2) is noted. In short, one case is reduced (theoretically) to a matrix of the size 2×34, in which each element can assume the value 0, 1 or 2.

This definition has the following basis. The immediate finding of CGH for a certain chromosomal region is usually expressed as a rational number, i.e., the fluorescence ratio r of tumour DNA to normal DNA. It is recommended to classify a finding as loss when $r < 0.8$, as normal when $0.8 \leqslant r \leqslant 1.25$, and as gain when $r > 1.25$ [5,8]. Our numerical values for loss, normal state and gain are thus obtained as step function $y(r) = 0$ for $r < 0.8$, $y(r) = 1$ for $0.8 \leqslant r \leqslant 1.25$, and $y(r) = 2$ for $r > 1.25$ on the basis of these thresholds. This function definition was used to express equal weights for losses and gains. The values thus represent equally strong deviations from the norm for loss and gain, as the Euclidean distance of both values to 1 amounts to 1. Using this step function integer values are obtained from primarily continuous data, which results in an ordinal scale. The statistical methods described below are not restricted to continuous variables but may be applied to integer values as well (for more details and examples see [6,23,24]).

Our task consists in the formation of a certain number of groups (clusters), to which the chromosomal regions are assigned in a biologically meaningful manner. This task has to be fulfilled without knowing further variables, which is usual for clustering methods, solely on the basis of the CGH data. The present paper is an example to achieve this for prostatic cancer, which has been intensively studied by CGH [1,2,15,31, 32]. Cluster analysis can be used to assign similar individual cases (samples) to different clusters of cases, where each cluster may be related to a different tumour biology [3,30]. Another approach consists in the clustering of chromosomal regions. The idea of clustering variables instead of cases (samples) is a classical option in cluster analysis in general, based on a transposition of the original data matrix [17]. While in sample clustering the cases are written as rows and the variables as columns, the variables are now written as rows and the cases as columns. Recently this approach has been widely applied to gene expression data [11]. Both approaches can be coupled in the same study [4,12].

The required grouping can be principally obtained by all kinds of clustering techniques. For example, hierarchical cluster analysis, k-means, fuzzy c-means and other techniques can be used. Here we concentrate on two implementations of an artificial neural network developed by Kohonen, the self-organizing map: SOM (Kohonen network) [24,25,45] in comparison with a classical hierarchical cluster analysis. Recently such neural networks were successfully used for cluster analysis in gene expression [40]. Our group has recently used a SOM for sample cluster analysis of CGH data [30], and we have applied related networks with a supervised learning rule for predictive purposes in prostate carcinoma research [25,28,29].

2. Materials and methods

2.1. Patient population

Group I. The archive of the Department of Pathology of the University of Ulm from 1990–99 was searched for all patients from the Department of Urology of the University with incidental prostatic cancer (tumour category T1). On the whole, this included 66 cases (resection specimens and adenomectomies), removed because of benign prostatic hyperplasia, and in which a prostatic carcinoma was incidentally found. Incidental carcinomas within cystoprostatectomy specimens were excluded from the study. The current TNM classification according to the UICC was used, and the series included cases in categories T1a and T1b [37,43]. All cases from which technically acceptable CGHs could be obtained, were selected for the study (43 TUR-specimens and 5 adenomectomies).

Group II. This material consisted of 50 prostatectomy specimens with preoperatively diagnosed prostate carcinomas. The pTNM classification was pT2N0 [37]. Small tissue blocks of tumor material and normal tissue from the same patient were flash-frozen in liquid nitrogen immediately after surgical removal. Five μm sections were cut from freshly frozen tumor and normal tissue blocks and stained with hematoxylin and eosin to ensure the histological representativeness of the samples. Based on microscopic evaluation the tumor region was selected and removed for DNA extraction with a scalpel [35]. For DNA isolation we used the Qiagen-Blood & Cell Culture-Kit (Qiagen GmbH, Hilden, Germany), following the instructions of the supplier.

Group III. This material consisted of 20 prostatectomies with prostate carcinomas. The pTNM classification was pT3N0. Preparation and DNA-isolation were performed in the same manner as described for group II.

Pathology. The specimens removed with incidental carcinomas and the prostatectomy specimens were step sectioned at 3–5 mm slice thickness. In prostatectomy specimens, at least 2 additional sections from the resection margins and at least 1 additional section from each seminal vesicle were taken. The tumor-bearing slides of all cases were reevaluated by the first author.

2.2. Comparative genomic hybridization

CGH was performed as described previously [9,18, 19,42,43] with minor modifications. We currently used standard nick translation for labeling genomic DNAs with biotin-16-dUTP (tumor DNA) and digoxigenin-11-dUTP (normal DNA). The fragment length of our genomic probes after nick translation was 500 to 1600 bp. One μg of labeled tumor and normal DNA were precipitated together with 70 g Cot-1 DNA. The mixture was dissolved in 12 μl of hybridization buffer: 50% formamide, 10% dextran sulfate, and $2 \times$ SSC ($1 \times$ SSC: 0.15 M NaCl, 0.015 M sodium citrate, pH 7). This hybridization mixture was denatured at 75°C for 6 minutes and pre-annealed for 30 minutes at 37°C and hybridized to a slide with normal male lymphocyte (46,XY) metaphase spreads, denatured separately in a formamide solution (70% formamide, $2 \times$ SSC, pH 7–7.2) for 2 minutes at 70°C and dehydrated through a series of graded solutions of 70%, 90% and 100% ethanol. The hybridization was performed for 2 to 3 days at 37°C in a moist chamber. After hybridization, the slides were washed 3 times in 50% formamide, $2 \times$ SSC, pH 7–7.2 for 5 minutes at 42°C and 3 times in $0.1 \times$ SSC, pH 7–7.2 for 5 minutes at 60°C. The tumor DNA was detected with a single layer of avidin-conjugated fluorescein isothiocyanate (FITC) (Vector Laboratories, Inc., Burlingame, CA), and the normal DNA was detected with anti-digoxigenin antibody conjugated to rhodamine (Roche Diagnostics GmbH, Mannheim, Germany) for 45 minutes at 37°C. The chromosomes were counterstained with 4,6-diamidino-2-phenylindole (DAPI) (Sigma-Aldrich Chemie GmbH, Steinheim, Germany) and embedded in antifade solution (Vector Laboratories, Inc., Burlingame, CA).

2.3. Digital image analysis

Three single-color images (matching DAPI = blue, FITC = green and rhodamine = red) were acquired from 15–20 metaphases using a Zeiss fluorescence microscope (Carl Zeiss, Oberkochen, Germany) and a Hamamatsu chilled charge-coupled-device (CCD) camera (Hamamatsu Photonics K.K., Tokyo, Japan) interfaced to a computer workstation. The selection of metaphases for CGH analysis was based on quality control criteria as known from the literature [20, 22]. The ISIS digital image analysis system (Metasystem GmbH, Altlussheim, Germany) was used with CGH analysis software (Version 3.02). Fluorescence ratio (green : red) for each chromosome type were derived for these metaphase cells. All of the ratio profiles from a chromosome type were averaged, and the standard deviation of the profile set was calculated at each point. For all the profiles, losses of DNA sequences are defined as chromosomal regions in which the mean green : red ratio is below 0.8 whereas gains are defined as chromosomal regions in which this ratio is above 1.25. These threshold values are symmetric cutoff values, 1.25 and its reciprocal value, 0.8 (see above). Interpretation of CGH-results followed previously described protocols [20]. Hybridization of two differentially labeled mismatched normal DNAs were used as a control experiment for each batch of hybridization. Heterochromatin blocks such as the distal long arm of the Y chromosome or the centromeres, and the near-centromere heterochromatic regions of chromosomes 1, 9, 16 were excluded from CGH analysis as well as centromeres and short arms of the acrocentric chromosomes (13p, 14p, 15p, 21p and 22p). Furthermore, chromosomal regions that showed no aberrations in any case (6p, 10p, 11p, 12p and 15q), were considered as uninformative.

2.4. Hierarchical cluster analysis

Hierarchical agglomerative cluster analysis is a classical deterministic method to find clusters in an n-dimensional space of data points (input vectors). Clusters are found among the data points according to the interpoint and intercluster distances (or squared distances). How these are measured in detail depends on the specific algorithm applied (see, e.g., [13]). The algorithm of Ward, to which the property is ascribed to provide ‘good and homogeneous’ clusters [39], was used to select a specific clustering from the dendrogram. Hierarchical cluster analyses were performed

in series of 1–15 clusters per case group. The mean quadratic distance between the data points and their corresponding cluster centre (quantization error) diminished for an increasing number of clusters. This behaviour does not necessarily indicate better clustering as the complexity of the system increases. In the extreme case, one could use the same number of cluster points as data points and obtain an error of zero. The number of clusters generated is to a certain degree arbitrary. However, plausible hints can often be obtained from the dendrogram [39]. For example, in the present study, the dendrograms appeared consistent with three major clusters of variables per case group (Fig. 1). We used the implementation of hierarchical cluster analysis by SAS [34].

2.5. *Self-organizing feature maps (SOM)*

Artificial neural networks (ANN) are information processing systems consisting of a number of units (neurons), communicating with each other through connections. Such systems 'learn' by processing external information adapting to a learning rule. They are classified into ANNs with supervised learning and with unsupervised learning. In most applications to biological material so far, networks with supervised training have been used. In our recent paper on case cluster analysis on the basis of CGH data we showed that also unsupervised networks can provide useful information from biological data [30]. Here we want to extend this view to the clustering of chromosomal regions.

Self-organizing maps (SOMs) belong to the ANNs with an unsupervised learning rule. To such networks only input vectors (input data, input information) are presented, and no output vectors. In this application, the input vectors are simply the CGH data. The task of a typical SOM consists in finding clusters of the input vectors, with similar vectors in the same clusters. A short nontechnical introduction to SOMs can be found in [40], whereas the subject is presented in depth in [24]. Briefly, the fundamental structure of a SOM is a layer of neurons with a simple geometric shape, e.g., a rectangle or a line (chain) in the plane. These neurons are connected to weight vectors, that lie in the n-dimensional space of the input vectors. These basic active neurons are called Kohonen neurons, the layer is the Kohonen layer. During the learning process the weight vectors are moved in the n-dimensional space until they have moved as close as possible to the input vectors. That neuron whose weight vector has come nearest to an input vector is called a winner neuron.

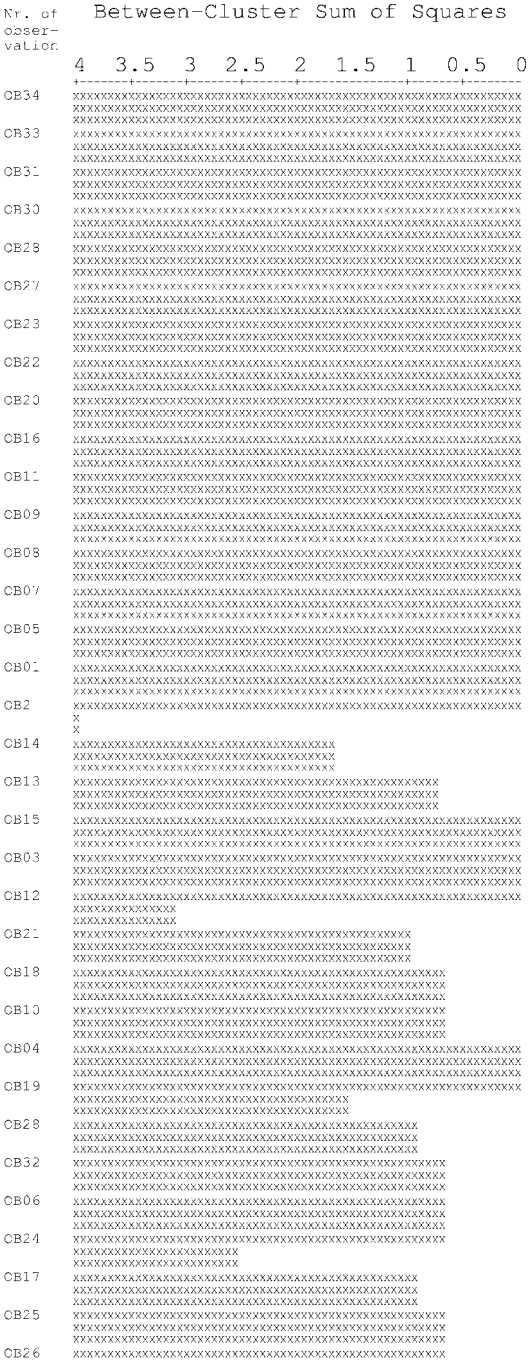


Fig. 1. Dendrogram of 34 chromosomal regions of group I (incidental prostate carcinomas), obtained by hierarchical cluster analysis by SAS-software using the algorithm of Ward. Vertical axis: number of chromosomal region extending from 1 (region 1p) to 34 (region 22q), compare Table 3. Horizontal axis: Sum of squares between clusters. The upper part of the dendrogram consists of chromosomal regions with no difference between them (a homogeneous cluster of normal chromosomal regions). In the lower part at least two further clusters are found. Similar results were found for group II and III.

On the whole, the learning process has the effect that properties of the n-dimensional input vectors are transferred to the low-dimensional space of the Kohonen neurons ($d = 1$ or 2). In particular, input vectors lying close to each other in space will generate a cluster in the Kohonen layer.

The classical implementation of a SOM is the package SOMPAK [24,26]. At http://www.cis.hut.fi/research/som-research/nnrc-programs.shtml it can be obtained as free academic software by internet. The programs are available as a set of source files for Linux or DOS, and as binary (executable) files for Windows. For our study only the implementation of SOM under Linux was used. The input variables are fed as ASCII file into the system, and the user has to enter a number of system parameters such as the number of neurons (nodes) in the Kohonen plane, the neighbourhood function and others. As result, the program provides the x, y-coordinates of the winner neurons.

An alternative to SOMPAK is the recently published SOM Genecluster [40]. The input data are also given as an ASCII data set. For the calculations only very few parameters have to be indicated to the system, most parameters are preset to standard values. The result consists in tables where each data point is assigned to one cluster; here we selected to ascribe the data to 1–25 clusters for first exploration, and concentrated on 3 clusters finally. Genecluster is free academic software and runs under Windows NT. It can be obtained by internet under http://genome.wi.mit.edu.

Attempts were made to compare the clustering results obtained by HCA and the two implementations of SOMs by statistical evaluation. The classifications of the same set of data by two clustering methods into n clusters can be represented as a contingency table of size $n \times n$, where the number of data (here: chromosomal regions) in the ith cluster according to method 1 and simultaneously in the jth cluster according to method 2 is $B_{i,j}$. The null hypothesis means here that the number of items B_{ij} for $i \neq j$ is symmetric to the main diagonal of the table. This test of symmetry can be seen as an extension of the well-known McNemar test for 2×2 tables to $n \times n$ tables [7,10,33,38]. Before the test was performed, the clusters were ordered solving the linear assignment problem [27]. The test statistic was estimated using the equation:

$$\hat{\chi}^2 = \sum_{i=1}^{j-1} \sum_{i>j} \frac{(B_{ij} - B_{ji})^2}{B_{ij} + B_{ji}}$$

with $n(n-1)/2$ degrees of freedom. The resulting value $\hat{\chi}^2$ itself may be considered as a measure for the dissimilarity of the two compared methods [33]. In practice, the clustering results of our 34 chromosome regions as provided by hierarchical cluster analysis, were compared with those provided by SOMPAK and Genecluster by this test. No difference between the methods was detected at a significance level of $p < 0.05$. Hence we decided to present the output of GeneCluster only and to omit the results of the other programs to save space. Also the quantization error decreased strongly with rising number of clusters when using the two types of SOMs, as reported above for hierarchical cluster analysis.

3. Results

3.1. Basic findings

The basic findings in the three groups are summarized in Table 1. There is a significant increase of mean

Table 1
Basic data of group I–III

	Group		
	I	II	III
Stage	T1	pT2	pT3
Number of cases	48	50	20
Mean number of losses per case	0.33 ± 0.88	0.90 ± 1.53	2.00 ± 2.96
Mean number of gains per case	0.46 ± 1.11	0.58 ± 1.26	1.50 ± 1.67
Mean Gleason score	5.06 ± 1.73	5.60 ± 1.09	7.25 ± 1.29
Mean WHO grade	1.69 ± 0.56	1.88 ± 0.42	2.43 ± 0.40
Percentage of cases with aberrations	29	46	70

In this table mean values with standard deviations are presented for our three case series. There is a monotonous increase of the number of losses and gains from T1 carcinomas to pT2N0 to pT3N0 cases. The percentage of cases with aberrations also rises monotonously from T1 to pT2N0 to pT3N0. Such an increase appears to apply also for Gleason score and WHO grade at first sight, however for the grading systems the difference is significant for pT3N0 versus pT2N0, but not for pT2N0 versus T1 (see Discussion).

Gleason score and mean WHO grade from pT2N0- to the pT3N0-carcinomas. Also we found a rising fraction of the number of cases with aberrations detectable by CGH per total number of cases from T1 to pT2N0 to pT3N0. Parallel to this global increase of aberrations, the mean number of losses and of gains per case increased with tumour stage. Summarizing, in our cases the histopathological grading parameters as well as the genetic aberrations detected by CGH rose with increasing tumour stage.

Table 2 shows the original data matrix of group III: cases with stage pT3N0. It is a rectangular matrix with 20 rows, representing the cases (patients) 1–20, and 34 columns representing the 34 chromosome regions 1p–22q as shown in the three lines above the table. Each number is a result of a CGH examination, where a '1' represents a normal content of chromosomal material, a '0' is a loss of chromosomal material in that region, and a '2' indicates a gain of chromosomal material there. One can apply directly some method of cluster analysis to such a matrix, in that case a clustering by cases (samples) is performed [11,30]. Each cluster consists of a number of patients, who can be further characterized in terms of clinical parameters.

3.2. *Cluster analysis of chromosome regions from CGH data*

Another idea of clustering is known from data analysis in gene expression [11,21,40]. The data matrix in many gene expression studies is a rectangular matrix of rational numbers indicating intensity of expression, where the rows indicate genes and the columns represent the samples, e.g., different patients or different time points in an experiment. By transposition of the matrix, one can either obtain a study for sample clustering or gene clustering [11]. In our present study we have CGH data relating to chromosomal regions and not genes, and natural instead of rational numbers, but otherwise the structure of the data is analogous.

The data for this analysis were obtained by transposition of Table 2, but this matrix was omitted here

Table 2

Original data matrix

	1 p	1 q	2 p	2 q	3 p	3 q	4 p	4 q	5 p	5 q	6 q	7 p	7 q	8 p	8 q	9 p	9 q	10 q	11 q	12 q	13 q	14 q	16 p	16 q	17 p	17 q	18 q	19 q	20 q	21 q	22 q
1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	2	1	1
3	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1	1	1	1	0	2	1	2	1	1	2	1	1	1	1	1	1	1	1	1	2	1	1
7	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1	1	2	2	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1
9	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
10	1	0	1	1	1	2	1	0	1	1	0	1	1	0	2	1	1	1	1	1	0	1	1	0	1	1	2	1	1	1	1	1
11	1	1	1	1	1	1	1	1	1	0	0	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1
14	1	1	0	1	1	1	0	1	1	1	1	2	1	0	2	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
15	0	0	1	0	1	1	1	0	2	0	0	1	1	0	0	1	2	1	1	1	0	1	1	0	1	1	0	0	1	1	1	1
16	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	1	1	1	1	2	1	1
19	1	1	1	1	1	1	1	2	0	1	2	2	0	1	1	2	0	0	1	0	1	1	1	1	1	1	1	1	1	1	1	1
20	1	1	1	1	1	2	1	1	1	1	0	2	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

The original data matrix of group 3 (20 pT3N0 cases). Each row represents the CGH data of 34 chromosome arms of one case, each column indicates the findings at the chromosome region listed at the top. In this form the data matrix is suitable for a cluster analysis of samples. Note the unsupervised approach with merely input vectors in contrast to supervised methods for classification, such as multilayer feedforward networks and learning vector quantization.

to save space. After transposition the chromosomal regions are represented as 34 rows and the case numbers represent the 20 columns. If a value in the ith row and in the jth column is denoted as o_{ij} in the original matrix and as t_{ij} in the transposed matrix, then $t_{ij} = o_{ji}$ and $t_{ji} = o_{ij}$. Now a cluster analysis applied to this matrix means clustering by chromosomal regions. While the aforementioned study by samples attempted more at the identification of different case groups, the present investigation is directed primarily towards a better understanding of losses and gains of chromosomal material at certain regions and the interrelations between them.

All studies in groups I–III were performed with hierarchical cluster analysis and the SOMs Genecluster and SOMPAK. There were 34 input variables each, corresponding to all chromosome arms that could be studied by CGH and in which at least one imbalance had occurred. In the following learning process, the SOM gets information solely on the input variables. When operating with the SOMs, standard settings were used. The number of clusters was set to 3 corresponding to a chain of 3 units length. Very high numbers of clusters are not informative for small sample sizes. Furthermore it is to a certain extent *a priori* plausible to divide cancer cases into three clusters, which may be understood as low, intermediate and high degrees of malignancy [30]. Plots of dendrograms of the hierarchical cluster analysis were also consistent with three clusters (Fig. 1). The number of iterations per run was set to 5000, as thereafter one could not see significant changes of the error. As the neighbourhood function of the SOM the step function (bubble) was used. The initial and final learning rate values were $\alpha_i = 0.1$ and $\alpha_f = 0.005$, and the initial and final radius values of the step function were $r_i = 5$ and $r_f = 0.2$. The map was initialized using random vectors. We performed 10 repeated runs for each network. The results were mostly identical within repetitions, and sometimes there was one clustering result different from the others (see also [30]). The programs were applied to the groups I, II and III separately (no pooling), and the outcome is shown in Tables 3, 4 and 5, respectively.

In Table 3, we see the results of clustering the 34 selected chromosomal regions of 48 patients with incidental carcinomas evaluated with Genecluster. The table is sorted by the three clusters indicated by the numbers 1, 2 and 3. The first cluster consists only of chromosomal regions with at least one gain. Only one region shows losses (region 8p with 5 losses). The second and largest cluster consists mainly of domains with no genetic aberrations at all, plus 4 regions with low numbers of losses and gains. In the third cluster we find regions all of which show losses. In general, the system has generated three clusters which may be characterized as nearly normal, rich in gains, and rich in losses.

In Table 4 (pT2N0-cases) the clustering achieved is similar, however the succession of the clusters is reversed on the Table. The first cluster consists of 6 regions with losses, in some of these regions infrequent gains could also be found. Again the system found a large cluster of cases with normal genome plus a few cases with a low number of losses and gains (cluster 2). The last cluster contains regions with a rather high number of gains. The most frequent losses occur in cluster 1 on 13q, 8p and 6q. The most frequent gains are found in cluster 3 on 17q, 20q and 9q.

Finally, Table 5 shows the results for the pT3N0 cases. The first cluster consists of regions with gains only. The second cluster contains normal regions and regions with a single loss and 1–3 gains. The third cluster consists of regions with losses only. Clearly the percentage of regions without alterations is lowest in this group. The most frequent losses are localized on regions 8p, 6q, 13q and 18q. The most frequent gains are found on 7p, 8q, 9q and 20q.

4. Discussion

In the present paper, we have applied self-organizing maps for the first time to perform cluster analysis of chromosomal regions studied by comparative genomic hybridization. The small datasets used here (20–50 cases) are typical for CGH, because evaluation of a single case is rather laborious. In order to make SOMs for CGH popular, we used also the program ‘Genecluster’, equipped with a graphical interface of Windows-type and preset parameters (which can be changed if desired, nevertheless). The comparative studies with two SOMs and hierarchical cluster analysis gave no evidence for significant differences between methods in clustering the chromosome regions when the same numbers of clusters were used. This view can be strictly only stated for the present data sets from prostate cancer cases. In other papers it has been reported that different results may be obtained when hierarchical or neuronal methods are used, in particular when the data sets are noisy [16,40,44].

We have applied cluster analysis for the second time to CGH data from prostate carcinomas. The clustering itself was performed by means of the SOM Geneclus-

Table 3
Results of cluster analysis in Group I

Region	Cluster	Number of losses	Number of gains
2p	1	.	1
7p	1	.	1
7q	1	.	2
8p	1	5	1
8q	1	.	1
9q	1	.	3
17p	1	.	4
17q	1	.	6
1p	2	.	.
1q	2	.	.
3p	2	.	.
3q	2	.	1
4p	2	.	.
4q	2	.	.
5p	2	.	.
6q	2	.	.
9p	2	.	.
12q	2	.	.
14q	2	.	.
16p	2	.	.
16q	2	1	.
18p	2	.	.
18q	2	2	.
19p	2	.	.
19q	2	.	.
20p	2	.	.
20q	2	.	1
21q	2	.	.
22q	2	.	.
2q	3	1	.
5q	3	2	.
10q	3	1	1
11q	3	1	.
13q	3	3	.

Here the result of cluster analysis with the SOM Genecluster for group 1 (incidental carcinomas) is presented. The data are now sorted by clusters numbered from 1–3. Cluster 1 has only one region with losses, in each region has occurred at least one gain. Cluster 2 has many regions which show no deviation from the norm according to CGH, and a few regions with 1–2 losses and gains. There is no region with a combination of gains and losses, however. Cluster 2 can thus be characterized as normal plus a few regions with a low number of isolated losses or gains. In cluster 3 we have only domains with one or more losses, in one region in combination with a single gain.

Table 4
Results of cluster analysis in Group II

Region	Cluster	Number of losses	Number of gains
5q	1	3	.
6q	1	5	.
8p	1	9	.
13q	1	11	1
16q	1	4	1
18q	1	3	.
1p	2	2	.
1q	2	.	.
2p	2	.	.
2q	2	1	.
3p	2	.	.
3q	2	.	1
4p	2	.	.
4q	2	.	.
5p	2	1	.
7p	2	.	.
7q	2	.	1
9p	2	1	.
10q	2	1	1
11q	2	.	.
12q	2	.	.
14q	2	1	.
16p	2	.	.
18p	2	.	.
19p	2	.	.
19q	2	.	.
20p	2	2	1
21q	2	.	1
22q	2	.	.
8q	3	1	3
9q	3	.	5
17p	3	.	3
17q	3	.	6
20q	3	.	5

The tabulated result of cluster analysis with Genecluster for group II (pT2N0 carcinomas) is shown. The data are sorted by clusters numbered from 1–3, and cluster 2 consists of domains without aberrations or with maximally one isolated loss or gain detectable by CGH. The other clusters have similar characteristics as in group I, but here the succession is reversed: cluster 1 at the top is rich in regions which show losses and contain no or maximally 1 gain, whereas cluster 3 has regions with multiple gains and no more than 1 loss. On the whole the table shows only 74 aberrations in 50 cases, which is presumably due to the rather limited tumor stage with restriction of the tumor cells to the prostate gland and lack of lymph node metastases.

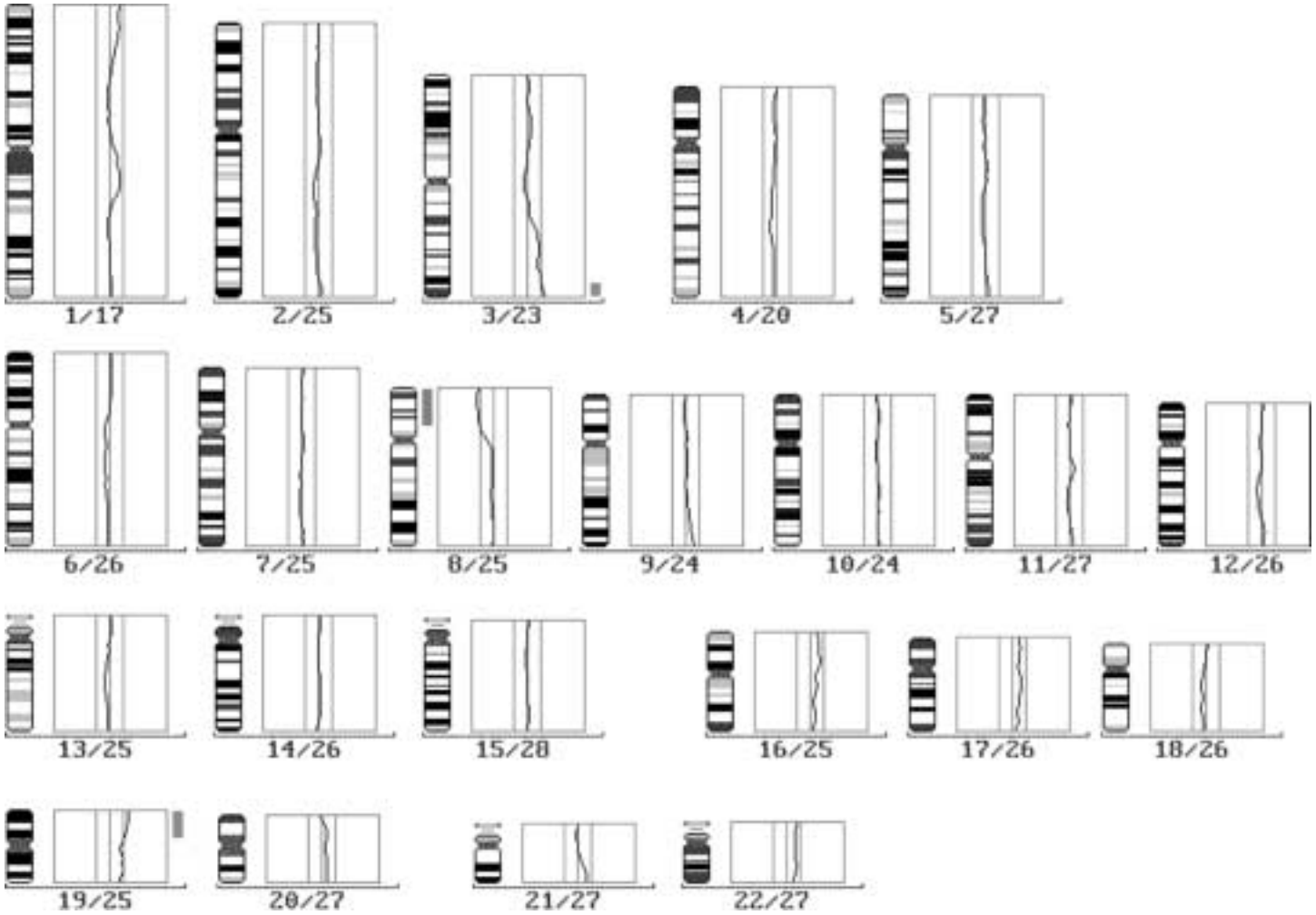
Table 5
Results of cluster analysis in Group III

Region	Cluster	Number of losses	Number of gains
5p	1	.	2
7p	1	.	4
7q	1	2	2
8q	1	1	4
9q	1	.	3
12q	1	.	2
20p	1	.	1
2p	2	1	.
3p	2	.	.
3q	2	.	2
4p	2	1	.
9p	2	.	.
10q	2	1	.
11q	2	1	.
14q	2	.	.
16p	2	.	.
17p	2	1	2
17q	2	.	2
18p	2	1	1
19p	2	.	.
19q	2	.	.
20q	2	.	3
21q	2	.	.
22q	2	.	.
1p	3	1	.
1q	3	2	.
2q	3	2	.
4q	3	2	.
5q	3	6	.
6q	3	4	.
8p	3	7	.
13q	3	3	.
16q	3	3	2
18q	3	4	.

Cluster analysis of the 20 pT3N0 cases shows basically the same results as for the lower tumor categories. However, the number of chromosome regions in cluster 2, with predominantly normal and slightly changed domains, has diminished, whereas the number of regions in cluster 1 and 3 with multiple gains and losses, respectively, have considerably increased. This behaviour obviously results from the increased frequency of aberrations and loss of differentiation with increasing stage (see Table 1).

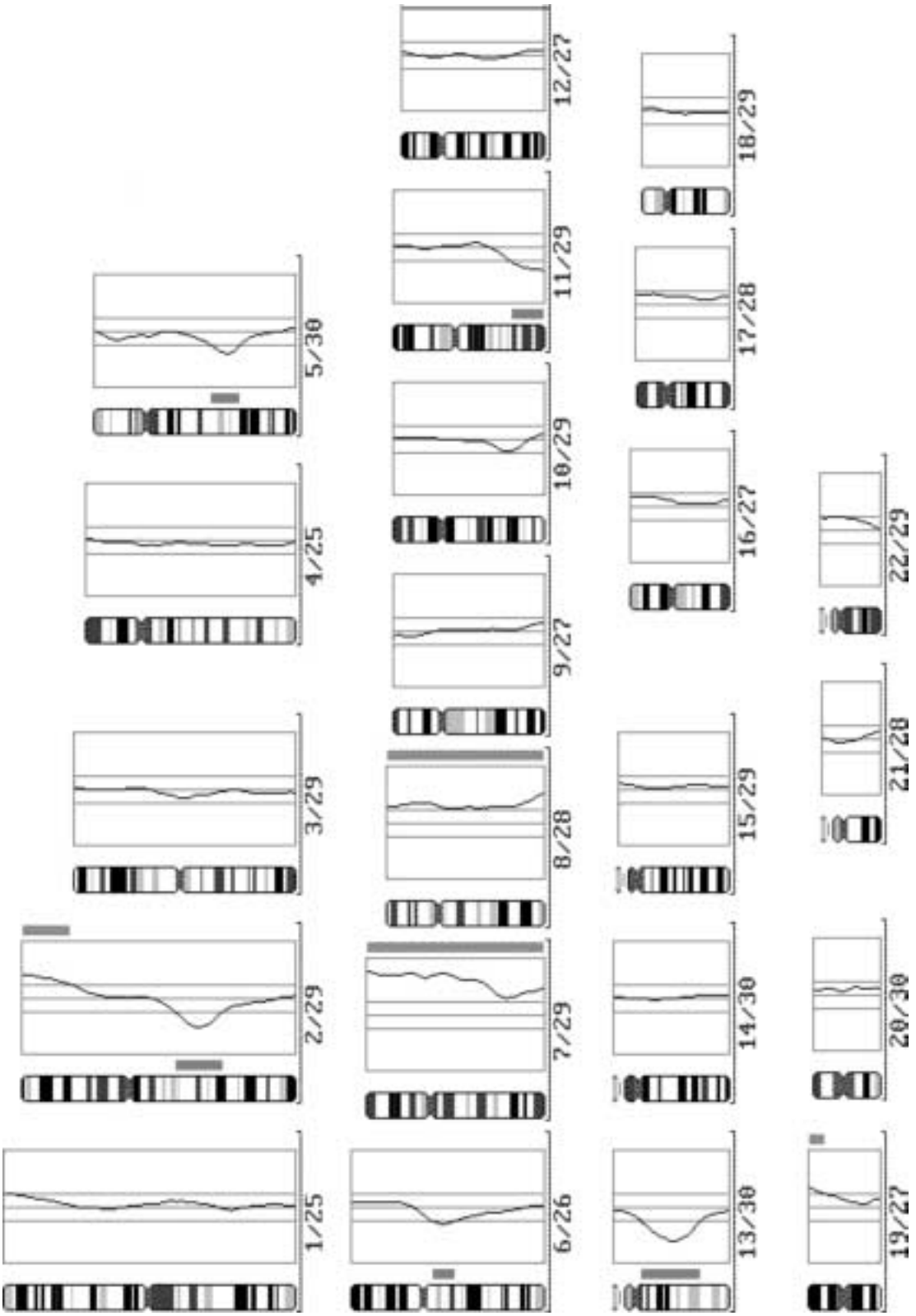
ter, the well-known SOMPAK package and hierarchical cluster analysis. In contrast to the previous investigation, where cases were clustered, we have now performed a clustering of changes of chromosomal regions. Both approaches have different goals. In case clustering the individual case is the unit, of which usually many other clinical data are known, such as histopathological grade, PSA values, tumour progression and many others. Hence mean values of these data can be computed for the clusters, and the usefulness of the clustering can be judged from the homogeneity of the individual clusters and the contrasts between the clusters. Clustering of chromosomal regions as performed here is different inasmuch as series of chromosome domains of *all* cases contribute to *all* clusters. It is however possible to find relations between different chromosomal regions and the total number of aberrations per region. The strategy of our paradigms has been to construct three rather homogeneous clusters in all three groups: a cluster of normal or nearly normal chromosomal arms, a cluster of regions with predominant losses, and a cluster of regions where gains predominate. The fraction of these clusters is influenced by the tumor category: in T1 carcinomas the normal group is the largest, in the pT3N0 group the fraction of normal cases is much lower because we have more aberrations in these advanced cases, and the pT2N0-group behaves intermediate. While our results were obtained by genuine multivariate techniques such as HCA and SOMs, this excludes by no means that similar relations could have been obtained by other approaches which may be considered as simpler. For example, it is possible to estimate partial correlation coefficients between the variables, or to estimate contingency tables. By this approach it may be possible to identify groups of variables (clusters in a wide sense) that are significantly correlated to each other, but not to the other groups. An instructive example for this approach with graphical presentation of results is given in the literature [39].

Comparison of the three tumour categories is consistent with the general hypothesis that higher stages of tumours are correlated with higher numbers of genetic aberrations and higher grade. This is evident in particular when the mean Gleason score of group pT2N0 is compared to group pT3N0 (t-test: $p < 0.0001$). For the comparison between T1 to pT2N0, however, this difference is not significant (t-test: $p > 0.6$). This result may follow from the heterogeneity of the incidental carcinomas, which are defined clinically as a tumor category by their mode of incidental clinical presentation, but not as pathological stage. For example, when a prostatectomy is performed after an incidental carcinoma has been found, such a specimen may be free of tumour, but such a carcinoma may also be widely invasive resulting in a final stage of pT2 or pT3.



(a)

Fig. 2. Each figure represents the CGH data of a selected case of incidental prostate carcinoma. The small graphs show the findings on particular chromosomes. Below each graph the chromosome number and the number of studied metaphases are indicated. Left of the graph the ideogram of the corresponding normal chromosome is shown. The curves are the ratio profiles. The three vertical lines indicate the normal value of 1 and the upper and lower bounds for normal ratio profiles. Ratio profiles extending beyond the left border indicate loss of DNA in a tumour, whereas an extension beyond the right border indicates a gain of DNA in a tumour. Losses and gains are marked by gray bars. (a) This case shows a loss at 8p, moreover gains were found on chromosomes 3q and 19p. Most regions showed however no alterations by CGH. (b) In this case more drastic changes of the genome could be observed: a loss and a gain on chromosome 2, losses on 5q and 6q, extended gains on chromosomes 7 and 8 which practically involve the total DNA of both arms, a loss on 13q and a gain on 19p.



(b)

Fig. 2. (*Continued.*)

Finally, we note that in tumour stages pT2N0 and pT3N0, the clusters rich in losses and the clusters rich in gains, are not randomly agglomerated but there are regularities with respect to the most frequent aberrations per cluster. For example, in both groups we have the chromosome arms 6q, 8p and 13q in the clusters with predominant losses, and the chromosome arms 9q and 20q among the clusters with predominant gains. The results are in accordance with the literature [1, 2,14,15,18,31,32,41]. The coupling of the aforementioned losses and gains in the same clusters could only be documented in the manifest cancers with pT2N0 and pT3N0 stage, but not for the T1 carcinomas. This finding underlines the view that T1 carcinomas are not just a lower stage than pT2N0 and pT3N0 cases, but a special, heterogeneous tumour category. This heterogeneity was also reflected in the CGH findings, where 70% of the cases showed normal CGH profiles, whereas in the remainder of the cases drastic changes could not infrequently be documented (see Table 1 and Fig. 2a,b).

Acknowledgements

Thanks are due to Dr. Horst Hameister for fruitful discussions. The study was supported by the IZKF (interdisciplinary centre for clinical research), Ulm, grant number C8.

References

[1] J.C. Alers, J. Rochat, P.-J. Krijtenburg, W.C.J. Hop, R. Kranse, C. Rosenberg, H.J. Tanke, F.H. Schröder and H. van Dekken, Identification of genetic markers for prostatic cancer progression, *Lab. Invest.* **80** (2000), 931–942.

[2] J.C. Alers, P.-J. Krijtenburg, A.N. Vis, R.F. Hoedenmaeker, M.F. Wildhagen, W.C.J. Hop, T.H. Van der Kwast, F.H. Schröder, H.J. Tanke and H. Van Dekken, Molecular cytogenetic analysis of prostatic adenocarcinomas from screening studies, *Am. J. Pathol.* **158** (2001), 399–406.

[3] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.G. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown and L.M. Staudt, Distinct types of diffuse large B-cell lymphoma indentified by gene expression profiling, *Nature* **403** (2000), 503–511.

[4] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack and A.J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci. USA* **96** (1999), 6745–6750.

[5] T.F.E. Barth, A. Benner, M. Bentz, H. Döhner, P. Möller and P. Lichter, Risk of false positive results in comparative genomic hybridization, *Genes Chromos. Cancer* **28** (2000), 353–357.

[6] J.H. Beitchman, E.M. Adlaf, L. Douglas, L. Atkinson, A. Young, C.J. Johnson, M. Escobar and B. Wilson, Comorbidity of psychiatric and substance use disorders in late adolescence: a cluster analytic approach, *Am. J. Drug Alcohol Abuse* **27** (2001), 421–440.

[7] B.M. Bennett, Tests for marginal symmetry in contingency tables, *Metrika* **19** (1972), 23–26.

[8] M. Bentz, A. Plesch, S. Stilgenbauer, H. Döhner and P. Lichter, Minimal sizes of deletions detected by comparative genomic hybridization, *Genes Chromos. Cancer* **21** (1998), 172–175.

[9] M. Bentz, A. Plesch, L. Bullinger, S. Stilgenbauer, G. Ott, H.K. Müller-Hermelink, M. Baudis, T.F. Barth, P. Möller, P. Lichter and H. Döhner, Positive mantle cell lymphomas exhibit complex karyotypes and share similarities with B-cell chronic lymphocytic leukemia, *Genes Chromos. Cancer* **27** (2000), 285–294.

[10] A.H. Bowker, A test for symmetry in contingency tables, *J. Am. Stat. Assoc.* **43** (1948), 572–574.

[11] A. Brazma and J. Vilo, Gene expression data analysis, *FEBS Lett.* **480** (2000), 17–24.

[12] S.M. Dhanasekaran, T.R. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K.J. Pienta, M.A. Rubin and A.M. Chinnaiyan, Delineation of prognostic biomarkers in prostate cancer, *Nature* **412** (2001), 822–826.

[13] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein, Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA* **95** (1998), 14 863–14 868.

[14] A. El Gedaily, L. Bubendorf, N. Willi, W. Fu, J. Richter, H. Moch, M.J. Mihatsch, G. Sauter and T.C. Gasser, Discovery of new DNA amplification loci in prostate cancer by comparative genomic hybridization, *Prostate* **46** (2001), 184–190.

[15] W. Fu, L. Bubendorf, N. Willi, H. Moch, M.J. Mihatsch, G. Sauter and T.C. Gasser, Genetic changes in clinically organ-confined prostate cancer by comparative genomic hybridization, *Urol.* **56** (2000), 880–885.

[16] J. Herrero, A. Valencia and J. Dopazo, A hierarchical unsupervised growing neural network for clustering gene expression patterns, *Bioinformatics* **17** (2001), 126–136.

[17] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, 1988.

[18] S. Joos, U.S.R. Bergerheim, Y. Pan, H. Matsuyama, M. Mentz, S. Du Manoir and P. Lichter, Mapping of chromosomal gains and losses in prostate cancer by comparative genomic hybridization, *Genes Chromos. Cancer* **14** (1995), 267–276.

[19] A. Kallioniemi, O.P. Kallioniemi, D. Sudar, D. Rutovitz, J.W. Gray, F. Waldman and D. Pinkel, Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors, *Science* **258** (1992), 818–821.

[20] O.P. Kallioniemi, A. Kallioniemi, A. Piper, J. Isola, F.M. Waldman, J.W. Gray and D. Pinkel, Optimizing comparative genomic hybridization for analysis of DNA sequence copy number changes in solid tumors, *Genes Chromos. Cancer* **10** (1994), 231–243.

[21] P. Kalocsai and S. Shams, Uses of Bioinformatics in arrays, in: *Methods Molec. Biol. 170, DNA Arrays: Methods and Protocols*, J.B. Rampal, ed., Humana Press, Totowa, 2001.

[22] R. Karhu, M. Kähkönen, T. Kuukasjärvi, S. Pennanen, M. Tirkkonen and O. Kalloniemi, Quality control of CGH: impact of metaphase chromosomes and the dynamic range of hybridization, *Cytometry* **28** (1997), 198–205.

[23] M. Kendall and A. Stuart, The advanced theory of statistics, in: *Design and Analysis, and Time Series*, Vol. 3, 3rd edn, Griffin, London, 1976.

[24] T. Kohonen, *Self-Organizing Maps*, 2nd edn, Springer, Heidelberg, 1997.

[25] T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen and K. Torkkola, LVQ_PAK: The learning vector quantization program package. Technical Report A30, Helsinki University of Technology, Laboratory of Computer and Information Science, Otaniemi, Finland, 1996.

[26] T. Kohonen, J. Hynninen, J. Kangas and J. Laaksonen, SOM_PAK: The self-organizing map program package. Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, FIN-02150 Espoo, Finland, 1996.

[27] S. Martello and P. Toth, Linear assignment problems, *Annals Discr. Math.* **31** (1987), 259–282.

[28] T. Mattfeldt, H.A. Kestler, R. Hautmann and H.W. Gottfried, Prediction of prostatic cancer progression after radical prostatectomy using artificial neural networks: a feasibility study, *BJU Int.* **84** (1999), 316–323.

[29] T. Mattfeldt, H.A. Kestler, R. Hautmann and H.-W. Gottfried, Prediction of postoperative prostatic cancer stage on the basis of systematic biopsies using two types of artificial neural networks, *Eur. Urol.* **39** (2001), 530–537.

[30] T. Mattfeldt, H. Wolter, R. Kemmerling, H.-W. Gottfried and H.A. Kestler, Cluster analysis of comparative genomic hybridization (CGH) data using self-organizing maps: application to prostate carcinomas, *Analyt. Cell. Pathol.* **23** (2001), 29–37.

[31] N. Nupponen, L. Kakkola, P. Koivisto and T. Visakorpi, Genetic alterations in hormone-refractory recurrent prostatic carcinomas, *Am. J. Pathol.* **153** (1998), 141–148.

[32] N. Nupponen and T. Visakorpi, Molecular cytogenetics of prostate cancer, *Microsc. Res. Techn.* **51** (2000), 456–463.

[33] L. Sachs, *Angewandte Statistik*, 5th edn, Springer, Berlin, 1978.

[34] SAS Institute, SAS/STAT Users Guide, Release 6.03 edition, SAS Institute, Cary, NC, 1988.

[35] H.P. Sattler, V. Rohde, H. Bonkhoff, T. Zwergel and B. Wullich, Comparative genomic hybridization reveals DNA copy number gains to frequently occur in human prostate cancer, *Prostate* **39** (1999), 79–86.

[36] G. Sherlock, Analysis of large-scale expression data, *Curr. Opin. Immunol.* **12** (2000), 201–205.

[37] L.H. Sobin and C. Wittekind, *International Union against Cancer (UICC): TNM Classification of Malignant Tumours*, 5th edn, Wiley-Liss, New York, 1997.

[38] R.R. Sokal and F.J. Rohlf, *Biometry*, 2nd edn, Freeman, San Francisco, 1981.

[39] D. Stoyan, H. Stoyan and U. Jansen, *Umweltstatistik*, Teubner, Stuttgart/Leipzig, 1997.

[40] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander and T.R. Golub, Interpreting patterns of gene expression with self-organizing maps: Methods and application to hemopoietic differentiation, *Proc. Natl. Acad. Sci. USA* **96** (1999), 2907–2912.

[41] T. Visakorpi, A.H. Kalloniemi, A.-C. Synänen, E.R. Hyytinen, R. Karhu, T. Tammela, J.J. Isola and O.R. Kalloniemi, Genetic changes in primary and recurrent prostatic cancer by comparative genomic hybridization, *Cancer Res.* **55** (1995), 342–347.

[42] H. Wolter, H.-W. Gottfried and T. Mattfeldt, Genetic changes in stage pT2N0 prostate cancer studied by comparative genomic hybridization, *BJU Int.* **89** (2001), 310–316.

[43] H. Wolter, D. Trijic, H.-W. Gottfried and T. Mattfeldt, Chromosomal changes in incidental prostatic carcinomas detected by comparative genomic hybridization, *Eur. Urol.* **41** (2002), 328–334.

[44] K.Y. Yeung, D.R. Haynor and W.L. Ruzzo, Validating clustering for gene expression data, *Bioinformatics* **17** (2001), 309–318.

[45] A. Zell, *Simulation neuronaler Netze*, Addison-Wesley, Bonn, 1994.