Combining Not-Proper ROC Curves and Hierarchical Clustering to Detect Differentially Expressed Genes in Microarray Experiments

Stefano Parodi[1], Vito Pistoia[2], and Marco Muselli[1](✉)

[1] Institute of Electronics, Computer and Telecommunication Engineering,
National Research Council of Italy, Via De Marini 6, 16149 Genoa, Italy
{parodi,muselli}@ieiit.cnr.it

[2] Laboratory of Oncology, G. Gaslini Children's Hospital,
Largo G. Gaslini, 16147 Genoa, Italy
vitopistoia@ospedale-gaslini.ge.it

Abstract. *TNRC* (Test for Not Proper ROC Curve) is a statistical tool recently developed to identify differently expressed genes in microarray studies. In previous investigations it was demonstrated to be able to separate hidden subgroups in a two-class experiment, but being a univariate technique it could not exploit the complex multivariate correlation naturally occurring in gene expression data. In this study we show as the combination of *TNRC* with a standard technique of hierarchical clustering may provide useful biological insights. An example is provided using data from a publicly available data set of 4026 gene expression profiles in 42 samples of lymphomas and 14 samples of normal B cells.

Keywords: ROC analysis · Hierarchical clustering · Feature selection · Gene expression

1 Introduction

In the last four decades Receiver Operating Characteristic (ROC) curve analysis has been extensively used in biomedical setting for the evaluation of the performance of tumour markers for diagnostic and prognostic purposes [1–4]. In recent years, ROC curve parameters, including the whole and the partial area under the curve (AUC and $pAUC$, respectively), have also been applied to feature selection tasks to identify potential markers from microarray experiments [5,6]. However, not-proper ROC curves that cross the ascending diagonal ("wiggly" curves) are in general discarded by standard methods of analysis, in that the value of AUC and often also of $pAUC$ tend to be similar to those of a not informative curve [7].

Some statistical tests, able to identify wiggly ROC curves, have been developed, including two algorithms based on the projected length and on the area

E. Formenti et al. (Eds.): CIBB 2013, LNBI 8452, pp. 238–247, 2014.
DOI: 10.1007/978-3-319-09042-9_17

swept out by the ROC curve [7], Pietra and Gini indices of the corresponding Lorentz curve [8], and a test on the highest vertical distance between the rising diagonal and the curve. This latter can be estimated either on the whole set of observed values (in that case corresponding to the Kolmogorov-Smirnov statistics [4]) or an *a priori* identified range of specificity [9]. However, these methods are all unable to separate proper ROC curves from not-proper ones and also tend to show a low statistical power [9].

A statistical test for not-proper ROC curves ($TNRC$) has been recently developed, which was demonstrated to be able to identify differently expressed genes that tend to escape common statistical methods of feature selection [5,10]. $TNRC$ is highly specific for not-proper curves and its statistical power clearly outperformed that of other statistical methods in a large simulation study [9]. Furthermore, differently from the above cited methods, a high level of the $TNRC$ statistics can reveal hidden subgroups inside either one class under study [10]. In particular, $TNRC$ was applied to a large dataset of gene expression profiles [11] and was able to identify 16 genes that had not been selected by two standard methods of analysis (namely, AUC and Student t statistics). Interestingly, 13 out of the 16 corresponding not-proper ROC curves allowed to separate either the two hidden subclasses of malignant lymphomas (namely, CLL and FL) or the two hidden subgroups of differently stimulated normal cells [10].

A limit of the application of not-proper ROC curves is that it is impossible to assess if a high value of the $TNRC$ statistics actually corresponds to hidden subclasses with clinical or biological meaning in the absence of some *a priori* information. In such case not-proper ROC analysis could take advantage from information derived from common methods of unsupervised data mining that have been largely applied in several biomedical fields including gene expression data analysis [12]. In the present investigation we will show how results from hierarchical clustering can contribute to the interpretation of not-proper ROC curves, comparing the expression profile of an apparently homogeneous group of diffuse large B-cell lymphoma (DLBCL) with that of a group of non-neoplastic B cells (NBC).

2 ROC Curve and the *TNRC* Statistics

Consider a sample of n subjects, classified into two classes (A and B, respectively) on the basis of a binary outcome Y taking values in $\{0, 1\}$. Suppose that a variable of interest (*e.g.*, the expression level of a given gene) is measured in all the n individuals of the study. If n_0 is the number of subjects belonging to class A ($Y = 0$), denote with $X_1, X_2, \ldots, X_{n_0}$ the values assumed by the variable of interest in this group of subjects, and denote with $W_1, W_2, \ldots, W_{n_1}$ the values measured in the n_1 individuals belonging to class B ($Y = 1$). The empirical ROC curve can then be defined by considering different threshold values c for the variable of interest and by computing the true and the false positive fractions, denoted by $TPF(c)$ and by $FPF(c)$, respectively, in the sample at hand [2,4]. It can be seen that:

$$TPF(c) = \frac{1}{n_1}\sum_{j=1}^{n_1} I(W_j \geq c), \quad FPF(c) = \frac{1}{n_0}\sum_{i=1}^{n_0} I(X_i \geq c) \tag{1}$$

where I is the indicator function providing $I(X_i = c) = 1$ if $X_i = c$ and $I(X_i = c) = 0$ otherwise. TPF is often called the *sensitivity* of a diagnostic test, while FPF corresponds to $1 -$ *specificity*.

Let AUC_k be the partial area under an ROC curve between the consecutive abscissa points $FPF(c_k{-}1)$ and $FPF(c_k)$, with $k = 1, \ldots, n$, computed according to the standard trapezoidal rule. The total area AUC under the ROC curve is then given by:

$$AUC = \sum_{k=1}^{n_0} AUC_k = \sum_{k=1}^{n_0} \frac{1}{2}(TPF(c_k) + TPF(c_k - 1))(FPF(c_k) - FPF(c_k - 1))$$

When $TPF(c_k) = FPF(c_k)$ for any k, every threshold c_k is not able to provide a valid classification for the two groups of subjects, *i.e.*, the class is assigned by chance. In this case we obtain a particular ROC curve, named the *chance line* (or *chance diagonal*) corresponding to the rising diagonal (Fig. 1, panel A). It should be observed that $AUC = 0.5$ for the chance line.

AUC is strictly related to the Mann-Whitney U statistics [13]. In particular, when referred to a gene expression profile, AUC corresponds to the probability that a subject randomly selected from class B has a higher gene expression than a subject randomly selected from class A [14]. In most cases, the greater is the value of AUC, the higher is the difference between the two distributions [2,4]. Figure 1 shows an example of a proper (concave) ROC curve (panel A) derived from two normal distributions (panel B, plot I). However, in some cases the ROC curve is not-proper and crosses the chance line in one or more points (curve II in Fig. 1, panel A). In this case, even if the value of AUC is close to 0.5, the two distributions can differ significantly (plot II in panel B). To recover these situations, the $TNRC$ statistics was introduced, by employing the following definition [10]:

$$TNRC = \sum_{k=1}^{n_0} |AUC_k - A_k| - |AUC - 0.5| \tag{2}$$

where A_k represents the partial area below the *chance line*.

When an ROC curve completely lies above (resp. below) the *chance line* we have $AUC_k \geq A_k$ (resp. $AUC_k < A_k$) for every $k = 0, 1, \ldots, n$, and (2) gives $TNRC = 0$. As a special case, this holds also for the *chance line*.

As shown in our previous paper [10], high values of $TNRC$ may correspond to a variety of not-proper ROC plots, including sigmoid and anti-sigmoid shaped curves. In particular, when a class of malignant cells samples is compared to non-neoplastic samples, considered as the referent (*i.e.*, corresponding to the class with $Y = 0$), sigmoid curves point out the presence of two hidden subclasses among normal cells, whereas anti-sigmoid curves indicate the presence of two hidden subclasses inside malignant cells. Finally, differently shaped not-proper

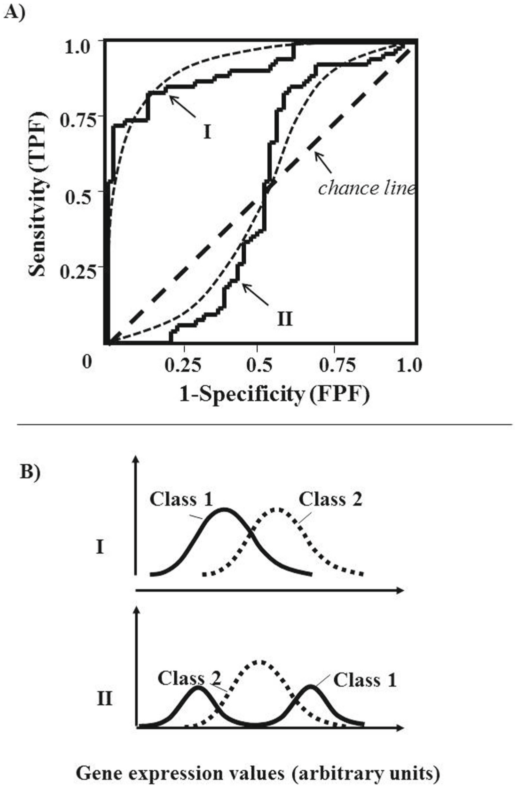

Fig. 1. Proper (concave, curve I) and not-proper (sigmoid, curve II) ROC curves (panel A) and the corresponding gene expression distributions (plot I and plot II, respectively, panel B).

curves can be occasionally observed. In general, they are difficult to interpret, and they may originate from multimodal distributions within either one class [10].

2.1 Properties of *TNRC*

It should be noted that the first part of the *TNRC* statistics in (2) corresponds to the area between the ROC curve and the chance diagonal ($ABCD$). Then, (2) can be rewritten as follows:

$$TNRC = ABCD - |AUC - 0.5|$$

Considering that $ABCD$ can be split into two subareas, namely the part above ($ABCD_a$) and below ($ABCD_b$) the *chance diagonal*, it can be easily shown that $TNRC$ corresponds to the minimum value between $ABCD_a$ and $ABCD_b$:

$$TNRC = 2\min(ABCD_a, ABCD_b) \tag{3}$$

As a matter of fact:

$$AUC = ABCD_a - ABCD_b + 0.5$$

that, replaced in (2), provides:

$$\begin{aligned} TNRC &= ABCD - ABCD_a + ABCD_b = 2ABCD_b, \quad \text{if} AUC \geq 0.5 \\ TNRC &= ABCD + ABCD_a - ABCD_b = 2ABCD_a, \quad \text{if} AUC < 0.5 \end{aligned}$$

Since $ABCD_a \geq ABCD_b$ (resp. $ABCD_a < ABCD_b$) if $AUC \geq 0.5$ (resp. $AUC < 0.5$), (3) follows.

2.2 Interpreting *TNRC* Using Information from Hierarchical Clustering

Hierarchical clustering represents a standard simple unsupervised method for the analysis of microarray data able to exploit the complex correlation inside gene expression profiles [11]. When applied to an apparently homogeneous class, the associated plot (dendrogram) can identify subsets of samples belonging to hidden distinct subclasses. Conversely, *TNRC* is a supervised method that is also able to discover hidden clusters of samples, but, as illustrated above, it needs a referent group to make a comparison between the cumulative distributions of each feature in two classes.

Accordingly, a dendrogram identifying two distinct clusters can be combined with a not-proper ROC curve simply by merging the two corresponding plots, as it will be illustrated in the example reported in the Results section. The concordance between the hidden subclasses identified by the two methods can be assessed by standard statistical methods of bivariate analysis (*e.g.*, Pearson χ^2 test, Fisher exact test or some index of concordance [15]).

A hierarchical clustering based on the Euclidean distance, was successfully applied to the large group of DLBCL considered for the present analysis (also including few samples from some selected lymphoma cell lines) and was able to identify two clusters with a signature characteristic of normal germinal cells (GC) and activated circulating B cells (AC), respectively [11]. Very interestingly, the two identified sub-classes corresponded to two groups of patients with a statistically significant different survival.

In order to combine information from *TNRC* and hierarchical clustering analysis we have classified our samples in over- and under-expressed on the basis of their location on the ROC curve, similarly to our previous investigation [10]. For this task, ROC curves identified by a high value of *TNRC* were separated into sigmoid shaped curves that can point out a bimodal distribution among NBC, and anti-sigmoid shaped curves, probably corresponding to a bimodal distribution among DLBCL samples. Other differently shaped curves were excluded from the analyses because considered as not-informative [10].

The two supposed hidden clusters were identified simply by splitting the ROC plot into two parts drawing a vertical line in the middle of the graph, thus crossing the x axis in the correspondence of 0.5 specificity. Samples lying at the left part of the graph were considered as over-expressed, whereas those lying at the opposite site were classified as under-expressed. It should be noted that, according to the above reported definition of *FPF*, in the presence of an anti-sigmoid curve, over-expressed (under-expressed) samples correspond to DLBCL with a gene expression higher (lower) than the median expression among NBC.

The association between over- and under-expression obtained from the not-proper ROC plots and the DLBCL classification from hierarchical clustering (namely, GC and AC) was assessed by the Pearson χ^2 test, and p-values <0.05 were considered as statistically significant.

3 Results

Analysis was performed on a subset of samples from the database by Alizadeh *et al.* [11], which included 4026 gene expression profiles in many different samples of lymphomas or non-neoplastic cells. For the present analysis we selected a

Table 1. Comparison between the gene expression of 14 samples of normal circulating B cells and 42 samples of diffuse large B cell lymphomas by the *TNRC* test. The first 20 selected features, corresponding to the highest *TNRC* values, are listed.

N	Gene ID	Gene name	$TNRC$	P_{TNRC}	AUC
1	GENE1358X	*c-fos*	0.1667	0.275	0.481
2	GENE563X	*Similar proteasome sub. p112*	0.1616	0.230	0.477
3	GENE3494X	*ribonuclease 6 precursor*	0.1553	0.170	0.511
4	GENE3968X	*Deoxycytidylate deaminase*	0.1536	0.220	0.490
5	GENE289X	*Unknown*	0.1514	0.330	0.453
6	GENE789X	*KIAA0052*	0.1446	0.135	0.487
7	GENE813X	*Unknown*	0.1429	0.340	0.438
8	GENE790X	*Unknown*	0.1417	0.210	0.532
9	GENE173X	*Unknown*	0.1406	0.075	0.489
10	GENE1226X	*LD78 beta*	0.1372	0.055	0.499
11	GENE2474X	*Unknown*	0.1366	0.100	0.498
12	GENE1860X	*KSR1*	0.1366	0.190	0.474
13	GENE3493X	*ribonuclease 6 precursor*	0.1361	0.140	0.533
14	GENE1225X	*MIP-1 alpha*	0.1338	0.160	0.537
15	GENE1086X	*LYL-1*	0.1321	0.175	0.520
16	GENE2335X	*Unknown*	0.1315	0.120	0.475
17	GENE295X	*FBP1*	0.1315	0.165	0.474
18	GENE3967X	*Deoxycytidylate deaminase*	0.1298	0.090	0.499
19	GENE827X	*Unknown*	0.1281	0.125	0.474
20	GENE904X	*cote1*	0.1253	0.065	0.503

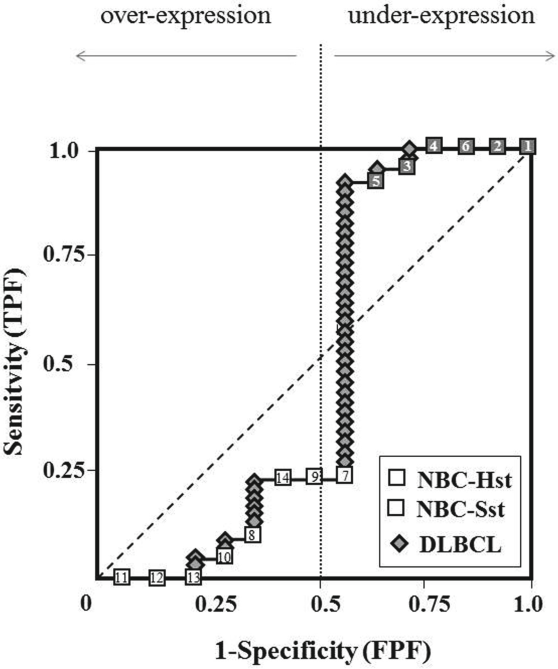

Fig. 2. Not-proper ROC curve corresponding to the expression of GENE1358X (c-fos) in Table 1. Hst = Highly stimulated; SSt = Slightly or not stimulated. NBC samples are numbered according to Alizadeh et al. (2000) [11].

class of 14 NBC, stimulated in different ways (6 heavily and 8 slightly or not stimulated) and a class of 42 DLBCL.

Feature selection was performed using the $TNRC$ statistics. The first 20 genes corresponding to the highest $TNRC$ values were retained. An estimate of the false discovery rate (FDR) was obtained from 200 permutations, using the method by Tusher *et al.* [16], while the probability for each gene to be included in the first 20 ones (P_{TNRC} or P_{AUC}) was estimated by the method by Pepe *et al.* [6] using 200 bootstrapped samples.

Similarly to our previous analysis [10], selected genes were grouped on the basis of their function as follows: lymphocyte related genes (group 1), major histocompatibility complex related genes (group 2), genes involved in malignant cell transformation (group 3), genes related to nucleic acid metabolism or DNA transcription (group 4), and gene encoding various enzymes/kinases (group 5). In spite of some overlap, this classification allows to subdivide the tested genes according to their functional features.

Table 1 shows the results of the comparison between DLBCL and NBC. Seven genes had an unknown function at the time of the microarray experiment (genes n. 5, 7, 8, 9, 11, 16, 19), while the remaining fell in group 1 (genes n. 10, 14),

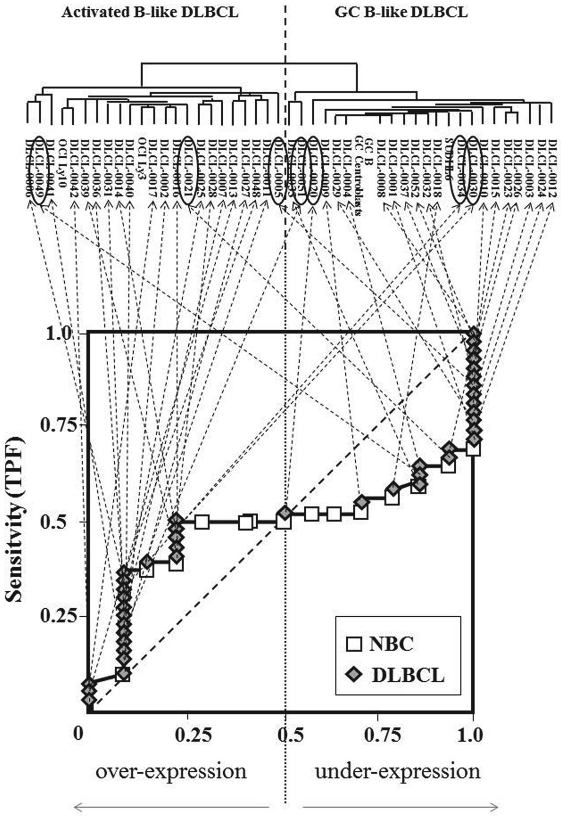

Fig. 3. Not-proper ROC curve corresponding to the expression of GENE3968X (Deoxycytidilate deaminase) in Table 1, and sample classification by hierarchical clustering. The corresponding dendrogram above the ROC curve was adapted from the original Fig. 3 in Alizadeh et al., 2000 [11] (http://www.nature.com/nature/journal/v403/n6769/full/403503a0.html), with permission.

group 3 (genes n. 15, 17), group 4 (genes n. 1, 3, 4, 6, 13, 18) or group 5 (genes n. 2 and 12).

AUC estimates were all close to the expected value of 0.5 and, accordingly, the corresponding selection probabilities equal 0 for any comparison (not shown in Table 1). Conversely, values of *TNRC* statistics ranged from 0.1253 to 0.1667 and the corresponding selection probability P_{TNRC} varied between 0.065 to 0.275. *FDR* estimate was 16.9 %. Anti-sigmoid shaped curves were observed in five cases (genes n. 4, 11, 12, 15 and 18), whereas the remaining curves all had a rather regular sigmoid shape, which indicated the existence of at least two hidden subclasses within the NBC class. As an example, ROC curve for GENE1358X (*c-fos*) is shown in Fig. 2. All the six samples lying in the right side of the plot corresponded to highly stimulated B cells, while the remaining eight samples, corresponding to slightly or not stimulated B cells, lied at the opposite side. As expected, anti-sigmoid curves allowed to separate DLBCL in (allegedly) over- and under-expressed groups. Figure 3 shows the ROC curve corresponding to the *Deoxycytidilate deaminase* expression profile (gene n. 4 in Table 1). The large majority of samples in the right part of the curve corresponded to GC B-like DLBCL identified by the hierarchical clustering in the original paper by Alizadeh *et al.* [11], with only three exceptions (namely: DLCL-005, DLCL-0021 and DLCL-0049), while all samples but four (DLCL-0030, DLCL-0034, DLCL-0020, and DLCL-0051) in the left part of the curves (over-expressed respect to NBC) corresponded to Activated B-like DLBCL. This association was highly statistically significant ($\chi^2 = 18.7, p < 0.001$).

With regards to the other four anti-sigmoid curves, no association was found between gene n. 11, gene n. 12 and the GC/AC status. Over-expressed DLBCL samples for gene n. 15 were mostly AC (16 out of 21) and under-expressed samples were mostly GC (11 out of 21), similarly to that observed for gene n. 4, but in this case statistically significance was borderline ($\chi^2 = 3.635, p = 0.057$). Finally, DLBCL expression for gene n. 18 was strongly associated to GC/AC status ($\chi^2 = 11.96, p = 0.001$). Interestingly, gene n. 18 is a clone of gene n. 4 (*Deoxycytidilate deaminase*), thus indicating that a chance finding due to multiple testing is very unlikely.

4 Conclusions

TNRC represents a new methodology of ROC analysis, which belongs to the supervised methods of feature selection. The main limitation of ROC analysis is that it cannot take into account the complex multivariate correlation between features that is commonly encountered in gene expression databases. Conversely, hierarchical clustering is an unsupervised methodology that can identify hidden subgroups of genes and/or samples exploiting the distance between features in a multivariate Euclidean space [17]. The main limit of this technique is the tendency to find pseudo-clusters also in data sets of randomly generated features. Results from the present investigation, even if still explorative, indicates that the combination of not-proper ROC analysis with traditional hierarchical clustering can provide useful insights for the interpretation of gene expression data.

References

1. Kampfrath, T., Levinson, S.S.: Brief critical review: statistical assessment of biomarker performance. Clin. Chim. Acta **419**, 102–107 (2013)
2. Alemayehu, D., Zou, H.: Applications of ROC analysis in medical research: recent developments and future directions. Acad. Radiol. **19**, 1457–1464 (2012)
3. Parodi, S., Muselli, M., Carlini, B., Fontana, V., Haupt, R., Pistoia, V., Corrias, M.V.: Restricted ROC curves are useful tools to evaluate the performance of tumour markers. Stat. Methods Med. Res. 26 Jun 2012 [epub ahead of print]
4. Pepe, M.: The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, Oxford (2003)

5. Silva-Fortes, C., Turkman, M.A., Sousa, L.: Arrow plot: a new graphical tool for selecting up and down regulated genes and genes differentially expressed on sample subgroups. BMC Bioinf. **13**, 147 (2012)
6. Pepe, M., et al.: Selecting differentially expressed genes from microarray experiments. Biometrics **59**, 133–142 (2003)
7. Lee, W., Hsiao, C.: Alternative summary indices for the receiver operating characteristic curve. Epidemiology **7**, 605–611 (1996)
8. Lee, W.: Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorentz curve-based summary measures. Stat. Med. **18**, 455–471 (1999)
9. Kagaris, D., Yiannoutsos, C.: A multi-index ROC based methodology for high throughput experiments in gene discovery. Int. J. Data Min. Bioinf. **8**, 42–65 (2013)
10. Parodi, S., Pistoia, V., Muselli, M.: Not proper ROC curves as new tool for the analysis of differentially expressed genes in microarray experiments. BMC Bioinf. **9**, 410 (2008)
11. Alizadeh, A., et al.: Distinct types of diffuse large B-cell lymphoma identified by gene. Nature **403**, 503–511 (2000)
12. Michiels, S., Kramar, A., Koscielny, S.: Multidimensionality of microarrays: statistical challenges and (im)possible solutions. Mol. Oncol. **5**, 190–196 (2011)
13. Bamber, D.: The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J. Math. Psychol. **12**, 387–415 (1975)
14. Parodi, S., Muselli, M., Fontana, V., Bonassi, S.: ROC curves are a suitable and flexible tool for the analysis of gene expression profiles. Cytogenet. Genome Res. **101**, 90–91 (2003)
15. Gibbons, J.D., Chakraborti, S.: Nonparametric Statistical Inference, 4th edn. Marcel Dekker Inc, New York (2003)
16. Tusher, V., Tibshirani, R., Chu, G.: Significance analysis of microarrays applied to the ionizing radiation response. Proc. Nat. Acad. Sci. USA **98**, 5116–5121 (2001)
17. Hastie, T., Tibshirani, R., Friedman, J.: Hierarchical clustering. The Elements of Statistical Learning, 2nd edn, pp. 520–528. Springer, New York (2009)