Biometrical Journal **47** (2005) 4, 554–562 DOI: 10.1002/bimj.200410118

Tests for Differentiation in Gene Expression Using a Data-Driven Order or Weights for Hypotheses

Gerhard Hommel[*,1] and **Siegfried Kropf**[1,2]

[1] Institut für Medizinische Biometrie, Epidemiologie und Informatik, Universität Mainz, Germany
[2] Institut für Biometrie und Medizinische Informatik, Universität Magdeburg, Germany

Received 30 August 2004, revised 30 November 2004, accepted 5 January 2005

Summary

In the analysis of gene expression by microarrays there are usually few subjects, but high-dimensional data. By means of techniques, such as the theory of spherical tests or with suitable permutation tests, it is possible to sort the endpoints or to give weights to them according to specific criteria determined by the data while controlling the multiple type I error rate. The procedures developed so far are based on a sequential analysis of weighted p-values (corresponding to the endpoints), including the most extreme situation of weighting leading to a complete order of p-values. When the data for the endpoints have approximately equal variances, these procedures show good power properties.

In this paper, we consider an alternative procedure, which is based on completely sorting the endpoints, but smoothed in the sense that some perturbations in the sequence of the p-values are allowed. The procedure is relatively easy to perform, but has high power under the same restrictions as for the weight-based procedures.

Key words: Multiple tests; Closure test; Familywise error rate; Data-driven order for hypotheses; Data-driven weights for hypotheses; Gene expression.

1 Introduction

Many recent biomedical applications generate a lot of data coming from few subjects, but with very many variables. The most current example is the analysis of gene expression data in microarrays, where the number of hypotheses corresponds to the number of genes determining the variables (endpoints). The conductor of the experiment usually wishes to "screen" the hundreds or thousands of genes on the microarray chip for "interesting" ones. When a hypothesis for a gene is tested by means of a suitable statistical test, a p-value is obtained. Rejecting the corresponding hypothesis whenever its p-value is less than a usual bound α, e.g. $\alpha = 0.05$, leads to an excessive number of type I errors and cannot be recommended in a screening situation. A common solution is to control the multiple level α (control of the FWE = familywise error rate in the strong sense, Hochberg and Tamhane, 1987), i.e. it is guaranteed that the probability of committing any type I error is at most α. A simple procedure satisfying this condition is the sequentially rejective Bonferroni-Holm procedure (Holm, 1979); however, it is not suitable in the high-dimensional situation, in general, since it needs as starting condition for the rejection of at least one hypothesis a very small p-value, namely less than or equal to α/k, where k is the number of endpoints.

Another procedure controlling the FWE is based on an a priori order of the endpoints (Bauer et al., 1998) and requires no adjustment; hypotheses can be rejected as long as – within the given order – all p-values are $\leq \alpha$. But this procedure is not reasonable for screening situations, as well, since there is generally no meaningful a priori order. Of course, it is not allowed to sort the endpoints simply after looking at the data. Nevertheless, this strategy is admissible (while controlling the FWE), when

[*] Corresponding author: e-mail: hommel@imbei.uni-mainz.de

only partial information from the data is used according to specific rules. Such procedures are based either on the concept of "stable" tests (Läuter, 1996), or of "adaptive" rank tests (Randles and Hogg, 1973). In principle, these procedures do not use the information about the differences to be tested, but only about dispersions under the null hypothesis; however, if the distributions of the different endpoints have approximately equal variances, when the respective null hypothesis is true, then a large observed dispersion suggests that there is a difference concerning the corresponding endpoint.

When the strategy suggested by Bauer et al. is applied in that sense, the problem remains that the procedure has to stop with no further rejection as soon as a p-value greater than α occurs, irrespective of how many small or even very small p-values come in the sequel. Westfall et al. (2004) proposed not to use a complete order for the endpoints, but to determine weights for the endpoints formed by specific characteristica of the data and to perform the weighted version of Holm's (1979) procedure. Thus, they obtain a smoothed version of the procedure based on the complete order.

In Section 3 of this paper, we propose another type of smoothing. Our procedure is based on the complete order of endpoints, but we admit some perturbations of the p-values, i.e. it is allowed to "jump" over some large p-values when testing according to this order. By means of applying the closure test with suitable constituents (local tests), we show that for this procedure the FWE is controlled, as well.

2 Multiple Tests with Data-driven Order or Weights

We consider the two most common situations for testing multiple endpoints, namely the cases of one sample or of two samples, respectively. A global comparison of more than two samples can be performed with analogous methods as for two samples. The number of endpoints is k, corresponding to the number of hypotheses. All observation vectors are assumed to have continuous density.

In the one-sample situation, we have n data vectors

$$\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jk})', \quad j = 1, \ldots, n,$$

from n individuals. The $\boldsymbol{x}_j$ are independent and identically distributed, with multivariate density $f(\boldsymbol{x})$, symmetrical to a location vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$, i.e.

$$f(\boldsymbol{\mu} + \boldsymbol{x}) = f(\boldsymbol{\mu} - \boldsymbol{x}).$$

The null hypotheses of interest are $H_i\colon \mu_i = 0$, $i = 1, \ldots, k$.

In the two-sample situation, one has $n = n_1 + n_2$ independent data vectors

$$\boldsymbol{x}_{\ell j} = (x_{\ell j1}, \ldots, x_{\ell jk})', \quad \ell = 1, 2, \quad j = 1, \ldots, n_\ell.$$

The distributions of the $\boldsymbol{x}_{\ell j}$ follow a location model with densities $g_\ell(\boldsymbol{x})$, where

$$g_1(\boldsymbol{x} - \boldsymbol{\mu}_1) = g_2(\boldsymbol{x} - \boldsymbol{\mu}_2) = g(\boldsymbol{x}).$$

For $\boldsymbol{\mu}_\ell = (\mu_{\ell 1}, \ldots, \mu_{\ell k})'$, $\ell = 1, 2$, the null hypotheses $H_i\colon\ \mu_{1i} = \mu_{2i},\ i = 1, \ldots, k$, are to be tested.

2.1 Parametric analysis

For the one-sample case, in addition to the above conditions we assume that the $\boldsymbol{x}_j$, $j = 1, \ldots, n$, follow a multivariate normal distribution with expectation $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$ and a positive semidefinite covariance matrix $\boldsymbol{\Sigma}$. For all endpoints, one determines weights

$$w_i = \sum_{j=1}^{n} x_{ji}^2, \quad i = 1, \ldots, k.$$

The multiple test procedure is performed in the same way as for a priori ordered hypotheses, except that the order is determined by the order of the w_i: First, a level α test is performed for the hypothesis H_i with the largest value of w_i. If this H_i cannot be rejected, the procedure stops and and no hypothesis

is rejected. Otherwise, one continues with a level α test for the hypothesis with the second largest value of w_i. Again, the procedure stops if this hypothesis cannot be rejected. If it can be rejected, one continues with the third largest w_i, and so on. All tests to be performed are one-sample t-tests.

In the two-sample case, the $\boldsymbol{x}_{\ell j}$ are assumed to have multivariate normal distributions with expectations $\boldsymbol{\mu}_\ell = (\mu_{\ell 1}, \ldots, \mu_{\ell k})'$, $\ell = 1, 2$, and common positive semidefinite covariance matrix $\boldsymbol{\Sigma}$. The weights are here

$$w_i = \sum_{\ell=1}^{2} \sum_{j=1}^{n_\ell} (x_{\ell j i} - \bar{x}_i)^2 , \quad i = 1, \ldots, k ,$$

where $\bar{x}_i = (n_1 + n_2)^{-1} \sum_{\ell=1}^{2} \sum_{j=1}^{n_\ell} x_{\ell j i}$ is the total mean of all data for endpoint i. The strategy of the multiple test procedure is the same as in the one-sample case; the tests to be used are two-sample t-tests.

The proof that this multiple test procedure controls the multiple level α utilizes the theory of spherical tests ("stable tests", see Läuter, 1996) and can be found in Kropf and Läuter (2002).

2.2 Nonparametric analysis

We make the same assumptions as at the beginning of this section. The multiple testing strategy is the same as in 2.1.

In the one-sample situation, the weights can be chosen as the medians of the absolute values of the data for each endpoint, i.e.

$$w_i = \text{med}\,(|x_{1i}|, \ldots, |x_{ni}|) , \quad i = 1, \ldots, k .$$

The tests to be performed are exact or asymptotic Wilcoxon one-sample tests.

For the two-sample case, it is a reasonable choice to use the interquartile range

$$w_i = q_{3i} - q_{1i} , \quad i = 1, \ldots, k ,$$

where q_{1i} and q_{3i} are the 1st resp. the 3rd quartile of the data for endpoint i combined over both samples. As tests the usual (exact or asymptotic) Wilcoxon-Mann-Whitney test or any other two-sample rank test can be chosen.

For the nonparametric case, the argument of independence of order statistics and rank statistics (see Randles and Hogg, 1973, or Büning, 1991) can be used. A proof that the multiple level α is controlled is elaborated in Kropf et al. (2004).

2.3 A smoothed procedure

All multiple test procedures based on the strategy of testing a priori ordered hypotheses have the drawback that, once one is forced to stop, even hypotheses with very small p-values, but smaller weights, cannot be rejected. Following a proposal by Westfall et al. (2004) this drawback can be weakened: They proposed to perform the weighted procedure of Holm (1979) with weights $g_i = w_i^\eta$, where η is a smoothing parameter, which has to be chosen in advance. The w_i can be determined for the parametric as well as for the nonparametric case as described in 2.1 or 2.2. For this procedure, it is required to determine the weighted p-values $q_i = p_i / g_i$ and to sort them: $q_{(1)} \leq q_{(2)} \leq \ldots \leq q_{(k)}$. Corresponding to the order of the q_i's, the hypotheses and the weights are denoted by $H_{(1)}, \ldots, H_{(k)}$ resp. $g_{(1)}, \ldots, g_{(k)}$. Then $H_{(j)}$ is rejected as long as

$$q_{(j)} \leq \frac{\alpha}{\left(\sum_{i=j}^{k} g_{(i)} \right)} .$$

The value of η determines the degree of weighting. Limiting cases are $\eta = 0$, leading to the unweighted Bonferroni-Holm procedure, and $\eta \to \infty$, tending to the procedure with the most extreme type of weighting, namely the full a priori strategy.

3 An Alternative Procedure

Since the procedure by Westfall, Kropf and Finos is based on weighted p-values, it has the advantage that it may be possible to "jump" over a hypothesis with a large weight w_i or g_i, but also with a large p-value, and to reject a hypothesis with a smaller weight.

We show that this advantage can be obtained even when using the weights only for ordering the hypotheses, in the following way: Choose in advance an integer m, where $(m-1)$ is the number of allowed jumps. $m = 1$ means that no jump is allowed, corresponding to the ordering strategy of Sections 2.1 or 2.2, and $m = k$ implies that arbitrarily many jumps are allowed, resulting in the unweighted Bonferroni-Holm procedure.

We propose the following procedure (for $m > 1$):

Step 1 Consider the hypothesis H_i with the largest weight.
If the corresponding p-value p_i is $\leq \alpha/m$, reject H_i.
If $p_i > \alpha/m$, retain H_i, record it as a "failure", but continue with Step 2.

Step *g* (general step): Consider the hypothesis H_i with the g-largest weight w_i.
If the corresponding p-value p_i is $\leq \alpha/m$, reject H_i.
If $p_i > \alpha/m$, retain H_i and record it as a "failure".
If the number of recorded failures is m (or if $g = k$), stop and reject no further hypothesis. Otherwise continue with Step $(g+1)$.

The proof that this stepwise procedure controls the multiple level α is given in the Appendix.

4 Treatment of Ties

Though we have assumed that the underlying distributions are continuous, it may happen, due to rounding effects, that some parameters necessary for the decisions have equal values. For the application of the weighted procedure by Holm, equality of the q_i is no problem. However, when the procedures based on the complete order of the w_i are applied, some amendments for the testing strategy are required. In case of ties, it is possible to use an a priori order of the genes defined at the very beginning, or to select at random an order of the genes with tied w_i. It is also possible to use, in a second stage, other order criteria describing the dispersion of the data, or rejection is only admitted if *all* p-values belonging to the tied w_i satisfy the demanded criterion.

We propose a more elegant solution: if there are tied weights, then one can perform an "internal" Bonferroni-Holm procedure. E.g., assume that $(m-1)$ failures have occured, and the next two weights, say $w_{i_1} = w_{i_2}$ are equal. Then one can reject an H_{i_j}, $j \in \{1, 2\}$, provided $p_{i_j} \leq \alpha/(2m)$. In case of rejection, the remaining hypothesis can be tested at level α/m.

5 A Data Example

In a study at the University of Leipzig, Medical Department III (Prof. Paschke) $n = 15$ patients with cold thyroid nodules were investigated. For each patient the nodule tissue and the surrounding thyroid tissue were compared by means of mRNA expression. $k = 12625$ genes were analysed. Using the parametric analysis with the one-sample t-test for the differences of the logarithmic expression values, 989 p-values (7.83%) were found to be less than $\alpha = 0.05$, but no p-value was less than $\alpha/k = 0.000004$, hence the Bonferroni-Holm procedure could not find any significant difference.

In order to apply the procedures described in Section 2 and 3, we determined the weights $w_i = \sum_{i=1}^{n} x_{ji}^2$ for all genes. The results for the 20 genes with the largest weights are given in Table 1. If the strategy using the complete order (Section 2.1) is applied, the first three hypotheses can be rejected, because their p-values are ≤ 0.05. But since the fourth p-value $= 0.08710 > 0.05$, no further hypothesis can be rejected. The results of the smoothed procedure (Section 2.3) depend on η (see

Table 1 Results for the first 20 of 12625 genes after sorting for decreasing w_i for the thyroid study – test decisions for the procedure of Section 3 with $\alpha = 0.05$ and $m = 10$.

inverse rank of w_i	gene no.	weight w_i	p-value p_i	test decision
1	6746	144.4	0.00019	reject
2	6567	104.2	0.00059	reject
3	3568	98.7	0.00316	reject
4	3839	89.7	0.08710	failure #1
5	2148	87.8	0.02522	failure #2
6	4384	85.5	0.22274	failure #3
7	848	80.6	0.50013	failure #4
8	7361	76.7	0.13342	failure #5
9	6257	75.1	0.93653	failure #6
10	12224	74.4	0.10661	failure #7
11	11876	72.0	0.08576	failure #8
12	7465	72.0	0.00057	reject
13	8104	69.8	0.00002	reject
14	6518	65.2	0.09610	failure #9
15	5786	64.9	0.00039	reject
16	10503	64.7	0.00180	reject
17	12172	63.4	0.12351	failure #10 $\rightarrow$ STOP!
18	825	62.5	0.00028	
19	8135	61.2	0.00020	
20	4251	60.5	0.23972	

Table 2): for $\eta \rightarrow 0$ (Bonferroni-Holm) no rejections are obtained, for $\eta \rightarrow \infty$ three rejections as above, and a maximum number of 8 rejections for values of η about 4.

The decisions for the procedure of Section 3 and $m = 10$ can be followed up in Table 1; 7 rejections are obtained. For other m, the numbers of rejections are given in Table 3. Clearly, for $m = k$ (Bonferroni-Holm) no rejections are obtained, and for $m = 1$ (a priori strategy) three rejections. A maximum of 9 rejections is obtained for values of m about 20.

In the same study, the tissues of male ($n_1 = 6$) and female ($n_2 = 9$) patients were compared, in addition. Furthermore, these $n_1 = 15$ patients were compared with $n_2 = 15$ patients with hot nodules. Every type of comparison was performed using the parametric and the nonparametric analysis. The

Table 2 R = number of rejections for the data set for different values of η (procedure of Section 2.3).

η	0	0.5	1	2	4	8	16	32	∞
R	0	0	1	2	7	8	4	3	3

Table 3 R = number of rejections for the data set for different values of m (procedure of Section 3).

m	1	5	10	20	50	100	200	500	12.625
R	3	3	7	9	8	7	5	3	0

results of the two procedures in Sections 2.3 and 3 were very similar for all 6 types of comparison, provided the optimal choice of η resp. m would have been found. The maximal number of rejections was equal for four comparisons; for two comparisons the procedure of Section 3 rejected one hypothesis more than that of Section 2.3.

6 Discussion

6.1 General comments

We described multiple testing procedures suitable for small numbers of subjects, but high-dimensional data. An arbitrary continuous multivariate location model can be assumed, including the case of multivariate normal data. If the distributions of the k endpoints, corresponding to k genes, have approximately equal variances, good power properties can be expected. Even when the assumption of equal variances is not satisfied, leading to a loss of power, it is ensured that the multiple type I error is controlled. Moreover, for many situations with gene expression data it may be possible to achieve similar dispersions among the endpoints by means of variance stabilizing transformations, as described in Huber et al. (2002). For other areas of application, as the analysis of the same target variable under different conditions (e.g. time points or spatial structures) the assumption of approximately equal variances may often be plausible, in any case.

The ordering or weighting is data-dependent, and it is not based on importance or specific characteristics or neighborhood relations of the genes known in advance. This may be a disadvantage, in particular when hypotheses with rather small p-values cannot be rejected. Nevertheless, the stategies described in this paper can be expected to yield substantial gain in power since they use the additional information from the scale measures. It should also be noted that the detection of single genes is often not the final aim of research, but rather the detection of groups of genes; a possible treatment of this type of problem with similar statistical techniques has been described in Kropf and Läuter (2002) and Läuter et al. (2004).

Since a complete order testing strategy seems to be too restrictive, the procedures of Sections 2.3 and 3 are of main interest. Each of these procedures requires the choice of an additional smoothing parameter (η or m), in advance, either assuming that good choices are known from earlier studies, or by modifying the parameters within an adaptive design (see e.g. Bauer and Köhne, 1994). Moreover, both procedures have as limiting cases the case of no weighting (Bonferroni-Holm), and the case of extreme weighting, i.e. using a complete order. However, it should be mentioned that in between there is no complete correspondence between the two methods, though the decision patterns are very similar. It is not possible to determine an η or an optimal η leading to a corresponding m or optimal m, and vice versa. For the given data set, the modification of Section 3 found slightly more rejections than the procedure by Westfall et al.; another, though only marginal, advantage is that it is still easier to perform.

6.2 Choice of the error rate

All procedures described in Sections 2 and 3 control the FWE (multiple level) α. However, this property is usually too restrictive for screening investigations, such as the identification of relevant genes. Weaker criteria for the control of type I errors were described by van der Laan, Dudoit and Pollard (2004). In particular the control of the *generalized familywise error rate* (called "control of the multiple z-level" by Hommel and Hoffmann, 1988) demands that, for a given integer $z > 0$, the probability of committing no more than $(z-1)$ type I errors is at most α. For all procedures we have described this aim can easily be reached when rejecting, in addition to the hypotheses already rejected, $(z-1)$ of the remaining hypotheses, of course only those with reasonably small p-values.

Another weaker criterion for the control of type I errors is based on the concept of the *false discovery rate* (*FDR*, Benjamini and Hochberg, 1995), i.e.

$$E(Q) = E\left(\frac{V}{R}\right) \leq \alpha\,,$$

where R is the number of all rejections, and V the number of erroneous rejections of null hypotheses (for $R = 0$ one defines $Q = 0$). In fact, we found one weighted procedure concerning the FDR, described again by Benjamini and Hochberg (1997), Section 5. However, the criterion for the type I error control for this procedure is not to control the original FDR, but rather the *weighted FDR*, i.e.

$$E(Q(w)) = E\left(\frac{\sum_{i=1}^{k} w_i V_i}{\sum_{i=1}^{k} w_i R_i}\right) \leq \alpha\,,$$

where R_i (resp. V_i) are the indicator variables of a rejection (resp. erroneous rejection) of H_i. Hence, this procedure is not suitable for our situation, since the criterion for error control depends on the weights, but the weights are determined subsequently by the data.

Nevertheless, the procedures we described seem to have even better power properties than unweighted procedures controlling the FDR, provided the assumption of approximately equal variances is satisfied. For example, consider the "exploratory Simes procedure" (Simes, 1986) controlling the FDR also under weaker conditions than that of independence (Benjamini and Yekutieli, 2001). If this procedure is applied to the data set analysed in Section 4, one obtains that $\max\{k/i \cdot p_{(i)}\}$ is reached for $i = 1$, and therefore one obtains no rejections for $\alpha = 0.05$, as for the Bonferroni-Holm procedure (since $P_{(1)} = 0.00001712$, the adjusted p-value for the corresponding hypothesis is 0.216).

6.3 Improvements and other procedures

In the Appendix it is shown that the test we proposed in Section 3 is a conservative version of the closure test, based on local tests that all are based on the minimum p-value of their intersection hypothesis. The full closure test can lead to some more rejections, but only in the case where the very most of the test results become significant. As an example, consider $m = 2$, where $\alpha/2 < p_i \leq \alpha$ for the p-value belonging to the largest w_i. Then H_i is not rejected by our procedure, but by the closure test provided all other p-values are $\leq \alpha/2$. But this is certainly a very rare situation. In general, a complete application of the closure test is very difficult to perform.

Another possibility would be to improve the Bonferroni adjustment. In the parametric case, an adjustment based on the underlying multivariate t distribution could be chosen. One could also apply resampling procedures (Westfall and Young, 1993) or Simes tests (Simes, 1986) instead of local Bonferroni tests, provided the Simes tests are admissible tests (see Sarkar, 1998). However, these improvements do not seem to result in a substantial gain in power, either, while being much more computer-intensive. We have also tried to find other suitable local tests leading to a high power of the closure test (or a conservative version of it); but no procedure with similarly good properties was found.

Finally, one could consider an application of gatekeeping strategies, as an extension of the strategy based on a complete order of hypotheses (Westfall and Krishen, 2001). From the same reason as before, serial gatekeeping procedures cannot be recommended; however, parallel gatekeeping procedures, as described by Dmitrienko et al. (2003) seem to be promising. Further research is necessary, in order to decide whether a more elaborate use of the weights may lead to a substantial increase in power.

Acknowledgement The authors are grateful to Claudia Spix and two referees for helpful comments and suggestions.

7 Appendix

Proof that the procedure of Section 3 controls the FWE

The proof is carried out using the closed testing principle.

Let $H_1, \ldots, H_k$ be the hypotheses of interest. Then every intersection hypothesis can be written as

$$H_I = \cap\{H_i : i \in I\}\,,$$

where I is a unique non-empty subset of $\{1, \ldots, k\}$.

Let $I = \{i_1, \ldots, i_r\}$ be a set with r indices such that $w_{i_1} > \ldots > w_{i_r}$, and $v = \min\{r, m\}$. Then the local test of H_I is defined by the rule

"Reject H_I if $\min\{p_{i_j} : 1 \leq j \leq v\} \leq \alpha/v$",

where p_{i_j} is the p-value from the test of H_{i_j}.

This means that the test of H_I is a simple unweighted Bonferroni test when I has at most m elements. If I has more than m elements, only the first m indices determined by the order of the weights are used for the Bonferroni adjustment.

In order to prove that the multiple level is controlled, we have to show that 1. the tests of H_I are local level α tests, and 2. if the procedure of Section 3 rejects an H_i, then the closure test rejects it, too.

1. The tests of H_I are local level α tests:
   Each of the v tests yielding the p-values p_{i_j} can be interpreted as a local test of H_I, the choice of which is based only on the weights w_i. It has been shown for the parametric case (Kropf and Läuter, 2002) and for the continuous nonparametric case (Kropf et al., 2004, Section 3) that each of these tests is a level α/v test. By means of the Bonferroni inequality, the first part is shown.
2. The procedure of Section 3 is a conservative version of the closure test:
   Define the indices $i_1, \ldots, i_k$ such that $w_{i_1} > \ldots > w_{i_k}$. If an H_{i_j} can be rejected by the procedure of Section 3, then at most $(m-1)$ p-values out of $p_{i_1}, \ldots, p_{i_{j-1}}$ are greater than α/m. Consequently, the test of an intersection hypothesis H_I with $i_j \in I$ is based on v p-values, from which at most $(m-1)$ are greater than α/m and at least one is $\leq \alpha/m$. Therefore, H_I is rejected by the local test defined above.

Since this is true for every I with $i_j \in I$, H_{i_j} is rejected by the closure test.

Remark: The case of tied w_i (Section 4) can be treated with the following choice of local tests: If $r > m$ and $w_{i_{a+1}} = \ldots = w_{i_m} = w_{i_{m+1}} = \ldots = w_{i_b}$ (i.e. there are $(b-a)$ ties), then the rule is

"Reject H_I if $\min\{p_{i_j} : 1 \leq j \leq a\} \leq \alpha/m$ or $\min\{p_{i_j} : a+1 \leq j \leq b\} \leq \dfrac{m-a}{b-a} \cdot \alpha/m$".

References

Bauer, P. and Köhne K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**, 1029–1041.

Bauer, P., Röhmel, J., Maurer, W., and Hothorn, L. (1998). Testing strategies in multi-dose experiments including active control. *Statistics in Medicine* **17**, 2133–2146.

Benjamini, Y. and Hochberg, Y., (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Benjamini, Y. and Hochberg, Y. (1997). Multiple hypothesis testing with weights. *Scandinavian Journal of Statistics* **24**, 407–418.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* **29**, 1165–1188.

Büning, H. (1991). *Robuste und Adaptive Tests.* Walter de Gruyter, Berlin, New York.

Dmitrienko, A., Offen, W. W., and Westfall, P. H. (2003). Gatekeeping strategies for clinical trials that do not require all primary effects to be significant. *Statistics in Medicine* **22**, 2387–2400.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures*. Wiley, New York.

Holm, S. (1979). A simple sequentally rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65–70.

Hommel, G. and Hoffmann, T. (1988). Controlled uncertainty. In: Bauer, P., Hommel, G., Sonnemann, E. (Eds.), *Multiple Hypothesenprüfung – Multiple Hypotheses Testing*, 154–161. Springer-Verlag, Berlin/Heidelberg/ New York.

Huber, W., von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, Suppl. 1, S96–S104.

Kropf, S. and Läuter, J. (2002). Multiple tests for different sets of variables using a data-driven ordering of hypotheses, with an application to gene expression data. *Biometrical Journal* **44**, 789–800.

Kropf, S., Läuter, J., Eszlinger, M., Krohn, K., and Paschke, R. (2004). Nonparametric multiple test procedures with data-driven order of hypotheses and with weighted hypotheses. *Journal of Statistical Planning and Inference* **125**, 31–47.

Läuter, J. (1996). Exact t and F tests for analysing studies with multiple endpoints. *Biometrics* **52**, 964–970.

Läuter, J., Glimm, E., and Eszlinger, M. (2004). Search for relevant sets of variables in a high-dimensional setup keeping the familywise error rate. *Statistica Neerlandica*, submitted.

Randles, R. H. and Hogg, R. V. (1973). Adaptive distribution-free tests. *Communications in Statistics* **2**, 337–356.

Sarkar, S. K. (1998). Some probability inequalities for ordered MTP_2 random variables: a proof of Simes' conjecture. *Annals of Statistics* **26**, 494–504.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**, 751–754.

van der Laan, M. J., Dudoit, S., and Pollard, K. S. (2004). Multiple testing. Part III. Procedures for control of the generalized family-wise error rate and proportion of false positives. Technical Report 141, Division of Biostatistics, UC Berkeley.

Westfall, P. H. and Krishen, A. (2001). Optimally weighted, fixed sequence, and gatekeeping multiple testing procedures. *Journal of Statistical Planning and Inference* **99**, 25–40.

Westfall, P. H., Kropf, S., and Finos, L. (2004). Weighted FWE-controlling methods in high-dimensional situations. In: Benjamini, Y., Bretz, F., Sarkar, S. K. (Eds.), *Recent Developments in Multiple Comparison Procedures*, IMS Lecture Notes and Monograph Series, Vol. 47, 143–154.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing*. Wiley, New York.