JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 7, 2009

Pp. 959–967
DOI: 10.1089/cmb.2008.0226

Estimation of Gene Induction Enables a Relevance-Based Ranking of Gene Sets

KILIAN BARTHOLOMÉ,[1] CLEMENS KREUTZ,[1] and JENS TIMMER[1,2]

ABSTRACT

In order to handle and interpret the vast amounts of data produced by microarray experiments, the analysis of sets of genes with a common biological functionality has been shown to be advantageous compared to single gene analyses. Some statistical methods have been proposed to analyse the differential gene expression of gene sets in microarray experiments. However, most of these methods either require threshhold values to be chosen for the analysis, or they need some reference set for the determination of significance. We present a method that estimates the number of differentially expressed genes in a gene set without requiring a threshold value for significance of genes. The method is self-contained (i.e., it does not require a reference set for comparison). In contrast to other methods which are focused on significance, our approach emphasizes the relevance of the regulation of gene sets. The presented method measures the degree of regulation of a gene set and is a useful tool to compare the induction of different gene sets and place the results of microarray experiments into the biological context. An R-package is available.

Key words: statistics, gene sets, microarray, false discovery rate.

1. INTRODUCTION

THE INVESTIGATION OF EXPRESSION PATTERNS of different biological entities using microarray experiments has become a standard technique over the last years. The results of many of these studies are often lists of genes that are differentially expressed between the sample groups. The interpretation of these lists to understand the underlying biological phenomena is tedious and often depends on the experimentalists' experience, sometimes leading to subjective interpretations. Also, the resulting gene lists can exhibit large variations when studies between different laboratories are compared, showing the limitations in the reproducibility of the expression patterns for single genes (Shi et al., 2006; Järvinen et al., 2004). A better approach is to investigate the gene expression of biologically related sets of genes. These sets can for example be a collection of genes belonging to the same pathway or ontology, or having the same chromosomal location. Besides the integration of previously achieved biological knowledge, a further advantage of the analysis of sets of genes is the better reproducibility of the results.

A number of methods to analyse the expression of sets of genes have been proposed in the last couple of years. These methods range from simple combinatorial considerations (Al-Shahrour et al., 2004; Beissbarth and Speed, 2004; Hosack et al., 2003; Lee et al., 2005; Pehkonen et al., 2005; Yi et al., 2006; Zhang et al.,

[1]Institute of Physics, University of Freiburg, Freiburg, Germany.
[2]Freiburg Institute for Advanced Studies (FRIAS).

2004; Zeeberg et al., 2003; Breitling et al., 2004; Al-Shahrour et al., 2005) via the extension of the Significance Analysis of Microarray to gene sets (SAM-GS) (Dinu et al., 2007) and the comparison of empirical distributions as utilized in the Gene Set Enrichment Analysis (GSEA) (Mootha et al., 2003; Subramanian et al., 2005) to logistic regression models (Goeman et al., 2004; Mansmann and Meister, 2005). Among the large number of gene set analysis approaches proposed, the GSEA is the most widely used method so far.

Most of these methods require a threshold value to determine whether a single gene is differentially expressed or not. The results of these methods can strongly depend on the choice of this value.

In addition, many methods are competitive (Goeman and Bühlmann, 2007). That means, the expression of genes in a certain set is compared to the expression of genes in a reference set. This reference set can for example consist of all genes that are available on the microarray. Thus, for these competitive methods the significance of a gene set cannot be given independently, but only in comparison to the reference set, and an exchange of this reference set alters the final results. This issue has also been addressed in Allison et al. (2006), Goeman and Bühlmann (2007), and Damian and Gorfine (2004).

Moreover, most methods use permutations across gene labels for the determination of the significance. As pointed out by Allison et al. (2006) and Goeman and Bühlmann (2007), this can lead to inaccurate p-values, as it assumes transcripts to be independently regulated, which is often not the case.

Furthermore, for many methods the null hypothesis that is tested either is not clearly defined (Allison et al., 2006), or it is a weak null hypothesis such as *"no genes in the gene set are differentially expressed"* or *"the genes in the gene set are at most as often differentially expressed as the genes in the reference set"* (Goeman and Bühlmann, 2007). These hypotheses are weak in the sense that their rejection does not necessarily mean that a meaningful amount of genes in the gene set is differentially expressed. In other words, significance of a gene set in the sense of the above stated null hypotheses can mean in the extreme case, that just one gene in the set is differentially expressed. Thus, a significant p-value is either difficult to interpret in the case of an unclearly defined null hypothesis, or its significance is not necessarily suitable for the evaluation of the relevance of a gene set in the case of weakly stated null hypotheses.

In this article, we present a new algorithm to estimate the number of genes in a gene set that are differentially expressed between two groups. This estimation does not require a threshold value for the detection of differentially expressed genes. The confidence interval of this estimate is determined by bootstrapping the group samples. Based on these estimations, a regulation index is defined as a measure for the regulation of a gene set. In contrast to the p-value of hypothesis tests of many other gene set analysis approaches that focus on the significance of gene sets, this index is a measure for the degree of differential gene regulation of a gene set and is thus suitable for evaluating the relevance of a gene set. We demonstrate the applicability of our method for simulated data as well as for a microarray study comparing acute lymphoid leukemia (ALL) and acute myeloid leukemia (AML) published in Armstrong et al. (2002).

2. RESULTS AND DISCUSSION

2.1. Principle of the method

A standard approach in most microarray studies is to statistically test whether the expression value of a gene is significantly different between two groups using statistical tests like the Student's t-test or its non-parametric pendant the Mann-Whitney U-test. Available methods differ in the underlying statistical assumptions. Therefore, the choice of a test depends on the assumptions for the measurements. It is assumed in the following that the test for detecting differentially expressed genes was chosen appropriately. In this case, the p-value of a statistical test is the probability of obtaining a value of the test statistic as large or larger than the observed by chance. Like for any hypothesis test, the distribution of these p-values under the null hypothesis obtained by multiple tests is by definition a uniform distribution. On the other hand, p-values resulting from statistical tests where the null hypothesis does not hold accumulate around zero. In fact, the more power an hypothesis test has, the closer is the accumulation of p-values around zero for genes for which the null hypotheses is violated.

The method we present in this article employs these properties of statistical hypothesis testing to estimate the percentage of differentially expressed genes in a gene set. Hereby, the gene set can be any user-defined collection of genes, like a certain Gene Ontology (GO) association, a signaling or metabolic pathway affiliation or the set containing all genes present on the microarray.

This problem is closely related to the estimation of the false discovery rate (FDR), which has been addressed by several groups; for example, see Benjamini and Hochberg (1995), Storey and Tibshirani (2003), Pounds and Morris (2003), Scheid and Spang (2004), Broberg (2005), and Langaas et al. (2005). In principle, methods for estimation of the FDR can be adapted to the estimation of gene induction. However, the scope of these approaches is not the determination of the number of differentially expressed genes for gene sets of limited size but the estimation of the FDR for microarray experiments with several thousand genes. Therefore, most existing methods show limited applicability for the analysis of smaller gene sets. A comparison of the performance of existing methods with our new approach introduced below for gene sets of limited size is given in the appendix.

Given a gene set of size N with a total number of R differentially expressed genes and a number of $N - R$ genes that are not differentially expressed, the probability density for the resulting p-values $\rho(p)$ can be decomposed into a uniform distribution $\rho_{unif}(p)$ and an unknown, towards zero shifted, alternative distribution $\rho_{alt}(p)$:

$$\rho(p) = \frac{R}{N}\rho_{alt}(p) + \left(1 - \frac{R}{N}\right)\rho_{unif}(p) \tag{1}$$

As shown in Figure 1a, the height of the plateau resulting from the uniform distribution is $(1 - R/N)$ and corresponds to the percentage of genes that are not differentially expressed. Thus, an estimation of the height of the plateau of the uniform distribution leads to an estimate of the number of differentially expressed genes in the gene set. Since the empirical distribution function depends on the binning and is not continuous, an estimation of the probability distribution based on the histogram of p-values is tedious especially for small gene sets. The empirical cumulative probability density on the other hand is a continuous function and better suited for parameter estimation. The drawback for parameter estimation based on an empirical cumulative distribution is that here the data are correlated. This leads to overestimated, thus conservative confidence intervals, as will be described below.

The cumulative distribution function of p is given by:

$$\begin{aligned} cdf(p) &= \int_0^p \rho(p')dp' \\ &= \frac{R}{N}cdf_{alt}(p) + \left(1 - \frac{R}{N}\right)cdf_{unif}(p), \end{aligned} \tag{2}$$

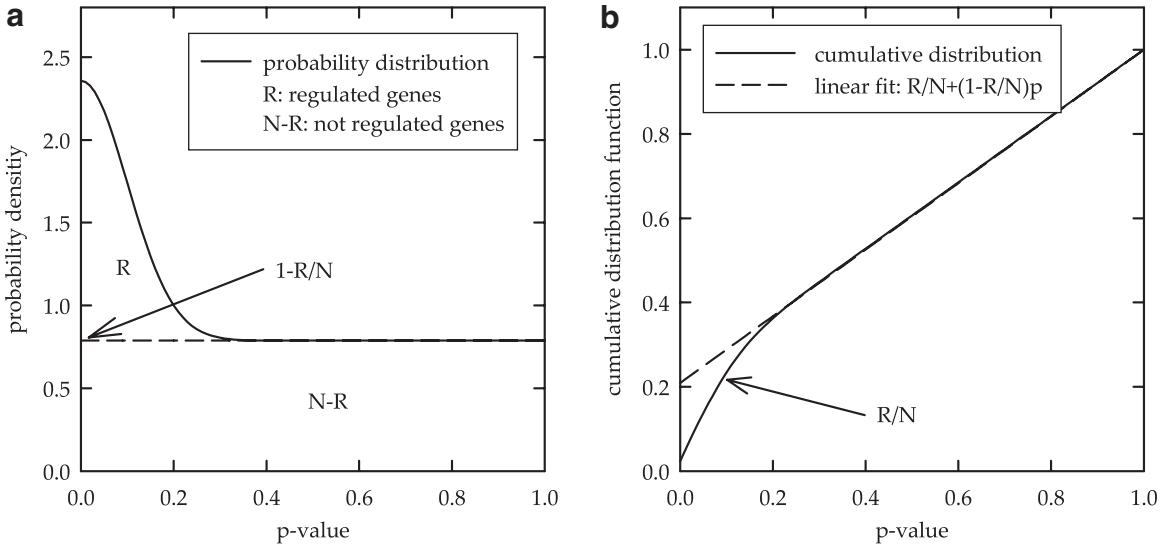

FIG. 1. Probability distribution. (**a**) Probability distribution of a sample of p-values resulting from Student's t-tests performed on a set of N genes with R differentially expressed genes. (**b**) Cumulative distribution function of the probability distribution shown in (a). The offset of the linear fit corresponds to the percentage of differentially expressed genes in the set.

where $cdf_{alt}(p)$ and $cdf_{unif}(p) = p$ denote the cumulative distribution of the probability density under the alternative hypothesis and the uniform probability density, respectively. Since the mass of $\rho_{alt}(p)$ is accumulated at small values of p, $cdf_{alt}(p)$ is a concave function and approximately equal to one for large p-values. Accordingly, for large p-values, where $cdf_{alt}(p) \approx 1$, the cumulative distribution function $cdf(p)$ can be approximated by

$$f(p) = \frac{R}{N} + \left(1 - \frac{R}{N}\right)p. \tag{3}$$

Thus, the percentage of differentially expressed genes corresponds to the y-axis offset of $f(p)$ (Fig. 1b).

In order to determine the number of differentially expressed genes in a gene set, Equation (3) is iteratively fitted to the realization of the cumulative distribution $cdf(p)$ obtained from testing for differential gene expression. In the first iteration step, all p-values are used for the linear fit and a first estimate of the number of differentially expressed genes R_{est} is obtained. In the second iteration step, the R_{est}-smallest p-values are omitted from the linear fit, and a new estimate R_{est}^{new} is achieved. This procedure is repeated until R_{est} converges, i.e., until $R_{est}^{new} = R_{est}$ (Fig. 2). This iteration process ensures that Equation (3) is only fitted to the uniform part of the cumulative distribution. Note that it is not necessary to set a threshold value that distinguishes between genes that are differentially expressed and genes that are not.

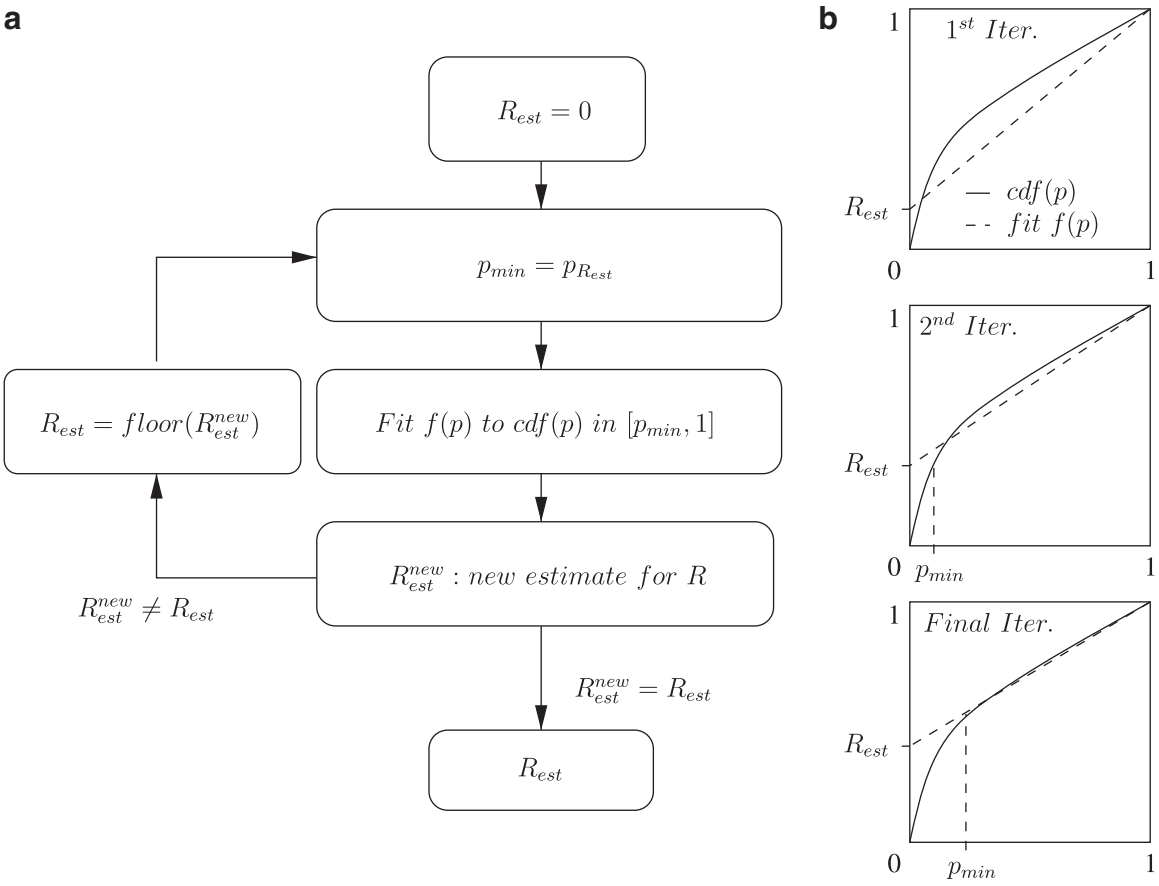

FIG. 2. Iteration algorithm. (**a**) Flowchart for the iteration algorithm: R_{est} denotes the estimated number of differentially expressed genes, $p_{\min}$ the minimum p-value entering the fitting algorithm and PR_{est} the R_{est} smallest p-value. Starting with an initial $R_{est} = 0$, $f(p)$, Equation (3), is fitted in the entire interval [0, 1] to empirical cumulative distribution and a first estimate for R_{est} is achieved. The inclusion of the p-values of the differentially expressed genes into the fitting procedure leads to an underestimation of the number of differentially expressed genes R. Thus, in the next iteration step, the R_{est}-smallest p-values are excluded from the fitting procedure, and only the resulting $N - R_{est}$ largest p-values are considered. Thus, the fitting interval is reduced to $[PR_{est}, 1]$ and a new estimate for R_{est} is calculated. This procedure is repeated until R_{est} converges. (**b**) Illustration of the algorithm. After each iteration step, the linear function converges closer to the uniformal part of the cumulative distribution function.

This iteration algorithm always has to converge for the following reason: Let R^i_{est} be the estimated number of regulated genes in iteration step i. Since the cumulative distribution function $cdf(p)$ is monotonic increasing, the p-values that are omitted in the next iteration step $i+1$ all contributed negative residues to the linear fit in the preceding iteration step i as illustrated in Figure 3. Thus, the linear fit in iteration step $i+1$ will yield an estimate R^{i+1}_{est} that cannot be smaller than R^i_{est}. The series of the estimated R^i_{est} is therefore monotonic increasing with an upper bound at 1 and has to converge.

The confidence interval of an estimated value for R is calculated by bootstrapping the group samples. Hereby, the microarrays are re-sampled with replacement by maintaining the sample group assignment. Then the number of differentially expressed genes R is estimated for each realization. As pointed out by Allison et al. (2006) and Goeman and Bühlmann (2007), bootstrapping the arrays and not the genes ensures the conservation of correlations between genes. Correlated genes have similar expression patterns and therewith similar p-values. Hence, the p-values of correlated genes accumulate in the probability distribution function, leading to a steeper slope in the cumulative probability function at this p-value. For a new bootstrap-sample, a new realization of the expression pattern is simulated, resulting in a new realization of p-values. Accordingly, for each bootstrap-realization, the point of accumulation in the cumulative distribution function is shifted, and concomitantly a different estimate for R is yielded. Thus, correlated genes lead to an increase in the estimated standard deviation.

In order to be able to compare the differential gene expression of several gene sets, we define the "Gene Set Regulation Index" (GSRI) η as the 5% quantile of the distribution of R_{est}/N determined by bootstrapping. Thus, the GSRI η is a measure of the regulation of a gene set, meaning that with a probability of 95% η or more genes are differentially expressed in this gene set. In contrast to many other gene set analysis approaches, the GSRI is not a statistical test that determines the significance for the rejection of some null hypothesis. The GSRI is rather a meaningful measure to quantify the regulation of a gene set and thereby to determine its relevance. Nevertheless, the GSRI can also be interpreted as a significance measure for the null hypothesis that no gene in the gene set is differentially expressed. By definition, a value of the GSRI larger than zero means that with a probability of 95% at least one gene is differentially expressed. This corresponds to the rejection of the above stated null hypothesis that no gene is differentially expressed with a significance level $\alpha = 0.05$.

2.2. *Simulation study*

To demonstrate the applicability of our method, a simulation study is performed for a gene set of size $N = 100$ with 20 microarray samples. The first 10 samples are the control group, and the second 10 samples are the treatment group. A series of 101 simulations was performed, increasing the number of differentially expressed genes from 0 to 100. The expression values for genes that are not differentially expressed were

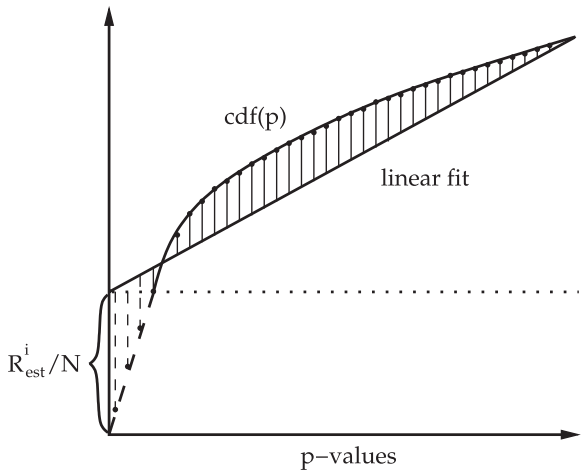

FIG. 3. Convergence of the algorithm. Let R^i_{est}/N be the offset of the linear fit in the i-th iteration of the algorithm. For the next iteration $i-1$, the R^i_{est}-smallest p-values are omitted. These p-values (dashed line) are by construction all below the dotted line. Therefore, the residues of these p-values to the linear fit in the i-th iteration are all below the linear fit. Omitting these values from the next iteration $i+1$ can under no circumstances lead to a decrease of the offset of the linear fit.

generated from a standard normal distribution $N(0,1)$ for both sample groups (control and treatment). The expression values for the differentially expressed genes on the other hand were drawn from $N(0,1)$ for the control group and from $N(1,1)$ for the treatment group, i.e. the mean value for differentially expressed genes in the treatment group was one standard deviation larger than the mean value in the control group. For these data, the cumulative distribution function $cdf(p)$ was calculated from the p-values obtained from a Student's t-test. Fitting Equation (3) to $cdf(p)$ using the iteration process illustrated in Figure 2 then yielded an estimate for the number of differentially expressed genes R. The confidence intervals were estimated by bootstrapping a thousand times from the sample groups.

The results of the simulation study are depicted in Figure 4. The left panel shows the GSRI η as well as the estimated percentage of differentially expressed genes R_{est}/N in the set plotted versus the true percentage of differentially expressed genes R/N, indicating a good agreement between estimate and true value. The right panel shows the histogram of the difference between the true percentage of differentially expressed genes R and the GSRI η. The 5%-quantile of this distribution being larger than zero shows that the estimation of R and its standard deviation $std(R_{est})$ is conservative.

There are two reasons for this. First, the result for the estimation of the number of differentially expressed genes R_{est} is in general a rational number. In each iteration step in the fitting procedure, the R_{est}-smallest p-values are disregarded from the optimization process. Thus, R_{est} has to be rounded to a natural number. Rounding off R_{est} in each iteration step leads to an underestimation and therefore a conservative estimate of R. Second, the data points in the cumulative distribution are correlated. Therefore, the bootstrapping algorithm overestimates the standard deviation, equivalent with a conservative estimate of the standard deviation. Together, this leads to an underestimation of the 5%-quantile of the distribution of R_{est} respectively the GSRI η, which is the reason for the 5%-quantile being larger than zero in Figure 3b.

2.3. Application to microarray datasets

To demonstrate the applicability of our approach, we applied our method to a study comparing the gene expression profiles of 24 ALL patients and 24 AML patients, published by Armstrong et al. (2002). For the analysis, we used gene sets with a minimum size of fifteen genes from the curated gene sets (c2) taken from the MSigDB, see Subramanian et al. (2005). This is a collection of gene sets consisting of pathways, ontologies, and lists of genes adopted from other publications.

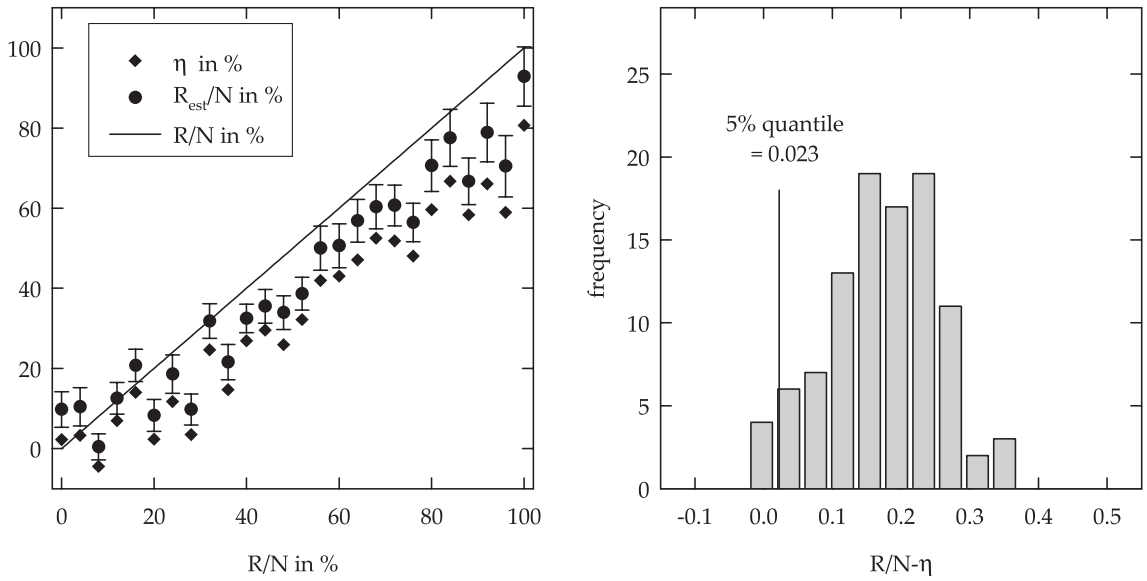

FIG. 4. Simulation study. (**a**) The estimated percentage of differentially expressed genes with confidence intervals plotted versus the real percentage of differentially expressed genes. For a better representation of the results, only every fourth simulated data point is shown. (**b**) Histogram of the difference between the true percentage of differentially expressed genes R/N and the GSRI η. The 5%-quantile of the distribution is 0.013, demonstrating the conservative estimation of the percentage of differentially expressed genes R_{est}/N and standard deviation $std(R_{est}/N)$. This leads to an underestimation and herewith conservative estimate of the GSRI.

We compared our analysis with the results of the GSEA introduced by Mootha et al. (2003). We used GSEA for comparison, since this is the most frequently used gene set analysis approach (Houstis et al., 2006; Krivtsov et al., 2006). The GSEA uses a weighted Kolmogorov-Smirnov statistics to test the null hypothesis that rankings of genes in a gene set are randomly distributed over the ranking of all genes.

For the Leukemia dataset, the GSEA is significant with a p-value smaller than 0.05 for 124 out of 1219 sets. When these p-values are corrected for multiple testing using FDR, only two gene sets are still significant, although about 60% of all genes in the experiment are differentially expressed. This lack of significance for gene sets containing a large amount of differentially expressed genes is a consequence of the competitive nature of the GSEA. The regulation of a gene set is compared with that of a reference set. This leads to decreasing power if a large amount of genes are differentially expressed in the reference set. Only those sets either containing a few highly differentially expressed genes or having disproportionate more differentially expressed genes than the reference set are significant.

The results for the ten gene sets with the highest GSRI are listed in Table 1. Both sets that were detected by the GSEA are also represented in this list. The GSRI for all these sets is larger than 79%, indicating a very large degree of regulation in these gene sets.

Eight of the top ten gene sets are gene sets that have previously been experimentally identified in other microarray studies that are closely related to the Armstrong experiment. These experiments dealt with cancer development (Schuringa et al., 2004; Iizuka et al., 2005), leukemia (Golub et al., 1999; Haslinger et al., 2004), or gene regulation in lymphocytes (Nakajima et al., 2001; Hu et al., 2004; Haddad et al., 2004). The strong association of mRNA processing and PIP3 signaling pathway in B-lymphocytes is also not surprisingly, since it has been shown that PIP3 plays an important role in tumorgenesis (Abubaker et al., 2007) and that altered mRNA processing is a feature of various kinds of cancer, including leukemia (Stoilov et al., 2002; Beghini et al., 2000; Perrotti and Neviani, 2007).

Summarizing, it can be stated that GSRI is capable of meaningfully measuring the degree of regulation of gene sets and is thus a useful tool for a further interpretation of the results of microarray experiments.

3. CONCLUSION

In this article, we introduced a new procedure to estimate the percentage of differentially expressed genes in a gene set for microarray experiments. It is based on the fact that the p-values resulting from a statistical test follow a uniform distribution when the null hypothesis is fullfilled and p-values under the alternative accumulate near zero. We introduced the Gene Set Regulation Index for quantifying the degree of differential gene expression in a gene set. In a simulation study, we have shown that this approach gives a conservative estimate of the number of differentially expressed genes and its standard deviation. We have

Table 1. Summary of the Top Ten Gene Sets for the Leukemia Data Set

Gene set	*N*	R_{est}/N	*GSRI*	*GSEA p-value*	*GSEA FDR*	*Reference*
Schuringa, STAT5A down	17	91.6%	84.7%	0.109	0.493	Schuringa et al. (2004)
mRNA, processing	40	94.8%	82.6%	0.108	0.325	MSigDB (Subramanian et al., 2005)
Golub, ALL vs. AML up	18	87.5%	81.7%	<0.001	0.249	Golub et al. (1999)
Nakajima, MCSMBP MAST	39	89.0%	81.1%	0.016	0.669	Nakajima et al. (2001)
Genotoxins, ALL 4hrs reg	17	92.9%	81.0%	0.252	0.455	Hu et al. (2004)
Haddad, HSC CD10 up	196	85.6%	81.0%	<0.001	0.039	Haddad et al. (2004)
Haddad, HPCLympho enriched	206	83.9%	79.3%	<0.001	0.020	Haddad et al. (2004)
PIP3 signaling, in B Lymphocytes	32	90.4%	79.2%	0.018	0.294	Sambrano et al. (2002)
Haslinger, B CLL 12	19	87.6%	79.2%	0.209	0.440	Haslinger et al. (2004)
Iizuka, G1 SM G2	22	90.6%	79.2%	0.244	0.440	Iizuka et al. (2005)

The ten gene sets with the highest GSRI for the leukemia data set. "N" is the number of genes that are in the gene set as well as on the microarray. "R_{est}/N" is the estimated percentage of differentially expressed genes in the set. "GSRI" is the Gene Set Regulation Index. "GSEA p-value" and "GSEA FDR" are the nominal p-values and FDR q-values calculated with the Gene Set Enrichment Analysis proposed in Subramanian et al. (2005). For all sets in this list, more than 79% of the genes are differentially expressed. For five out of the top ten gene sets, the GSEA determined a p-value smaller than 0.05, whereas after correcting for multiple testing, two sets were found to be significant by GSEA.

also shown that this method is applicable to experimental data and enables a meaningful ranking of the investigated gene sets. The advantages of our approach are (i) no thresholds are required, (ii) the method is self-contained (i.e., it does not require a reference gene set for comparison), and (iii) the result of the method is a meaningful measure of the differential gene expression of a gene set and can thus be applied for the determination of the relevance of a gene set.

AUTHORS' CONTRIBUTIONS

The original idea of the approach was introduced by K.B. with assistance of C.K. The idea for the iteration algorithm was introduced by C.K. J.T. reviewed the results and provided advice and guidance. K.B., C.K., and J.T. wrote the manuscript.

ACKNOWLEDGMENTS

We gratefully acknowledge financial support from grants BMBF HepatoSys 0313074D and BMBF FRISYS 0313921, and the FP6 EU-grant COSBICS LSHG-CT-2004-0512060.

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Abubaker, J., Bavi, P.P., Al-Harbi, S., et al. 2007. PIK3CA mutations are mutually exclusive with PTEN loss in diffuse large B-cell lymphoma. *Leukemia*.

Al-Shahrour, F., Díaz-Uriarte, R., and Dopazo, J. 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20, 578–580.

Al-Shahrour, F., Díaz-Uriarte, R., and Dopazo, J. 2005. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics* 21, 2988–2993.

Allison, D.B., Cui, X., Page, G.P., et al. 2006. Microarray data analysis: from disarray to consolidation and consensus. *Nat. Rev. Genet.* 7, 55–65.

Armstrong, S.A., Staunton, J.E., Silverman, L.B., et al. 2002. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.* 30, 41–47.

Beghini, A., Ripamonti, C.B., Peterlongo, P., et al. 2000. RNA hyperediting and alternative splicing of hematopoietic cell phosphatase (PTPN6) gene in acute myeloid leukemia. *Hum. Mol. Genet.* 9, 2297–2304.

Beissbarth, T., and Speed, T.P. 2004. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* 20, 1464–1465.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B. Methodol.* 57, 289–300.

Breitling, R., Amtmann, A., and Herzyk, P. 2004. Iterative group analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinform.* 5, 34.

Broberg, P. 2005. A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinform.* 6, 199.

Damian, D., and Gorfine, M. 2004. Statistical concerns about the GSEA procedure. *Nat. Genet.* 36, 663.

Dinu, I., Potter, J.D., Mueller, T., et al. 2007. Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinform.* 8, 242.

Goeman, J.J., and Bühlmann, P. 2007. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 23, 980–987.

Goeman, J.J., van de Geer, S.A., de Kort, F., et al. 2004. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics* 20, 93–99.

Golub, T.R., Slonim, D.K., Tamayo, P., et al. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537.

Haddad, R., Guardiola, P., Izac, B., et al. 2004. Molecular characterization of early human T/NK and B-lymphoid progenitor cells in umbilical cord blood. *Blood* 104, 3918–3926.

Haslinger, C., Schweifer, N., Stilgenbauer, S., et al. 2004. Microarray gene expression profiling of B-cell chronic lymphocytic leukemia subgroups defined by genomic aberrations and VH mutation status. *J. Clin. Oncol*. 22, 3937–3949.

Hosack, D.A., Dennis, G., Sherman, B.T., et al. 2003. Identifying biological themes within lists of genes with EASE. *Genome Biol.* 4, R70.

Houstis, N., Rosen, E.D., and Lander, E.S. 2006. Reactive oxygen species have a causal role in multiple forms of insulin resistance. *Nature* 440, 944–948.

Hu, T., Gibson, D.P., Carr, G.J., et al. 2004. Identification of a gene expression profile that discriminates indirect-acting genotoxins from direct-acting genotoxins. *Mutat. Res.* 549, 5–27.

Iizuka, N., Oka, M., Yamada-Okabe, H., et al. 2005. Self-organizing-map-based molecular signature representing the development of hepatocellular carcinoma. *FEBS* Lett. 579, 1089–1100.

Järvinen, A.-K., Hautaniemi, S., Edgren, H., et al. 2004. Are data from different gene expression microarray platforms comparable? *Genomics* 83, 1164–1168.

Krivtsov, A.V., Twomey, D., Feng, Z., et al. 2006. Transformation from committed progenitor to leukaemia stem cell initiated by MLL-AF9. *Nature* 442, 818–822.

Langaas, M., Lindqvist, B., and Ferkingstad, E. 2005. Estimating the proportion of true null hypotheses, with application to dna microarray data. *J. R. Stat. Soc. B Stat. Methodol.* 67, 555–572.

Lee, H.K., Braynen, W., Keshav, K., et al. 2005. ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinform.* 6, 269.

Mansmann, U., and Meister, R. 2005. Testing differential gene expression in functional groups. Goeman's global test versus an ANCOVA approach. *Methods Inf. Med.* 44, 449–453.

Mootha, V.K., Lindgren, C.M., Eriksson, K.-F., et al. 2003. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34, 267–273.

Nakajima, T., Matsumoto, K., Suto, H., et al. 2001. Gene expression screening of human mast cells and eosinophils using high-density oligonucleotide probe arrays: abundant expression of major basic protein in mast cells. *Blood* 98, 1127–1134.

Pehkonen, P., Wong, G., and Törönen, P. 2005. Theme discovery from gene lists for identification and viewing of multiple functional groups. *BMC Bioinform.* 6, 162.

Perrotti, D., and Neviani, P. 2007. From mRNA metabolism to cancer therapy: chronic myelogenous leukemia shows the way. *Clin. Cancer Res.* 13, 1638–1642.

Pounds, S., and Morris, S.W. 2003. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics* 19, 1236–1242.

Sambrano, G., Chandy, G., Choi, S., et al. 2002. Unravelling the signal-transduction network in B lymphocytes. *Nature* 420, 708–710.

Scheid, S., and Spang, R. 2004. A stochastic downhill search algorithm for estimating the local false discovery rate. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 1, 98–108.

Schuringa, J.J., Chung, K.Y., Morrone, G., et al. 2004. Constitutive activation of STAT5A promotes human hematopoietic stem cell self-renewal and erythroid differentiation. *J. Exp. Med.* 200, 623–635.

Shi, L., Reid, L.H., Jones, W.D., et al. 2006. The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* 24, 1151–1161.

Stoilov, P., Meshorer, E., Gencheva, M., et al. 2002. Defects in pre-mRNA processing as causes of and predisposition to diseases. *DNA Cell Biol.* 21, 803–818.

Storey, J.D., and Tibshirani, R. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* 100, 9440–9445.

Subramanian, A., Tamayo, P., Mootha, V.K., et al. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* 102, 15545–15550.

Yi, M., Horton, J.D., Cohen, J.C., et al. 2006. WholePathwayScope: a comprehensive pathway-based analysis tool for high-throughput data. *BMC Bioinform.* 7, 30.

Zeeberg, B., Feng, W., Wang, G., et al. 2003. GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.* 4, R28.

Zhang, B., Schmoyer, D., Kirov, S., et al. 2004. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinform.* 5, 16.

Address correspondence to:
Dr. Kilian Bartholomé
Institute of Physics
University of Freiburg
Herrmann-Herderstr. 3a
79104 Freiburg, Germany

E-mail: bartholo@fdm.uni-freiburg.de