# Gene Ranking Using Bootstrapped P-values

S. N. Mukherjee*
Department of Engineering Science
University of Oxford
Oxford OX1 3PJ, U.K.
sach@robots.ox.ac.uk

P. Sykacek
Department of Engineering Science
University of Oxford
Oxford OX1 3PJ, U.K.
psyk@robots.ox.ac.uk

S. J. Roberts
Department of Engineering Science
University of Oxford
Oxford OX1 3PJ, U.K.
sjrob@robots.ox.ac.uk

S. J. Gurr
Department of Plant Sciences
University of Oxford
Oxford OX1 3RB, U.K.
sarah.gurr@plants.ox.ac.uk

## ABSTRACT

Recent research has shown that it is possible to find genes involved in the pathogenesis of a particular condition on the basis of microarray experiments. Genes which are differentially expressed, for example between healthy and diseased tissues, are likely to be relevant to the disease under study. Some of the properties of microarray datasets make the task of finding these genes a challenging one. This paper proposes a gene-ranking algorithm whose main novelty is the use of bootstrapped P-values. We present an analysis of the algorithm, showing how it takes account of small-sample variability in observed values of the test statistic, in a way conventional statistical tests cannot. Experimental results show that our algorithm outperforms the widely-used two-sample $t$-test on challenging artificial data. Gene ranking is then performed on two well-known microarray datasets, with encouraging results. For example, a number of genes from one of the datasets, whose differential expression was subsequently confirmed by a more reliable biochemical analysis, are found to be ranked higher by the bootstrapped algorithm than by the conventional $t$-test, suggesting that the proposed algorithm may be better able to exploit the limited data available to infer biologically useful information.

## Keywords

microarrays, differential expression, $t$-test, bootstrap

## 1. INTRODUCTION

In recent years a number of seminal studies [7; 9; 12] have demonstrated the feasibility of using global expression analyses to better understand various diseases. Genes relevant to the pathology under investigation are expected to be up- or down-regulated between healthy and diseased tissues. An important task in microarray data analysis is therefore identifying genes which are differentially expressed in this way. Statistical analysis of gene expression data relating to complex diseases is of course not really expected to yield results of the form 'gene $X$ causes disease $Y$'. A realistic goal is to narrow the field for further analysis, to give geneticists a short-list of genes which are worth investing hard-won funds into analysing.

What makes it hard to find differentially expressed genes? Quite simply, experimental noise and biological variability. Experimental noise including errors in fabrication, hybridization, image analysis and so on, mean that the real-valued expression levels returned by a microarray experiment do not exactly reflect true mRNA levels. Biological variability refers to the natural variation we would expect to encounter even under ideal experimental conditions. That is to say, even if we could sidestep experimental issues, magically looking inside the cell and counting the RNA molecules of interest, we would still expect some variation in counts between cells in the same category.

All this means we cannot simply look at expression levels of genes in diseased and healthy tissues and choose the ones which are most different, but must treat those values as random variables, and the task of gene selection as essentially statistical. A variety of two-sample statistical tests have been applied to microarray data, including conventional [13] and non-parametric [15; 16] tests. However, with typically many thousands of genes to choose from and perhaps a few dozen to be selected, this can be a little like looking for a needle in the proverbial haystack [1].

In this paper, taking a classical two-sample test as our starting point, we focus on accounting for small-sample variability in the observed value of the test statistic. Canonical tests do not explicitly address this issue, even when parametric assumptions hold: in light of the properties of microarray data we argue that the consequences of such variability may be considerable. We use the bootstrap [5] to take account of this variability. The method developed is based on the two-sample $t$-test, which is widely used in microarray analysis [13], but we emphasise that our algorithm, and many of

*to whom correspondence should be addressed

[1]It turns out that the scale of this mismatch means that it computationally entirely infeasible to actually consider every possible subset of genes as a candidate solution. Most research in this area (ours included) essentially looks at one gene at a time.

the observations made here, generalise to other two-sample tests.

Making a brief digression, we note that the task addressed here is subtly different from that of feature selection for gene expression based classifiers [10; 18]. Two-sample tests aim to find all genes which are significantly up- or down-regulated between tissue classes; feature selection algorithms try to find genes which best explain class labels. As an example, consider a hypothetical dataset where a single gene fully explains the class labels, but a hundred genes are nonetheless consistently up-regulated in one set of tissues. A two-sample test will aim to identify all the up-regulated genes, while a feature selection algorithm should return the single explanatory gene. The distinction is biologically important: all hundred genes may have pathological effects of interest to the investigator, despite the fact that a single gene captures the class information.

## 2. BACKGROUND AND MOTIVATION

Let us introduce some notation to state more clearly the questions we wish to answer. Consider microarray slides (or chips) belonging to two classes, say, healthy and diseased, with $G$ gene expression levels measured on each slide. Recent work has shown that microarray data from higher organisms are very close to log-normally distributed [11]; in order to justify the assumptions of the $t$-test we therefore work in a log space. The data consists of $m$ $G$-dimensional vectors $\mathbf{x}_i$ (collectively referred to as $\mathcal{X}$), and $n$ $G$-dimensional vectors $\mathbf{y}_i$ (collectively referred to as $\mathcal{Y}$). $m$ and $n$ are the number of slides in each class, and the vector elements are log expression levels. We now assume these data are drawn from two (possibly different) multivariate normal distributions $q$ and $r$ respectively:

$$\begin{aligned} \mathbf{x}_i &\sim q(\mathbf{x}) \\ \mathcal{X} &= [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m] \quad (1) \\ \mathbf{y}_j &\sim r(\mathbf{y}) \\ \mathcal{Y} &= [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n] \quad (2) \end{aligned}$$

Thus, each gene has a pair of true (but unknown) class means, one from each of the distributions $q$ and $r$. Our task is to rank the genes according to how likely it is that these means are distinct. One way of accomplishing this is through the use of a test statistic, such as the $t$-statistic. In this paper we use the canonical form of the $t$-statistic throughout, assuming normally distributed data with equal but unknown variances in the two classes. We briefly present the essentials of the $t$-test below, emphasising the functional relationships between the data, statistic and P-value. A comprehensive account of the test can be found in most statistics textbooks, e.g. [4].

Let $t_k$ represent the $t$-statistic for gene $k$. $t_k$ is of course just a function of the data for the $k^{th}$ gene ($\mathcal{X}_k$ and $\mathcal{Y}_k$ respectively). Let $\mu_{\mathcal{X}}$ and $\mu_{\mathcal{Y}}$ represent the sample means of $\mathcal{X}_k$ and $\mathcal{Y}_k$ respectively, $\sigma^2_{\mathcal{X}}$ and $\sigma^2_{\mathcal{Y}}$ the (unbiased) sample variances [2], and $\mathcal{T}(\cdot)$ the $t$-statistic function. Then $t_k$ is given by:

$$\begin{aligned} t_k &= \mathcal{T}(\mathcal{X}_k, \mathcal{Y}_k) \quad (3) \\ &= \frac{\mu_{\mathcal{X}} - \mu_{\mathcal{Y}}}{(\frac{1}{m} + \frac{1}{n})^{\frac{1}{2}} (\frac{\sigma^2_{\mathcal{X}}(m-1) + \sigma^2_{\mathcal{Y}}(n-1)}{m+n-2})^{\frac{1}{2}}} \quad (4) \end{aligned}$$

The form of Equation 4 means that it is possible to analytically obtain the distribution of the statistic, which in turn allows the probability of type I errors (false positives) to be calculated. This probability is called the P-value. Under the assumptions of the canonical test, $t_k$ has a non-central $t$-distribution [4], with degrees of freedom $v = (m + n - 2)$ and non-centrality parameter $\psi_k$. In the special case of the distribution under the null hypothesis, $\psi_k = 0$ and $t_k$ has the familiar $t$-distribution with degrees of freedom $v$.

The observed value of the statistic is thus mapped to a P-value ($p_k$) by a function (which we shall call $f$) which depends on the $t$-distribution. For the two-sided test being used, $f$ is given by:

$$f(t_k) = 2[1 - \mathcal{C}_v(|t_k|)] \quad (5)$$

where $\mathcal{C}_v(\cdot)$ represents the cumulative distribution function (cdf) for a $t$-distribution with $v$ degrees of freedom. The method proposed in this paper is motivated by the following observations about the $t$-test:

- If the assumptions of the test hold, the function $f$ truly represents an error probability: $p_k$ is the probability of making a type I error, or false positive, if the significance level of the test is set just high enough to include gene $k$.
- But the ranking of a particular gene depends on the observed value $t_k$, which itself represents a single draw from a non-central $t$-distribution with unknown parameter $\psi_k$. A series of microarray experiments pertaining to the same clinical condition, with a fixed number of slides in each case, will produce varying $t$-statistics for the genes under study, and consequently quite different P-values and rankings.
- Thus, although the $t$-test captures the variability of the statistic and P-value under the null hypothesis, it cannot tell how reliably the observed values $t_k$ and $p_k$ actually represent gene $k$. In particular, if the observed statistics are atypical values under their distributions, the conclusions drawn from them may not generalise well to subsequent microarray experiments.

In classical hypothesis testing settings, the number of datapoints is relatively high, making the statistic $t_k$ and corresponding P-value $p_k$ good representatives of the $k^{th}$ feature. In contrast, the imbalance between slides and genes in microarray experiments places a considerable burden on the ability of ranking algorithms to discriminate between relevant and irrelevant genes.

[2]For typographic simplicity we have taken the liberty of dropping the subscript $k$ from the means and variances in Equation 4.

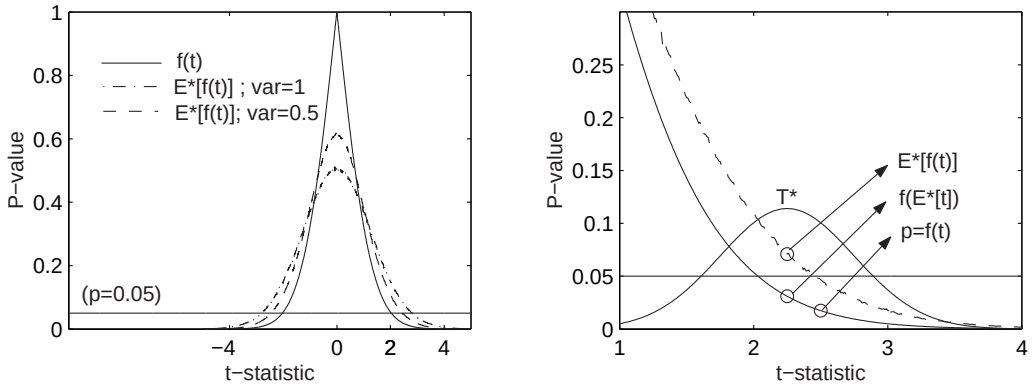


Figure 1: The effect of considering variation in the value of the observed $t$-statistic: the figure on the left shows the tail probability mass curve $f$, which maps a $t$-statistic to a P-value (Equation 5), and an illustration of the bootstrap P-value function $E^*[f(t)]$ (approximated following Equation 10), as functions of the bootstrap mean $t$-statistic $E^*[t]$, with two different variances. The figure on the right shows in detail how the bootstrap P-value relates to the estimated distribution of the $t$-statistic in the region of interest of the curve. In this illustrative example, the observed statistic $t$ is located to the right of the bootstrap mean $E^*[t]$ - considering the variability in $t$ we thus find a P-value considerably higher than the conventional one.

## 3. METHODS

As shown in Equation 5, the P-value is a statistic of the data; we use the bootstrap [5] to obtain an estimate of its value. The bootstrap is a widely-used resampling technique, by which an empirical estimate of the distribution of a statistic of interest can be obtained by repeatedly computing its value from datasets sampled with replacement from the original.

Let $E^*[\mathcal{F}(Z)]$ represent the bootstrap average of a function $\mathcal{F}$ of data $Z$:

$$E^*[\mathcal{F}(Z)] \equiv \lim_{B\to\infty} \frac{1}{B}\sum_{b=1}^{B} \mathcal{F}(Z^{*b}) \quad (6)$$

where the $Z^{*b}$'s are datasets obtained by resampling $Z$ with replacement. In practice, $B$ is set to a large finite value; in all our experiments $B = 500$.

Following Equations 5 and 6, the bootstrap estimate of the P-value for gene $k$, $p_k^*$, is given by:

$$p_k^* \equiv E^*[f(t_k)] = \frac{1}{B}\sum_{b=1}^{B} f(\mathcal{T}(\mathcal{X}_k^{*b}, \mathcal{Y}_k^{*b})) \quad (7)$$

where $\mathcal{X}_k^{*b}$ and $\mathcal{Y}_k^{*b}$ represent data for gene $k$ from the $b^{th}$ bootstrap iteration and $\mathcal{T}(\cdot)$ the $t$-statistic function. The bootstrap mean of $t_k$ may be obtained in a similar manner. We note that as the tail probability mass $f$ (Equation 5) is a non-linear function, its bootstrap average will not, in general, be identical to either the observed P-value $p_k$, or the tail probability mass corresponding to the bootstrap mean of $t_k$, $f(E^*[t_k])$.

Statistical power analysis [2] is superficially similar to our method, in that it explicitly deals with the distribution of the test statistic $t_k$. However, the aim of such analysis is quite different, namely to quantify the probability of type II error at a given significance level.

Previous applications of resampling, and the bootstrap in particular, to testing, have included P-value adjustments and non-parametric tests [3], as well as multiplicity corrections [17] (also, in the context of microarray analysis [6]). Our algorithm is closest in spirit to bootstrap P-value adjustments [1; 8], insofar as it treats the P-value itself as the statistic of interest.

### Analysing the bootstrapped P-value

A useful approach to understanding bootstrapped P-values is to explicitly think of the P-value $p$ (we drop the subscript $k$ for clarity) and its bootstrap estimate $p^*$ as realisations of random variables. Let $P$ represent the P-value, and $g(P)$ its density function; $P^*$ represents the bootstrap estimate and $h(P^*)$ the corresponding density. Following [14] we approximate $g(P)$ by a beta distribution with parameters $\xi$ ($0 < \xi \leq 1$) and 1. Thus:

$$g(p|\xi) = \xi p^{\xi-1} \quad (8)$$

Viewed in this way, conventional tests draw a conclusion, as to whether a gene is differentially expressed or not, on the basis of a single draw from the P-value density $g(P)$; errors in the procedure can be thought of as arising due to the uncertainty in $g(P)$[3]. Under the null hypothesis, $\xi = 1$, and the P-value is uniformly distributed in the range $[0, 1]$; consequently the probability of type I error at a given significance level $\alpha$ does not depend on any unknowns and is simply equal to $\alpha$. Under the alternative hypothesis, the parameter $\xi$ is unknown. The probability of type II error then depends on $\xi$ and is given by:

$$Pr(P > \alpha|H_1) = 1 - F_g(\alpha;\xi) \quad (9)$$

[3]A simple example of a less uncertain density: if $g'$ was a delta function located at the true mean of $P$, a single draw would unerringly reveal which hypothesis was correct.

where $F_g$ is the cdf corresponding to the P-value density $g$ in Equation 8. In general, for fixed significance level $\alpha$, this second error probability rises with the parameter $\xi$.

The effect of taking account of variation in the $t$-statistic via the bootstrap, is that a draw from the density function of the bootstrapped P-value, $h(P^*)$, is more likely to be close to the true mean of $P$ than a draw from the density of the standard P-value, $g(P)$. Indeed, we find via simulation that the average deviation of the bootstrapped P-value $P^*$ around the true mean of $P$ (i.e., $E[P^* - E[P]]^{\frac{1}{2}}$) is lower than than the corresponding figure for the conventional P-value (i.e., the standard deviation of $P$, $(E[P - E[P]])^{\frac{1}{2}}$).
We further illustrate the operation of the algorithm by making some simplifying assumptions, obtaining a qualitative picture of the interplay between the various quantities involved. We assume that the bootstrap distribution of the $t$-statistic is approximately Gaussian, with mean $\mu_t^*$ (defined as the bootstrap mean of $t$, $E^*[t]$) and variance $\sigma^2$. Under these assumptions, the bootstrap P-value $p^*$ can be thought of as being the integral of the product of the Gaussian with the tail probability mass $f$ (Equation 5). The integral is then itself a function, $\hat{f}^*$, of $\mu_t^*$ and $\sigma^2$; for a given variance, $\hat{f}^*$ can be thought of as mapping a bootstrap mean $t$-statistic $\mu_t^*$ to an expected P-value:

$$\hat{f}^*(\mu_t^*, \sigma) \quad = \quad \int_{-\infty}^{\infty} f(x)\mathcal{N}(x; \mu_t^*, \sigma^2)\mathrm{d}x \qquad (10)$$

We simulate Equation 10 directly, by sampling from the Gaussian and computing mean P-values via the function $f$ (Equation 5). Figure 1 shows the integral version of the bootstrapped P-value $\hat{f}^*$ (with variances 1 and 0.5) and $f$, as functions of the bootstrap mean $t$-statistic $\mu_t^*$. The right side of the figure shows an illustrative example for a single (hypothetical) gene with the observed $t$-statistic $t = 2.5$ and the corresponding bootstrap mean $\mu_t^* = 2.25$. The form of the function $f$ makes it clear why variation around high absolute values of $t$ has little effect on P-values, but for moderate values, fluctuations in $t$ can profoundly effect the final P-value and ranking of the gene. Many genes of interest have moderate $t$-statistics and are highly sensitive to the exact observed value. For example, if the 3000 genes analysed from the colon cancer dataset [12] are arranged in descending order of absolute bootstrap mean $t$-statistics $|\mu_t^*|$, only the 72 highest ranked genes have $|\mu_t^*|$ in excess of 6.
The bootstrap distribution of the $t$-statistic on which the P-value of Equation 7 is based is purely empirical. Under the assumptions of a canonical $t$-test, however, the form of the distribution is known. We note that an alternative approach to the one taken above would thus be to estimate the non-centrality parameter $\psi$ via the bootstrap, using the estimated value to obtain the appropriate non-central $t$-distribution. Using suitable approximations, the integral of Equation 10 could then be evaluated to obtain a P-value for each gene. One advantage of the approach we have taken is that it generalises easily to the non-parametric case: the cdf $C_v$ used in Equation 5 need only be replaced by an empirical (for example, permutation-based) cdf. The P-value could then be obtained as before from Equation 7.

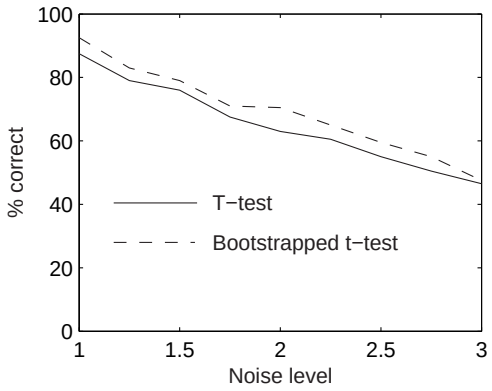


Figure 2: Results of bootstrapped and conventional $t$-tests on artificial data. The score reported is the proportion of two hundred iterations in which the algorithm in question ranked the correct features in the top two places. The bootstrapped test is able to take account of small-sample variability in the observed statistic and outperforms the conventional test at various noise levels.

# 4. RESULTS

## 4.1 Artificial data

We assess the ability of the proposed algorithm to detect differentially expressed genes on artificially generated data. Six-dimensional data in two classes are generated from six pairs of univariate Gaussians, only two of which have distinct means. The class variances are equal and are made to vary incrementally between 1 and 3, to simulate increasing noise levels; the number of samples in the two classes are 10 and 5. We choose a relatively small number of samples to mimic microarray data; to account for small-sample effects in the results we report average scores over 200 iterations. At each iteration, 15 samples are drawn from the Gaussians and passed to two ranking algorithms: a conventional $t$-test and the proposed algorithm, performance being assessed in terms of the proportion of runs in which the two highest ranked features are the correct ones. Repeatedly drawing data in this way provides an estimate of generalisation accuracy under small-sample conditions: the proposed algorithm scores $\sim$6.5% higher than the $t$-test, averaged over a range of variances and all 200 runs. Results are shown in Figure 2.

## 4.2 Microarray data

### *Datasets*

We select two well-known and widely analysed microarray datasets - the colorectal cancer data of Notterman *et al.* [12] and the leukaemia data of Golub *et al.*[7].

The colorectal cancer dataset consists of 36 labeled slides with 6600 complementary DNAs (cDNAs) and expressed sequence tags (ESTs) represented on each.

The leukaemia data consists of a training set of bone marrow samples taken from patients suffering from Acute Myeloid Leukaemia (AML) and Acute Lymphoid Leukaemia (ALL), and a separate test set with bone marrow as well as periph-

Table 1: The top five genes identified by the bootstrapped $t$-test from a colon cancer dataset: as a consequence of the extremely high absolute values of the $t$-statistic for these genes, the rankings do not vary much between the two algorithms.

| Rank | Acc. No. | Description | Bootstrapped P-value | Rank under $t$-test |
|---|---|---|---|---|
| 1 | M77836 | Human pyrroline 5-carboxylate reductase mRNA | $8.41 \times 10^{-9}$ | 2 |
| 2 | M83670 | Human carbonic anhydrase IV mRNA | $1.32 \times 10^{-8}$ | 1 |
| 3 | U17077 | Human BENE mRNA | $2.67 \times 10^{-8}$ | 3 |
| 4 | T96548 | Actin, Gamma-enteric smooth muscle (*Homo sapiens*) | $5.07 \times 10^{-8}$ | 4 |
| 5 | T64297 | Fatty acid binding protein, liver (*Homo sapiens*) | $6.07 \times 10^{-8}$ | 5 |

Table 2: Genes, whose differential expression was subsequently confirmed by RT-PCR, ranked by bootstrapped and conventional P-values. In most cases, the bootstrap algorithm ranks these genes higher than the $t$-test; the average number of places between the ranks is 25.5.

| Bootstrap rank | Rank under $t$-test | Acc. No. | Description |
|---|---|---|---|
| 76 | 117 | X54489 | Human gene for melanoma growth stimulatory activity (MGSA) |
| 43 | 60 | U22055 | Human 100 kDa coactivator mRNA |
| 166 | 261 | L23808 | Human metalloproteinase (HME) mRNA |
| 158 | 212 | H50438 | M-Phase inducer phosphatase 2 (*Homo sapiens*) |
| 52 | 72 | X54942 | *Homo sapiens* CKSHS2 mRNA for CKS1 protein homologue |
| 6 | 5 | M97496 | *Homo sapiens* guanylin mRNA |
| 47 | 62 | X64559 | *Homo sapiens* mRNA for tetranectin |
| 50 | 49 | L02785 | *Homo sapiens* colon mucosa-associated (DRA) mRNA |
| 93 | 83 | X86693 | *Homo sapiens* mRNA for hevin like protein |

eral blood samples. The training set contains data from 38 samples taken from patients with AML or ALL, and the test set 34 patient samples. Expression levels are given for 7129 genes/ESTs. For this work we use only the test dataset.

### *Pre-processing*

We pre-process the gene expression data according to current practice [19], removing within-slide location by changing from absolute to relative expression values. As noted previously, microarray data from higher organisms are very close to log-normally distributed [11]; we therefore transform the data into a log-space.

### *Results*

**Colorectal cancer data:** Table 1 shows P-values and identities of the five highest ranked genes identified by the bootstrap algorithm. These genes have extreme $t$-statistics, so as expected, their ranks under both algorithms are similar. Rankings returned by our algorithm and the $t$-test are noticeably different across the entire dataset: on an average, there are 65 places between the positions of the same genes; when the top 100 genes returned by the bootstrap algorithm are considered, the average displacement reduces to 15.

The difference between the results of the tests thus lies in the positions assigned to genes with high, but not extreme, $t$-statistics. These genes may be of great practical importance: indeed, from a biological perspective the aim of microarray experiments (which are high-throughput but noisy) is essentially to guide further investigation. More accurate transcript abundance analyses, for example quantitative real-time RT-PCR[4] can be used to confirm differential expression. RT-PCR is too expensive to be used to assess every gene; thus one major objective of microarray data analysis is to identify a subset of genes for such assessment. Table 4.2 compares the ranks of genes whose differential expression was subsequently confirmed by RT-PCR [12]. In most cases we find the proposed algorithm ranks these genes higher than the $t$-test, suggesting that if used to select a subset for further assessment it is more likely to uncover relevant genes.

**Leukaemia data:** Table 3 shows P-values and identities of the five highest ranked genes identified by the bootstrap algorithm on the leukaemia dataset. Once again, the extreme $t$-statistics of these top-ranked genes mean that their ranks under both algorithms are similar. For this dataset there are, on an average, 244 places between the ranks assigned to the same genes by the two algorithms; when only the top 100 genes returned by the bootstrap algorithm are considered, the average displacement is 9. Table 4 shows a selection of highly ranked genes whose ranks were higher under the bootstrapped test than the $t$-test. Some of these genes have been implicated in other studies too: for example, the Human myeloperoxidase gene has recently been found to be ranked much higher, compared to the $t$-test, by well-founded methods including information gain and a one-dimensional support vector machine [20].

[4]Reverse transcription polymerase chain reaction

Table 3: The top five genes identified by the bootstrapped $t$-test from the leukaemia dataset: once again, the extremely high absolute values of the $t$-statistic for these genes mean that the rankings do not vary much between the two algorithms.

| Rank | Acc. No. | Description | Bootstrapped P-value | Rank under $t$-test |
|---|---|---|---|---|
| 1 | D26361 | *Homo sapiens* KIAA0042 | $2.09 \times 10^{-4}$ | 1 |
| 2 | X64330 | *Homo sapiens* ATP-citrate lyase | $5.00 \times 10^{-4}$ | 3 |
| 3 | U28758 | Human NMDA receptor subtype 2B subunit (GRIN2B) mRNA, partial cds | $5.05 \times 10^{-4}$ | 4 |
| 4 | J03171 | Human interferon-alpha/beta receptor alpha chain precursor | $5.06 \times 10^{-4}$ | 5 |
| 5 | U95006 | Human D9 splice variant A mRNA | $6.34 \times 10^{-4}$ | 2 |

Table 4: Comparative results using the bootstrap and conventional $t$-tests on the leukaemia dataset. The genes shown were in the top 100 under the bootstrap test, and were ranked at least fifteen places higher than by the $t$-test.

| Bootstrap rank | Rank under $t$-test | Acc. No. | Description |
|---|---|---|---|
| 39 | 56 | U49957 | LIM protein (LPP) mRNA, partial cds |
| 41 | 60 | Y10313 | Nerve growth factor-inducible PC4 homologue |
| 58 | 79 | AB000450 | *Homo sapiens* VRK2 |
| 62 | 102 | X63578 | *Homo sapiens* gene for Parvalbumin |
| 66 | 84 | L38951 | Importin beta subunit |
| 72 | 103 | U20536 | Cysteine protease Mch2 isoform alpha (Mch2) |
| 75 | 107 | Z46376 | HK2 mRNA for hexokinase II |
| 77 | 140 | S70609 | Glycine transporter type 1b [human, substantia nigra, mRNA, 2364 nt] |
| 83 | 98 | U26648 | STX5A Syntaxin 5A |
| 85 | 124 | U61734 | Human protein trafficking protein (clone S31i125) |
| 96 | 144 | M19508 | MPO from Human myeloperoxidase gene |
| 99 | 135 | U07563 | Human proto-oncogene tyrosine-protein kinase (ABL) gene, exon 1b and intron 1b, and putative M8604 Met protein (M8604 Met) gene |

## 5. CONCLUSIONS

In this paper we have proposed a novel gene ranking method based on bootstrapped P-values, and shown that it can successfully account for small-sample effects in the observed test statistic for a gene. While it is premature to draw definitive biological conclusions from our results, experiments on both artificial and real data suggest that our algorithm is better able to deal with the level of uncertainty inherent in microarray data than a classical two-sample test. In particular, results concerning the comparative ranks of genes from the colon cancer dataset [12] whose differential expression was confirmed using RT-PCR (Table 4.2) are encouraging, and suggest that the proposed algorithm may be able to guide further investigation more accurately than the $t$-test. In essence, our method obtains a more accurate P-value at the cost of computational efficiency, but we feel that in this particular domain compute-time should not be an over-riding concern - with sometimes millions of dollars being spent on designing experiments and acquiring data, a few extra minutes or even hours of processing should be acceptable if better results can be obtained!

Clearly, many questions remain to be addressed. Further theoretical analysis is required to fully understand the distributional properties of the bootstrapped P-value. Our results are promising but preliminary - a thorough empirical evaluation of the proposed algorithm, potentially using a different two-sample test, as well as further investigation of the biological impact of the results reported here, will be informative. Also, the extension of the method proposed to a fully non-parametric setting may prove useful in analysing data, for instance from lower organisms, which do not conform to the assumptions made here.

## Acknowledgements

SNM gratefully acknowledges the support of the Biotechnology and Biological Sciences Research Council (BBSRC); thanks also to Dr. Sayan Mukherjee.

## 6. REFERENCES

[1] R. Beran. Prepivoting Test Statistics: A Bootstrap View of Asymptotic Refinements. *Journal of the American Statistical Association*, 83(403):687–697, 1988.

[2] J. Cohen. *Statistical power analysis for the behavioral sciences*. Lawerence Erlbaum, Hillsdale, NJ, 1988.

[3] A. C. Davison and D. V. Hinckley. *Bootstrap Methods and their Applications*. Cambridge University Press, 1997.

[4] M. H. DeGroot and M. J. Schervish. *Probability and Statistics*. Addison Wesley, 3rd edition, 2002.

[5] B. Efron. Bootstrap methods: another look at the jackknife. *Ann Stat.*, 7:1–26, 1979.

[6] Y. Ge, S. Dudoit, and T. P. Speed. Resampling-based multiple testing for microarray data analysis. Technical Report 633, Department of Statistics, University of California, Berkeley, 2003.

[7] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[8] P. H. Hall and M. A. Martin. On Bootstrap Resampling and Iteration. *Biometrika*, 75:661–671, 1988.

[9] I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi, B. Wilfond, A. Borg, and J. Trent. Gene-Expression Profiles in Hereditary Breast Cancer. *N. Engl. J. Med.*, 344(8):539–548, 2001.

[10] S. Hochreiter and K. Obermayer. Feature Selection and Classification on Matrix Data: From Large Margins to Small Covering Numbers. In S. T. Suzanna Becker and K. Obermayer, editors, *Advances in Neural Information Processing Systems 16*, Cambridge, MA, 2003. MIT Press.

[11] D. C. Hoyle, M. Rattray, R. Jupp, and A. Brass. Making sense of microarray data distributions. *Bioinformatics*, 18:576–584, 2002.

[12] D. Notterman, U. Alon, A. J. Sierk, and A. J. Levine. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. *Cancer Research*, 61(7):3124–30, 2001.

[13] W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, 18:546–554, 2002.

[14] T. Sellke, M. J. Bayarri, and J. O. Berger. Calibration of P-values for testing precise null hypotheses. *The American Statistician*, 55:62–71, 2001.

[15] O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein, and R. B. Altman. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18:1454–1461, 2002.

[16] V. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, 98:5116–5121, 2001.

[17] P. H. Westfall and S. S. Young. *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. John Wiley & Sons, 1993.

[18] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature Selection for SVMs. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 668–674. MIT Press, 2001.

[19] Y. H. Yang, S. Dudoit, P. Luu, and T. P. Speed. Normalization for cDNA microarray data. Technical report 589, Department of Statistics, University of California, Berkeley, 2001. Available at: http://stat-www.berkeley.edu/tech-reports/index.html.

[20] Yang Su, T. M. Murali, V. Pavlovic, M. Schaffer, and S. Kasif. RankGene: identification of diagnostic genes based on expression data. *Bioinformatics*, 19(12):1578–1579, 2003.