SCIENTIFIC REPORTS

OPEN

Statistically invalid classification of high throughput gene expression data

Shahar Barbash & Hermona Soreq

The Edmond & Lily Safra Center for Brain Sciences and the Department of Biological Chemistry at the Hebrew University of Jerusalem.

SUBJECT AREAS:

COMPUTATIONAL NEUROSCIENCE

HIGH-THROUGHPUT SCREENING

FUNCTIONAL CLUSTERING

MICROARRAYS

Received
29 November 2012

Accepted
27 December 2012

Published
22 January 2013

Correspondence and requests for materials should be addressed to H.S. (hermona.soreq@mail.huji.ac.il)

Classification analysis based on high throughput data is a common feature in neuroscience and other fields of science, with a rapidly increasing impact on both basic biology and disease-related studies. The outcome of such classifications often serves to delineate novel biochemical mechanisms in health and disease states, identify new targets for therapeutic interference, and develop innovative diagnostic approaches. Given the importance of this type of studies, we screened 111 recently-published high-impact manuscripts involving classification analysis of gene expression, and found that 58 of them (53%) based their conclusions on a statistically invalid method which can lead to bias in a statistical sense (lower true classification accuracy then the reported classification accuracy). In this report we characterize the potential methodological error and its scope, investigate how it is influenced by different experimental parameters, and describe statistically valid methods for avoiding such classification mistakes.

Rapidly increasing numbers of high throughput studies in life sciences in general and in neuroscience[1] in particular use classification analysis for comparing two or more biological states and identifying differences between them. The growing impact of these analyses is evident by their use in the fields of neurobiology, cell biology, immunology and oncology, to name a few[1–3]. Perhaps the most common comparison is between samples from patients and those from matched healthy individuals. Such analyses often serve for identifying a 'molecular signature' or 'fingerprint genes' – biomarkers that elucidate the underlying mechanism of the studied disease[4] and lead to diagnostic and drug discovery applications[5]. The continuously reduced costs, improved technological accuracy, and shortened experimental time of high throughput gene expression studies predict a future increase in similar uses of high throughput platforms. However, many studies, including classification analysis, apply statistically invalid methods that can lead to erroneous conclusions.

Results

Defining the classification inaccuracy. The most common invalid protocol starts with the selection of a subset of transcripts, usually a couple hundred out of the (roughly) 25,000 human genes studied in different high throughput platforms, which are altered between the two test groups. It then uses this subset of genes for differentiating between the two groups. This approach usually yields a clear distinction between the two groups; however, this is not surprising given the fact that the genes on which the classification was based were initially selected BECAUSE of their differential expression between the two groups. The classification of such groups is often presented as one of the main findings of these studies; furthermore, some studies argue explicitly that, based on the clear distinction between groups, these genes may serve as a 'signature' or 'fingerprint' for diagnosis of clinical conditions in the global population. Others make a more subtle argument, albeit also erroneous, that the 'successful' classification highlights the biological difference between groups. For instance, one such study stated that "Distinctions in the pattern of expression of this group of genes (the subset of significantly altered genes) between the cells of group A and the cells of group B are clearly evident", even though such a distinction would also occur for randomly selected matrices of numbers, as we demonstrate in Fig. 1.

The reason for such seemingly successful classifications being erroneous is that they are based on a circular argument, and classifying two groups based on a subset of genes that were selected to be different between the groups unavoidably yields a 'successful' classification. The correct alternative, trying to classify based on the full matrices or on a subset of randomly selected genes, will often fail regardless of the classifier being used. Figure 2 presents an example for such 'failed' classifications and the statistical tools that were employed. First, we present the chances for correct classification as dendrogram plots (see Methods for details) of real sets of experimental data and comparable random datasets. These aim to classify between gene sets of groups of 6 patients and 6 healthy matched controls (Fig. 2 a–c) or between two normal random datasets of a similar size (Fig. 2 d–f).

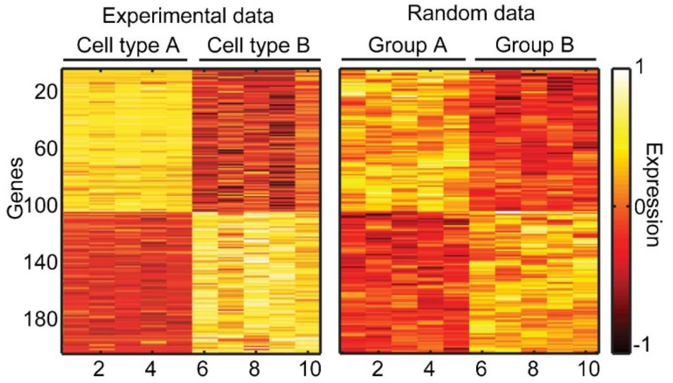

Figure 1 | Experimental classification and classification based on arbitrary numbers from a normal random matrix yield comparable classifications. Normalized expression matrices for a subset of genes taken from a larger list of deposited datasets are shown. Rows are genes and columns are biological replicates. The left matrix is based on real experimental data whereas the right one is based on random, normally distributed data. Normalized expression is color coded.

Classifications were based on the 100 most drastically altered genes (Fig. 2a and d), which in both cases emerged as 'successful', presenting statistically significant (but nevertheless invalid) differences between the two tested groups. However, comparisons which tested 100 randomly-selected genes (Fig. 2b and e) and the full expression matrices (2c and f) both emerged as 'failed' for the experimental as well as the random datasets.

Of note, there are cases where comparison of full datasets yields significant, and hence valid, classification. Several methods are readily available which can correctly differentiate between groups based on high throughput datasets. 47% of the 111 studies we screened and which performed high-throughput classification analyses of gene expression datasets used one of these methods and avoided this statistical bias in classification experiments. Below we describe such valid classification methods, which are based on unbiased classification of the expression levels of all the tested genes, on a selection rational for identifying groups of candidate genes, or on splitting the cohort into training and test datasets.

The outcome of selecting the most drastically altered genes based on a statistical test performed per gene is that the variance between the members of a group becomes much smaller than the variance between members of different groups. This is shown in Figure 2g, where the decision regions for classification are based on the minimal distance between the coordinate and the mean of the groups in the plane composed of components one and two of the Principal Component Analysis (PCA) (see 'methods' for more information). The first principal component which captures the maximal variance, and which is presented on the X axis in figure 2g, easily discriminates between the two groups. Although the results presented in figure 2g look 'real' or 'significant', the distribution of the total number of misclassifications (figure 2j–l) demonstrates that the outcome is always zero when using the 100 'most discriminative genes'. The chance of perfectly correct albeit arbitrary classification, for this scenario, is actually 100%. In other words, the P-value that should be assigned is 1.

Using statistical terminology the issue is one of selection (ascertainment) and errors in estimating effect sizes. Selecting a set of genes out of all genes based on a treatment or group differences when individual effects on gene expression are estimated with error leads to a correlation between treatment difference and random error. Creating a predictor out of the 'best' genes and fitting these back in the same data then amplifies these errors and may lead to serious bias.

Scope of the invalid analysis in representative journals. To define the frequency of misclassification errors we reviewed 250 articles encompassing different life science disciplines that employed various high throughput assays. These papers were published between 2002–2011 in high-impact journals (e.g. Nature, Nature Neuroscience, Cell and J. of Neuroscience). In about 50% of the articles the researchers performed some sort of classification; however, 53% out of these used statistically invalid classification. Figure 3 shows the prevalence of usage of this statistically invalid procedure and demonstrates consistent numbers of studies that perform classification and a consistent frequency of the statistically invalid classification.

Statistically biased classification depends on sample size and P value. The data of any biological experiment will include a stochastic element, due to either biological or technical reasons. Therefore, every biological experiment includes in its background a random matrix of this stochastic element. Most of the time this random matrix is harmless in the sense that the chances of it producing false results are very small. This is exactly why P-values are calculated – to estimate the probability of the observed result occurring by chance. In classification analysis, on the other hand, this random matrix in the background may produce statistically biased results. There are two major experimental parameters that influence the probability of the stochastic element to have an effect on the statistical rigorousness of the analysis. These are the sample size and the P-value that is selected as a threshold. To estimate the origins of such statistical bias, we studied the performance of random datasets for different sample sizes and P-values (Figure 4). This analysis showed that the smaller the group size, the less valid the classification – the reason being that random data produce more false positives (erroneous 'correct' classifications). Thus, groups of less than 20 samples are likely to yield biased classification results much more frequently than larger ones. In comparison, the effect of P-values on the chance of misclassification was maximal for medium size values (e.g. P= 0.05 and 1000 tested genes) and similarly low for very small or very large P-values, below 0.001 or above 0.7.

A high P-value implies that a large fraction of the genes analyzed in the experiment was taken into consideration, but these genes show a small difference between the groups. This is the reason for the drop in the chance of correct classification in random data (as shown in the right-hand panel in Figure 4). At the extremity of P-value approaching 1 we actually do not select any subset of genes, but rather try to classify based on all of them. And, as expected, this is where classification fails. In comparison, in the left-hand side domain of low P-values, taking too few genes into consideration hampers our ability to classify. This is where 10 genes are more informative than 5. The distinction is best evident in the central region of Fig. 4: for example, in groups including around 20 samples each and for a P-value threshold of 0.001, classification of random data is 100% successful. Unfortunately, studies of such size are rare (they are expensive and technically demanding) and lower sample sizes are often employed - which would make the statistically biased definition of 'correct' classification due to randomness even more robust (compare groups of 6 samples to those of 20 in Figure 4). This implies that the stochastic element of the experiment is by itself sufficient to separate the groups when using a subset of mostly altered genes. The 'successful' classification, therefore, cannot be attributed to the biological signal of the experiment and hence has no biological meaning. Using a valid statistical method - Leave One Out Cross Validation (LOOCV) for example - reduces the values to around 50% correct classifications across all group sizes and P-values (grey line in Figure 4), demonstrating the robustness of this validating procedure.

Discussion

Several possible reasons for the high frequency of statistically biased classification studies are presented, as follows. First, while studies in

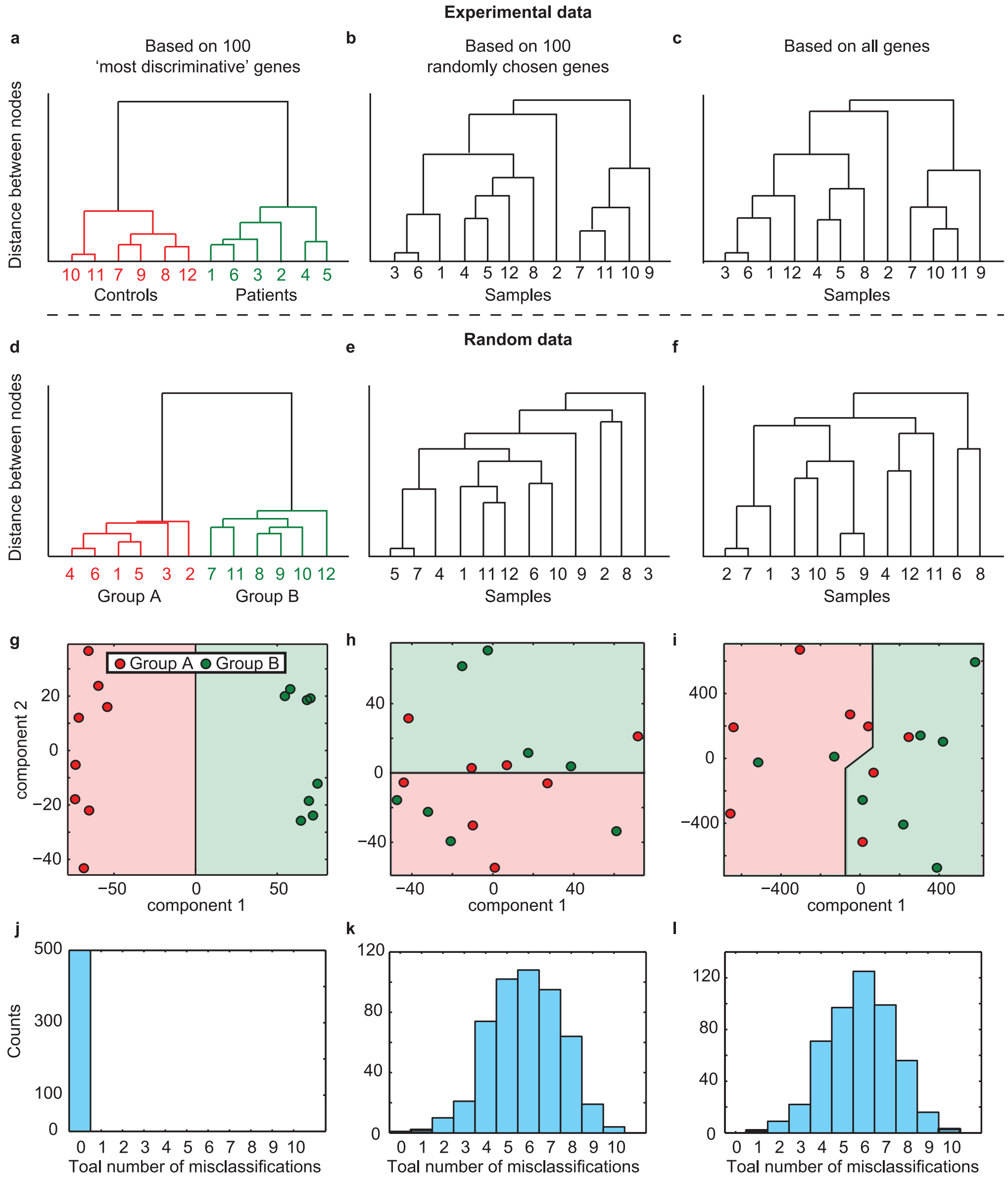

Figure 2 | Comparison of classifications based on a subset of 100 'discriminative genes' (or numbers), 100 randomly-selected ones, and full matrices. Dendrogram plots are shown for groups of 6 patients and 6 healthy matched control volunteers, based on the 100 most altered genes (a), on 100 randomly-selected genes (b) and on the full expression matrices (c). (d–f) Dendrogram plots based on normally distributed random data of similar sizes. (g–i) Scatter PCA plots and decision regions based on the 'minimal distance to group mean' classifier, on the plane of PC1 and PC2. (j–l) Distributions of the number of misclassifications based on the 'minimal distance to group mean' classifier for the three protocols described in (a–c). Number of iterations = 500. See Methods for further details.

the life sciences show a growing tendency to use computational tools, these are often internet-based software tools in which the user's control over the applied parameters is very limited and the full picture of the underlying algorithm that served to develop this software tool is inaccessible to the typical user. Second, the seemingly unbiased automatic clustering of samples, which provides a 'reliable' distinction between groups and presents a dramatic difference between different experimental conditions or patient groups. This points at the very attractive prospect of possible diagnostics or drug target applications. Of note, the conflict between the increasingly simple usage of large databases and the effort required for validation of single target genes can lead to limiting the number of validated genes that are selected for further focused studies. This, in turn leads to statistically biased selection of "signature" genes and is likely to lead to tedious, costly and frustrating efforts in translational studies.

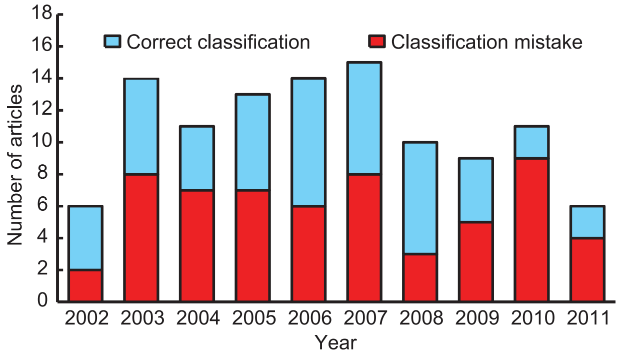

Figure 3 | Scope of invalid analyses in the literature. Stacked bar graphs represent the number of studies in which statistically invalid (red) or valid (blue) classification protocols were used, in the noted years.

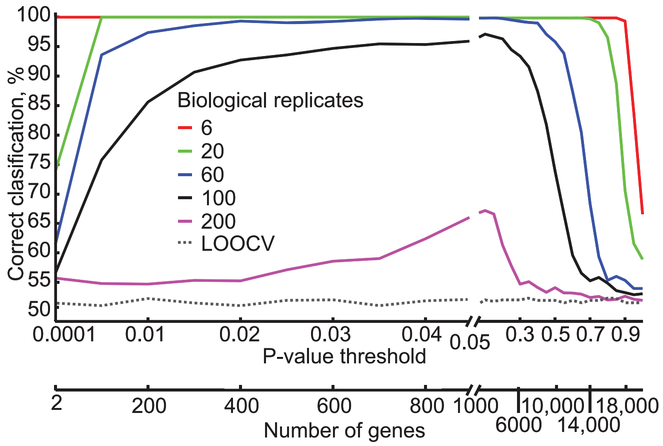

Figure 4 | Effect of sample size and selected P-value threshold on false classification. The percent of correct classification (Y axis) was based on random data using a subset of mostly altered genes. The values are means for 30 iterations. The P-value used as a threshold for the subset is on the upper X axis, the number of genes in the subset is on the lower X axis. Percent values are drawn for 5 sample sizes, called 'biological replicates' in the graph and represented as lines with different colors, for each of the two groups. For example, 'biological replicates' =20 indicates classification between 2 groups of 20 datasets each. The dotted grey line represents values for the LOOCV for 20 biological replicates (Other sizes of biological replicates give the same LOOCV result).

We have shown that clustering based on a list of 'informative genes' presenting large expression differences between the studied groups leads to statistical bias and may thus draw biologically irrelevant classifications. However, several strategies exist for avoiding such errors and reaching valid clustering analyses[6,7]. First, using an *a priori* list of genes for clustering tests, as done in[8], is absolutely essential, because the list would not depend on the specific experiment that is being assessed. For example, defining all of the genes involved in inflammation as a sub-set is a valid strategy for classification of groups with predictably distinct inflammatory features. Second, one can test for clustering based on the entire set of data, without any filtering, as done in[9]. This method has actually been the preferred approach among those studies that avoided the statistically biased classification which we discuss here. Finally, one can divide the sample population into two sets: a training set in which the separating line is searched for, and a test set, in which the clustering performance is assessed, as done in[5].

Recently, a paper by Nieuwenhuis et al. published in Nature Neuroscience[10] showed that 50% of the articles examined made a statistical mistake by which the interaction between compared groups was not taken into consideration. In some of the examined articles the implication of this mistake was that the main claim of the article could not be statistically supported. We believe that the statistically biased classification we have drawn attention to is of similar significance, and that it may have a direct impact on scientific arguments that are made in various research fields.

Methods

Dendrogram plot. The nodes at the bottom of a dendrogram plot represent samples. While climbing up the dendrogram tree the nodes are joined together according to their proximity in a multidimensional space. In the beginning samples are scattered in a multidimensional space according to their expression values (of whole transcripts or subsets of it). In the first step, each pair of samples is joined together to form a new node whose coordinates are the means of the pair. In the second step a second level of nodes is formed similarly from the nodes in the previous step. This continues until all nodes have been assigned to a pair. The distance between each two nodes is the peak on the Y axis of the inverse U shaped line that connects them.

'Minimal distance to group mean' classifier. Decision regions are determined as follows: the samples are scattered on a two dimensional plane (PCA1 and PCA2 in Figure 2 g–I, but can be any other space as well). Each pixel on the plane is assigned as 'red' or 'green' according to its distance from the means of the groups. The color of the group with the closest mean to the pixel is the one that is assigned to it. Statistically biased classifications are counted as red dots in a green region, and vice versa.

1. Cooper-Knock, J. *et al.* Gene expression profiling in human neurodegenerative disease. *Nature reviews. Neurology* **8**, 518–530 (2012).
2. Luciani, F., Bull, R. A. & Lloyd, A. R. Next generation deep sequencing and vaccine design: today and tomorrow. *Trends in biotechnology* **30**, 443–452 (2012).
3. Roychowdhury, S. *et al.* Personalized oncology through integrative high-throughput sequencing: a pilot study. *Science translational medicine* **3**, 111ra121 (2011).
4. Barbash, S. & Soreq, H. Threshold-independent meta-analysis of Alzheimer's disease transcriptomes shows progressive changes in hippocampal functions, epigenetics and microRNA regulation. *Current Alzheimer research* **9**, 425–435 (2012).
5. Segman, R. H. *et al.* Peripheral blood mononuclear cell gene expression profiles identify emergent post-traumatic stress disorder among trauma survivors. *Molecular psychiatry* **10**, 500–513, 425 (2005).
6. Ambroise, C. & McLachlan, G. J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6562–6566 (2002).
7. Wood, I. A., Visscher, P. M. & Mengersen, K. L. Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics (Oxford, England)* **23**, 1363–1370 (2007).
8. Petermann, K. B. *et al.* CD200 is induced by ERK and is a potential therapeutic target in melanoma. *The Journal of clinical investigation* **117**, 3922–3929 (2007).
9. Lopes, A. R. *et al.* Bim-mediated deletion of antigen-specific CD8 T cells in patients unable to control HBV infection. *The Journal of clinical investigation* **118**, 1835–1845 (2008).
10. Nieuwenhuis, S., Forstmann, B. U. & Wagenmakers, E. J. Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature neuroscience* **14**, 1105–1107 (2011).

Acknowledgments

The authors are grateful to Prof. Israel Nelken and Dr. Estelle R. Bennett, Jerusalem, for fruitful and comprehensive discussions and to the Gatsby, Legacy Heritage, ISF and Rosetree foundations for support. S.B. is an incumbent of a PhD fellowship from the ELSC Brain Center.

Author contributions

H.S. and S.B. designed and wrote the main manuscript text. S.B. performed the simulations.

Additional information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Barbash, S. & Soreq, H. Statistically invalid classification of high throughput gene expression data. *Sci. Rep.* 3, 1102; DOI:10.1038/srep01102 (2013).