DNA expression microarrays may be the wrong tool to identify biological pathways

Mondry A.[1,]*, Loh M.[2], Giuliani A.[3]

[1]Oxford Radcliffe Hospitals NHS Trust, Oxford, United Kingdom

[2] Bioinformatics Institute, Agency for Science, Technology and Research, Singapore

[3]Environment and Health Department, Istituto Superiore di Sanità , Roma, Italy

Corresponding author:

Dr. Adrian Mondry

Consultant Physician,

Horton Hospital,

Oxford Radcliffe Hospitals,

Banbury OX16 9AL,

United Kingdom

Tel: (44) 1295 229027, Fax: (44) 1295 229603

E-mail: mondry@hotmail.com

Abstract

DNA microarray expression signatures are expected to provide new insights into patho- physiological pathways. Numerous variant statistical methods have been described for each step of the signal analysis. We employed five similar statistical tests on the same data set at the level of gene selection. Inter-test agreement for the identification of biological pathways in BioCarta, KEGG and Reactome was calculated using Cohen's κ-score. The identification of specific biological pathways showed only moderate agreement ($0.30 < \kappa < 0.79$) between the analysis methods used. Pathways identified by microarrays must be treated cautiously as they vary according to the statistical method used.

Introduction

DNA microarrays emerged as one of the dominant technologies in biomedical research during the past decade. A major promise of the method is thought to lie in the potential to identify genes actively involved in patho-physiological pathways[1]. Discovery of such pathways and their genetic regulation may lead to targeted efforts in drug discovery.

It has by now become clear that gene clusters ("signatures"), rather than the expression of individual genes, yield higher information contents. The ultimate aim is to move from signatures to models, that is, to analyze the data, integrate the information, and thus, gain knowledge[1].

The information contained in the expression data is made available through a sequence of statistical analyses, which are not yet standardized. Most reports on DNA microarray data feature a very unique methodology. Often, the information on the methods used is insufficient to allow replication.

When assessing the usefulness of a technology, much attention must be given to the robustness of the knowledge gain, i.e. different observers should come to the same conclusion by using this method.

The present study addresses this point. More precisely, we ask: if one and the same set of microarray data is analyzed using different statistical tests at the same analytic step, will the same pathways be identified? (The data set and statistical methods are described in detail in Appendix A.) This was done by calculating Cohen's κ - score[2], which allows to assess whether inter-test agreement is more than a chance product.

In addition, we assessed the inter-test mutual agreement at different levels of analysis of the data set, namely the biological level (in terms of both gene and pathway identification) and the clinical level (case sample clustering). The comparison of these test results allows to judge the robustness of the information contents, and the independence from the possible introduction of bias through onomastically variant, but de facto redundant entries in the protein databases.

Results

Each statistical test produced a gene set of 50 genes, according to the conditions chosen. There was only partial overlap (Table 1) with not more than three genes common to all five tests. The generalized κ-score[3] was $\kappa = -28.01$.

At the analytical step of gene selection, Cohen's κ-score was low ($0.15 < \kappa < 0.68$), indicative of only "fair" to at best "substantial" agreement (Table 2).

Following submission of the five sets of 50 genes selected by the different statistical methods to all three pathway databases as described in the methods section, some genes were found to be involved in more than one pathway described in one or more of the three databases, and in each set of selected genes, pathways were identified that involved more than one of these genes (Table 3). A total of 38 genes (Appendix B: Additional Table 1) could not be allocated to any pathway described in any of the three databases. We used Cohen's kappa score to calculate agreement at the level of pathway identification between the five statistical methods. The κ –scores ranged from 0.30 to

0.79 as displayed in Table 3, which indicates mostly "moderate", in three cases "substantial", and twice only "fair" agreement between the five test methods. Most of the higher agreement scores were seen between Golub's method and any one of the other methods.

In order to assess the consistency between gene and pathway based inter-test similarities, the matrices reported in Table 2 and Table 3 were correlated to each other by means of Pearson's correlation coefficient calculated between corresponding elements. The correlation coefficient of r = 0.916 points to a substantial equivalence between gene and pathway information and rules out possible bias due to differences in gene assignment to different pathways as a function of the databases used.

When the inter-test concordance matrix based on the diagnosis reported by Golub et al was correlated with the corresponding gene and pathway based matrices, we saw a complete lack of concordance between biological (gene and pathway) and clinical (diagnosis) levels (r = 0.055 and r=0.916 for gene-diagnosis and pathway-diagnosis consistency).

Table 4 reports the inter-test pairwise k-scores for these three levels of analysis while Figure 1 shows the mutual correlation between gene-pathway and gene-patient inter-test correlation matrices.

Discussion

DNA microarrays have much evolved over the past decade to become a dominant technology in the life sciences. The huge promise of this technology lies in its capacity to carry out high throughput analysis. As such, it is of great interest for both research in basic (patho-) physiology, as well as for screening assays towards "biomarkers" in an industrial setting. Microarrays produce vast amounts of data; analysis of this data provides information that ultimately serves to increase knowledge.

Data, information and knowledge are closely related, but separate entities. They may be defined as simple observation ("data", e.g. differential expression of signals on a microarray), data with relevance and purpose ("information" e.g. an expression signature used to classify samples) and valuable information from the human mind ("knowledge" e.g. different prognosis for AML and ALL cases classified by microarray through differential expression signals).[4]

In a situation where data is commonly analyzed by variant approaches, it becomes imperious to ascertain the robustness of information gain and knowledge creation. In clinical settings, the robustness of information is commonly assessed by Cohen's κ-score. This κ-score tests whether the inter-tester agreement is factual, or a product of chance. We used this simple and elegant method to assess to what extent the information gain from microarray data used for pathway discovery is more than a chance product.

The overall κ-score of -28.01 is indicative of below chance agreement only: the agreement between the selected signatures is arbitrary, despite the formal similarities between the tests. This indicates that a per-chance choice of method is more likely to

succeed. Golub's method is on average the 'most correlated' with the other tests, giving it a 'central' position. Both at the stage of gene selection, and of pathway identification, the inter-test agreement between the five statistical tests is, on average, only "fair" for selection and "moderate" for identification. Moreover, the gene and pathway based mutual similarities between tests are highly correlated, thus showing that the same basic information is carried by gene and pathway based analyses.

The low kappa scores at the selection stage can be explained by the large "marginal imbalance" between selected genes (n=50) and unselected genes (n=6095)[5,6]. This, however, is not the case at the stage of pathway identification which in turn was completely consistent with the gene selection procedure with regards to the inter-test mutual relations. Genes selected by different statistical methods may differ in name, but be involved in the same physiological functional systems of pathways because of redundancy of genes on the chip. This 'regularization' effect, suggested by the enhancement of average k-scores going from gene to pathway level (from 0.344 to 0.515), is in any case marginal and does not substantially alter the inter-test agreement structure (r = 0.916 between gene and pathway based k-scores).

The kappa scores for the pathway involvement show that the same data set, depending on which statistical method the researcher chooses to make use of, may provide considerably different "knowledge" gain. In other words, depending on which test is used on the same data set, different pathways are considered to be involved in the condition. Essentially the same conclusion, expressed in opposite terms, was drawn by Suarez- Farinas et al[7], who showed that a normalized assessment of raw data from the

same tissue analyzed on different microarray platforms increased the consistency of results.

Even more puzzling are the different findings obtained at the level of 'case sample clustering', i.e. the clinical (as opposed to the biological) level of appreciation of microarray data. In this case the inter-test concordance is much higher (average k-score 0.75, range between 0.60 to 0.89), pointing to a much more robust information content. The inter-test similarity structure arising from the case sample classification is completely independent from the gene and pathway based similarity structures (Figure 1 and Table 2). This points to a relative independence of the clinical judgement from the related biological explanation. In other words, it seems as if patients 'respond' with different expression signatures to the same disease condition, nevertheless maintaining an 'invariant gene expression signature' at the level of the whole expression pattern that cannot be further decomposed to the level of single genes.

In a recent study, Michiels et al. investigated the extent by which the composition of the "training set" (i.e. what samples are chosen to train the classifying algorithm) influences the final classification of "case samples"[8]. Depending on which "training samples" were chosen, the level of misclassification varied considerably. The larger the training set, the more robust the diagnosis became. In other words, an algorithm learns to correctly identify subclasses in the same way a physician does during his training: through studying as many training samples as possible. This opinion directly challenges the "robustness" of findings thought to be the main advantage of simply using gene clusters ("signatures") from microarray expression profiles as "biomarkers" in study designs with few samples, but high dimensionality of signals[9].

Do the observations demonstrated here disqualify microarray analysis as a valid technology to study biological phenomena? In our opinion, not at all. Technical problems concerning the reproducibility of findings even on the same platform have sparked research that resulted in more reliable data acquisition and analysis methods. By today, much information is gained from studies that use across- platform approaches and have advanced from the lab bench close to the bedside[10,11]. In particular, Suarez- Farina provided evidence that a normalized approach to data analysis will result in higher coherence of results[7], while Michiels has drawn attention to the need to critically appraise this information before eventually accepting the newly gained knowledge[8]. Miller et al., whose objective was to test the diagnostic value of an expression signature, moreover recognized from thorough analysis of their data the primary importance of the p53 functional status in predicting clinical breast cancer behavior[12]- that is, the quantitative or systemic behavior of the *pathway* rather than the common *biomarker* p53 predicts clinical outcome. In the setting of Miller et al's study, the biomarker p53 is inferior with regards to prognostic accuracy compared to the DNA microarray expression "signature" described. If considered as a diagnostic tool, DNA microarrays achieve high agreement scores on par with experienced clinicians[13]. The problem of the large difference in dimensionality between the number of signals and the number of samples, as is common in microarray data, necessitate the scientist to decide on a trade-off against other techniques (e.g. PCR) where there is less dimensionality difference. This problem is a current focus of research interest; one possible solution may be a reversal of the matrix, that is, to classify large numbers of expression signals based on small numbers of clinical samples. In this case, from a purely statistical point of view, the problem loses it

degeneracy, because the genes (now considered as statistical units) become much more numerous than the biological samples (now considered as variables). This last research avenue was opened (among others) by Landgrebe et al.[14] and by Tritchler et al.[15] who applied the method to Golub's data, too. This new perspective is extremely interesting as it implies a complete change of emphasis from considering single genes as 'efficient causes' of the disease towards understanding co-regulation networks of genes as 'measurable effects' of the disease.

Taken together, the analysis presented here shows that even after the critical analytic step of signal processing and normalizing, variation in analytical procedures may reduce the coherence of the conclusion.

The observations reported in the present study should serve two immediate purposes. Firstly, the findings remind both technology developers and users that only validated information becomes knowledge. In view of the low agreement scores for pathway discovery shown here, information from microarray analysis must be considered very critically and not be accepted as knowledge too easily.

Secondly, and in analogy to the technological refinements that have been put into place, our observations should encourage research into the refinement of statistical analysis methods.

On a more general level, our results substantiate the need to progressively abandon the 'single gene' or even 'pathway' level of analysis in order to look at a different level of physiological co-regulation modes.

Methods

We assessed the inter- test agreement as described by Cohen's κ-score between five statistical tests [16-20] used in the analysis of DNA microarray expression data[2]. The data set analyzed was from the well characterized study on acute leukemia by Golub et al. which has a detailed methodology section that allows stepwise replication of the analysis[16]. The authors have updated their methodology and allow download from a dedicated website[21]. Appendix Part A gives a brief explanation of the five tests used, of Cohen's kappa score, and of the original dataset. At the level of gene selection, Golub et al. employ an algorithm that requires specification of the number of genes one hopes to select from the "training" set of samples. Accordingly, we used alternative statistical approaches[17-20] with the same formal requirement for comparison. The software GenePattern[21] was used for data preprocessing. The preprocessed data was imported into the program R (The R Foundation for Statistical Computing, Version 2.1.1), in which all downstream analysis was performed (see Appendix Part B for the program codes). We submitted the gene sets ("signatures") selected by each of the five classification methods to three databases that provide pathway information: BioCarta[22], KEGG PATHWAY[23], and Reactome[24] (Appendix C). The gene sets selected from the training samples used by Golub and the pathways identified were then compared by calculating the kappa scores. κ-scores are reported at the level of gene-selection and pathway identification on the "training" samples.

The matrices reporting the pairwise inter-tests κ-scores relative to the concordance were correlated at the level of gene-selection and pathway identification so

as to assess the consistency of the observed inter-test similarities at the gene and pathway levels. Both these matrices were then compared with the "case sample clustering" matrix reporting the inter-tests κ-scores agreement with clinical diagnosis.

Authors' contributions

AM conceived the study design and wrote most of the manuscript. ML did the statistical analysis and wrote part of the methods section. AG contributed to statistical analysis and wrote part of the discussion section.

Acknowledgments

The study was financed by an intramural research grant from Singapore's Biomedical Research Council. The authors thank Mr. Liu Xiao Xing for his contributions in the earlier stages of this study, in particular his work on the data preprocessing.

References:

1. Segal, E., Friedman, N., Kaminski, N., Regev, A. & Koller, D. From signatures to models: understanding cancer using microarrays. *Nat Genet* **37 Suppl**, S38-45 (2005).
2. Cohen, J. A coefficient for agreement of nominal scales. *Educational and Psychological Measurement* **20**, 37-46 (1960).
3. Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**, 378-382 (1971).
4. Davenport, T.H. *Information Ecology*, (Oxford University Press, New York, 1997).
5. Feinstein, A.R. & Cicchetti, D.V. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* **43**, 543-9 (1990).
6. Cicchetti, D.V. & Feinstein, A.R. High agreement but low kappa: II. Resolving the paradoxes. *J Clin Epidemiol* **43**, 551-8 (1990).
7. Suarez-Farinas, M., Noggle, S., Heke, M., Hemmati-Brivanlou, A. & Magnasco, M.O. Comparing independent microarray studies: the case of human embryonic stem cells. *BMC Genomics* **6**, 99 (2005).
8. Michiels, S., Koscielny, S. & Hill, C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* **365**, 488-92 (2005).
9. Grimwade, D. & Haferlach, T. Gene-expression profiling in acute myeloid leukemia. *N Engl J Med* **350**, 1676-8 (2004).
10. Rabson, A.B. & Weissmann, D. From microarray to bedside: targeting NF-kappaB for therapy of lymphomas. *Clin Cancer Res* **11**, 2-6 (2005).

11. Lam, L.T. et al. Small molecule inhibitors of IkappaB kinase are selectively toxic for subgroups of diffuse large B-cell lymphoma defined by gene expression profiling. *Clin Cancer Res* **11**, 28-40 (2005).

12. Miller, L.D. et al. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A* (2005).

13. Loh, M., Liu, X.X., Dong, P. & Mondry, A. Diagnostic performance of five statistical methods commonly used in the analysis of DNA microarrays. in *4th Asia-Pacific Bioinformatics Conference (APBC2006)* (Taiwan, 2006).

14. Landgrebe, J., Wurst, W. & Welzl, G. Permutation-validated principal components analysis of microarray data. *Genome Biol* **3**, RESEARCH0019 (2002).

15. Tritchler, D., Fallah, S. & Beyene, J. A spectral clustering method for microarray data. *Computational Statistics & Data Analysis* **49**, 63-76 (2005).

16. Golub, T.R. et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-7 (1999).

17. Ma, X.J. et al. Gene expression profiles of human breast cancer progression. *Proc Natl Acad Sci U S A* **100**, 5974-9 (2003).

18. Liu, X., Krishnan, A. & Mondry, A. An entropy-based gene selection method for cancer classification using microarray data. *BMC Bioinformatics* **6**, 76 (2005).

19. Tusher, V.G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-21 (2001).

20. Efron, B., Tibshirani, R., Storey, J.D. & Tusher, V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151-1160 (2001).

21. GenePattern. (http://www.broad.mit.edu/genepattern).

22. BIOCARTA.

23. KEGG PATHWAY Database.

24. Reactome.

25. Broad Institute. (http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_menu.cgi).

26. PubMed. (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi).

Tables

	t-test	Golub	SAM	EBAM	nMigs
t-test	50	38	26	31	8
Golub		50	38	25	5
SAM			50	18	3
EBAM				50	11
nMigs					50

Table 1: Overview of number of genes selected by each pair of the five statistical methods

	Gene	Pathway	Patient
Gene	1.00	0.92 (0.0002)	0.06 (0.8788)
Pathway		1.00	0.18 (0.6275)
Patient			1.00

Table 2: Kappa scores for agreement of "gene selection from training set" between the five methods (average = 0.34; "fair")

	t-test	Golub	SAM	EBAM	nMigs
t-test	1.00	0.61	0.30	0.79	0.55
Golub		1.00	0.60	0.51	0.53
SAM			1.00	0.34	0.44
EBAM				1.00	0.48
nMigs					1.00

Table 3: Kappa scores for agreement of "pathway involved" between the five methods (average = 0.52; "moderate")

Pairname	Gene	Pathway	Patient
t-Golub	0.52	0.61	0.85
t-SAM	0.15	0.3	0.6
t-EBAM	0.68	0.79	0.75
t-nMigs	0.25	0.55	0.75
Golub-SAM	0.54	0.6	0.65
Golub-EBAM	0.4	0.51	0.79
Golub-nMigs	0.31	0.53	0.68
SAM-EBAM	0.15	0.34	0.78
SAM-nMigs	0.19	0.44	0.78
EBAM-nMigs	0.25	0.48	0.89

Table 4: Inter-test pairwise k-scores at the levels of gene, pathway and patient identification

Figures

Figure 1: Mutual correlation between (a) gene-pathway and (b) gene-patient inter-test correlation matrices.