Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Biology and Chemistry 28 (2004) 3–10
Computational Biology and Chemistry
www.elsevier.com/locate/compbiolchem

The operons, a criterion to compare the reliability of transcriptome analysis tools: ICA is more reliable than ANOVA, PLS and PCA

Anne-Sophie Carpentier*, Alessandra Riva, Pierre Tisseur, Gilles Didier, Alain Hénaut

Laboratoire Génome et Informatique, UMR 8116, Tour Evry2, 523 Place des Terrasses, 91034 Evry, France

Received 5 September 2003; received in revised form 3 December 2003; accepted 3 December 2003

Abstract

The number of statistical tools used to analyze transcriptome data is continuously increasing and no one, definitive method has so far emerged. There is a need for comparison and a number of different approaches has been taken to evaluate the effectiveness of the different statistical tools available for microarray analyses.

In this paper, we describe a simple and efficient protocol to compare the reliability of different statistical tools available for microarray analyses. It exploits the fact that genes within an operon exhibit the same expression patterns. In order to compare the tools, the genes are ranked according to the most relevant criterion for each tool; for each tool we look at the number of different operons represented within the first twenty genes detected. We then look at the size of the interval within which we find the most significant genes belonging to each operon in question. This allows us to define and estimate the sensitivity and accuracy of each statistical tool.

We have compared four statistical tools using *Bacillus subtilis* expression data: the analysis of variance (ANOVA), the principal component analysis (PCA), the independent component analysis (ICA) and the partial least square regression (PLS). Our results show ICA to be the most sensitive and accurate of the tools tested.

In this article, we have used the protocol to compare statistical tools applied to the analysis of differential gene expression. However, it can also be applied without modification to compare the statistical tools developed for other types of transcriptome analyses, like the study of gene co-expression.

Keywords: Operon; Criterion of comparison; Transcriptome; Expression analysis; ANOVA; ICA; PCA; PLS

1. Introduction

1.1. A word about microarrays

1.1.1. Definition of microarrays

A microarray consist of a solid support on which a series of DNA segments is arranged and fixed in a regular pattern. These segments are incubated with a labeled nucleic acid sample. When a nucleic acid sequence in the sample is complementary to a DNA segment present on the support, it will bind and hybridize to this, specific segment. This hybridization is recorded and analyzed.

1.1.2. The historical background

As Jordan (2002) points out, DNA arrays were already being used in the seventies, in the form of dot blots and slot blots. Ekins et al. developed microspot fluorescent immunoassays in the late eighties and early nineties, proving that the sensitivity of these miniaturized assays was comparable to that of "macroscopic" ones and introducing the concept of micro-array (Ekins, 1989; Ekins et al., 1990; Ekins and Chu, 1991). The concept of miniaturization was also applied to DNA arrays, using two different approaches. One was to deposit the DNA (or complementary DNA) on glass plates, leading to the first publication of a gene expression microarray article in 1995 (Schena et al., 1995). The second approach was that of the oligonucleotide array, where the DNA is directly synthesized onto the support (Fodor et al., 1991; Southern et al., 1992).

* Corresponding author. Tel.: +33-1-60-87-38-74; fax: +33-1-60-87-38-97.
E-mail address: carpentier@genopole.cnrs.fr (A.-S. Carpentier).
URL: http://195.221.65.10:1234/~carpenti/.

1476-9271/$ – see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2003.12.001

1.1.3. Today's microarrays

In the following, "probe" denotes the immobilized DNA on the support and "target" the mobile DNA, cDNA or mRNA. Some authors, however, use the terms the other way round.

The *supports* used for microarrays today are either glass (microscope) slides, (nylon) membranes or silicon chips. The *material fixed* on the support ("probe") can be:

- DNA, representing coding sequences or, more generally, pieces of genomic DNA.
- complementary DNA, obtained from the mRNA of specific genes or expressed sequence tags (ESTs). The latter is usually used for organisms not yet completely sequenced.
- Oligonucleotides; in the case of oligonucleotide arrays the oligos are synthesized directly onto a silicon chip; this process has been pioneered by Affymetrix (see Lipshutz et al. (1999) for a comprehensive review on oligonucleotide arrays).

The *mobile* "*target*" can be:

- DNA,
- complementary DNA (cDNA), obtained from mRNA by reverse transcriptase-PCR (RT-PCR),
- mRNA; this can be used although cDNA is generally preferred.

A hybridization mixture is obtained by labeling the target fluorescently or radioactively. This mixture is then incubated with the prepared microarray and allowed to hybridize with the probe. Finally, the resulting signal intensity, that correlates with the amount of captured probe, is measured, stored in a computer and then analyzed.

Recently, efforts have been made to extend the microarray technology to the field of proteins. The interested reader may refer to the review written by Templin et al. (2002) for a comprehensive introduction to this field.

1.2. Applications

Microarrays can be used for the detection of mutations, DNA sequencing and the analysis of gene expression. The latter application has been gaining in importance and we will focus our attention on this aspect. As microarrays allow measuring the expression levels of thousands of genes at the same time, this opens the possibility to identify differentially expressed genes (Callow et al., 2000) and to cluster those genes sharing similar expression patterns (Heyer et al., 1999). They have become a widespread tool for analyzing the relative transcription levels of genes.

Microarrays have a widespread use, including:

- clinical medicine (see Joos et al. (2003) for a review on this subject);
- the study of the cell-cycle (see for example McCune and Donaldson, 2003);
- the study of the circadian rhythm in animals (see for example Stanewsky, 2003) and plants (see for example Davis and Millar, 2001); and
- the study of plant metabolism (see for example Buckhout and Thimm, 2003).

For further information on microarray technology, the reader may refer to recent review articles (Barrett and Kawasaki, 2003; Vrana et al., 2003); he may also refer to a related NCBI web page (http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html).

1.3. The analysis of the microarray data

Different tools have been developed for or adapted to the analysis of the huge amount of data created in microarray experiments (Draghici, 2002). The number of tools is continuously increasing and no one, definitive method has so far emerged, as is exemplified by the web-site maintained by Li, which has a continuously growing collection of articles on microarray data analysis (http://www.nslij-genetics.org/microarray/).

There is a need of comparing the tools, but identifying an unbiased and biologically relevant criterion for the comparison is difficult (He et al., 2003). A number of different approaches has been taken to compare the effectiveness, or reliability, of the different statistical tools available for microarray analyses.

Some are based on artificial data to define precisely the specificity and sensitivity of these statistical tools (Pan, 2003; Reiner et al., 2003).

Others are based on experimental data. The quality of a statistical tool can be measured by the number of differentially expressed genes which it reveals. A statistical parameter like the *P*-value may be used (Pan, 2002).

Finally some authors combine two criteria, the number of identified genes and their physiological coherence, based on an a priori knowledge of the biological phenomenon studied (Troyanskaya et al., 2002).

1.4. This paper

In this paper, we try to establish a protocol for the comparison of statistical tools (available for microarray analysis) which is objective, reflects a biological reality and is not bound to one, particular set of experimental conditions. It is based on the expression coherence of genes belonging to the same operon. In bacteria, a number of genes are organized in operons, that is to say clusters of contiguous genes transcribed from one promoter.

A good and reliable statistical tool is one that, when detecting an over- or under-expression for a gene belonging to an operon, also detects this pattern for the other genes belonging to this operon. Indeed, it has been shown that the genes within an operon exhibit the same expression patterns (Murray et al., 2001; Sabatti et al., 2002; Wei et al., 2001; Zimmer et al., 2000).

This criterion, based on the expression coherence of genes belonging to the same operon, therefore reflects a biological property that is not bound to a particular set of experimental conditions. Furthermore, it is independent of the statistical laws (for example Gaussian) governing the variations of the gene expression.

We have tested this criterion on four statistical tools using *Bacillus subtilis* expression data (Sekowska et al., 2001): The analysis of variance (ANOVA), the principal component analysis (PCA), the independent component analysis (ICA) and the partial least square regression (PLS). Note: ANOVA and PLS need the a priori definition of factors, which could influence the level of gene expression; ICA and PCA do not need the definition of any factor for their use.

Two of these tools (ANOVA and PCA) are frequently used for microarray analyses. The other two methods tested (ICA and PLS) have only been recently applied to the analysis of microarray data. All of these methods are used in many other fields.

- The analysis of variance is a classical statistical method for the analysis of fully crossed factorial designs. Its use on microarray data has allowed the identification of differentially expressed genes (Kerr and Churchill, 2001; Kerr et al., 2000).
- The principal component analysis is used to reduce gene space dimension and allows the detection of the major sources of variation (Landgrebe et al., 2002; Peterson, 2003).
- Originally developed for chemometric data (Wold, 1973), the term partial least square regression regroups several methods. PLS has been used in proteome and transcriptome analysis to classify benign and malignant tumours (Alaiya et al., 2000; Cho et al., 2002; Musumarra et al., 2001) or to reduce gene space (Nguyen and Rocke, 2002). In this article, we use PLS to identify differentially expressed genes.
- Independent component analysis (ICA) was originally developed (Comon, 1994) for analyses related to the "cocktail party problem". Its applications in transcriptome analysis include the identification of groups of genes implicated in cancer, the study of the cell cycle (Liebermeister, 2002) and to identify genes that are potentially co-regulated (Chiappetta et al., 2002(personal communication), http://www.cmi.univ-mrs.fr/~torresan/publi.html).

In this article, we set out to compare the four statistical tools mentioned above. However, our method of comparison may be applied to any other statistical tool used in the analysis of microarray data.

2. Methods

2.1. Data

The microarray data used in this study stem from experiments on the sulphur metabolism of *B. subtilis* (Sekowska et al., 2001). The experiments were carried out using Panorama nylon filters *B. subtilis* gene arrays (Sigma–GenoSys Biotechnologies); each array contained all of *B. subtilis'* genes and one gene is represented by one spot. Each gene spot is represented twice on the array.

The aim of these experiments was to identify the genes differentially expressed when the bacteria are grown with methionine or methyl-thioribose as sulphur source. The experiments followed a fully crossed factorial design (Fig. 1) with four factors (sulphur source, day of experiment, amount of RNA used and duplicate of each spot). The data (raw levels of expression) were gathered in an array of 4107 rows (all *B. subtilis* genes) and 16 columns (experimental conditions).

We have used the logarithm (base 10) of these raw data in order to remove much of the proportional relationship between random error and signal intensity (Nadon and Shoemaker, 2002). We have normalized the data (mean equal to 0 and variance equal to 1 for each experimental condition) because two methods (PCA and ICA) need normalized data.

In some parts of the article, the data will be referred to as a cloud of 4107 points (the genes) in a 16-dimensional space (the experimental conditions). In this paper, we will not exploit the dual representation (the 16 experiments in the 4107-dimensional space).

2.2. Programs used

ANOVA, PLS and PCA were carried out using a program called GeneANOVA (Didier et al., 2002). ICA is an adaptation of FASTICA Hyvarinen's fixed-point algorithm (Hyvarinen, 1999) made by Chiappetta and Torrésani (Chiappetta et al., 2002 (personal communication), http://www.cmi.univ-mrs.fr/~torresan/publi.html).

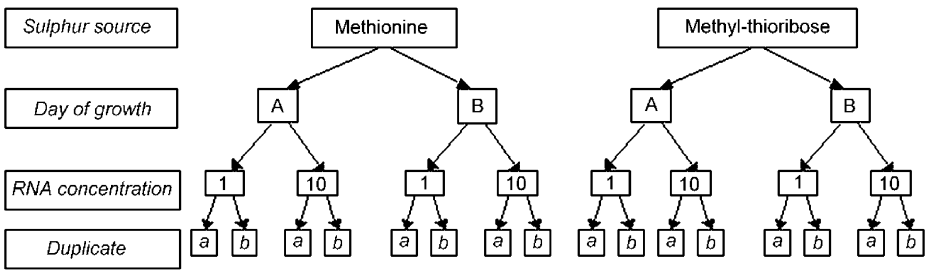

Fig. 1. Experimental design of the transcriptome analysis on *Bacillus subtilis* (Sekowska et al., 2001).

2.3. Choice of parameters

We have chosen to analyze the expression data for the two experimental factors "sulphur source" and "day of experiment".

For ICA and PCA, the axes which correspond to these two factors are determined a posteriori: one determines the relative weight of each of the 16 components whose combination defines the axes; the axes retained are those where either the component "sulphur source", or the component "day of experiment" plays a major role. The factor "day" corresponds to the third axis and the factor "sulphur source" to the fifth. The fourth axis corresponds to an interaction between these two factors.

For each gene, the equation used for ANOVA is the following:

$$Y_{ijkl} = \mu + S_i + J_j + C_k + D_l + \varepsilon_{ijkl}$$

where Y_{ijkl} is the gene intensity; μ the mean of the intensities of expression measured for the gene; S_i, J_j, C_k and D_l are, respectively, the effects of sulphur source i, experiment day j, RNA concentration k and duplicate l on the gene intensity; and ε_{ijkl} is the residual error.

There are 16 measurements per gene. Five degrees of freedom are lost for the estimation of the mean and the variances of the four factors. The residual variance ε_{ijkl} has 11 degrees of freedom. It encompasses all interactions: between two factors (6), between three factors (4) and between four factors (1).

F = "variance of the factor of interest"/"residual variance" with one degree of freedom in the numerator and 11 degrees of freedom in the denominator.

The interactions between the factors were not estimated because of the experimental design and the low degree of freedom obtained.

2.4. Operons

We need to know how the genes of *B. subtilis* are organized into operons. A presumed operon is defined as a group of contiguous genes that are on the same reading strand delimited either by a promoter and a terminator (predicted or not) or a gene, which lies on the other DNA strand. This allowed to find the operons in *B. subtilis* (Subtilist, http://genolist.pasteur.fr/SubtiList/). We compiled a list in which each gene is either assigned to an operon or defined as an isolated gene. This list may be consulted at http://195.221.65.10:1234/∼carpenti/. Even if some predicted operons will prove to be artefacts, this will only introduce a systematic bias for all the statistical tools tested. This will not raise any problem for the comparison of the statistical tools and it will not influence our conclusions about the quality of these tools with respect to each other.

2.5. Evaluating procedure

To compare statistical tools, one needs to define quantitative criteria that will measure the "tool reliability": sensitivity, accuracy and the detection of false positives need to be evaluated.The following procedure was applied:

1. The genes are ranked as a function of their expression changes (rank #1 is the most significant).
2. "Detected Operons" are identified based on the ranks (one gene with rank ≤ 20 and another gene with rank ≤ 100).
3. The most significant interval (MSI) is determined.
4. False positives are evaluated (MSI ≥ 700).
5. "Relevant detected operons" are identified (MSI < 700).
6. The accuracy of a "relevant detected operon" is evaluated (MSI < 150).
7. The sensitivity of a tool is evaluated.

2.5.1. Ranking of the genes

In order to compare the four tools under the best possible conditions, the genes are ranked according to the most relevant criterion for each tool, that is to say, the one that gives the most coherent results for the tool:

- for ANOVA, the *P*-value obtained for each gene;
- for PLS, the weight of the gene for the axis determination; and
- for PCA and ICA, the remoteness from the cloud center of the projection of the gene on the axis studied.

We thus obtain for each tool a list of genes, ranked according to a specific criterion; the most significant gene has rank #1. The order of the genes on the lists obtained may differ from each other.

2.5.2. Identification of "detected operons"

We define an operon to be detected ("detected operon") by a tool if at least one of its genes has a rank $\leq$20 and another of its genes a rank $\leq$100. For the assignment of genes to operons, we have used the list which may be consulted at http://195.221.65.10:1234/∼carpenti/. It should be noted that a priori the "detected operons" may be different for the various tools tested.

A possible bias of this method presents itself in one particular case: If one of the "detected operons" is very large, a considerable proportion of the genes with a rank $\leq$20 will belong to this particular operon, leaving "no place" for the other operons to be detected. The same problem may arise if a large number of isolated genes (not belonging to an operon) are highly relevant. As this possible bias will be present for all four statistical tools tested, it will not raise any problem for the comparison of the statistical tools and it will not influence our conclusions about the quality of these tools with respect to each other.

Note: the choice of "20; 100" is an arbitrary one. In order to establish whether this choice might affect the results and thus the conclusions of this paper, we have also run through

the procedure using, successively "10; 50" and "40; 200" for the identification of "detected operons". The results may be consulted at http://195.221.65.10:1234/~carpenti/ (see also Section 3).

2.5.3. Determination of the most significant interval

In order to facilitate the analysis and comparison of the statistical tools, we introduce the most significant interval (MSI). It is calculated for each "detected operon" in the following manner:

$$\mathrm{MSI}_j = \mathrm{median}_j - \mathrm{first}_j$$

where MSI_j is the MSI of "detected operon" j, median_j is the median of the rank values of the genes belonging to "detected operon" j, and first_j is the smallest rank value within "detected operon" j

2.5.4. Evaluation of false positives

The reliability of a statistical tool will also be measured by the absence of false positives.

For the definition of false positives, we exploit the fact that each gene spot had been duplicated on the microarrays and any difference measured for two spots belonging to the same gene cannot have a biological cause. We ranked the genes according to this "duplicate factor '', as described under point 1 and identified "detected operons" as described under point 2. As there is no biological cause for this detection, we find ourselves with false positives.

As before, a priori the false positives detected may be different for the various tools tested.

The results of this analysis lead us to conclude that a "detected operon" is a false positive when MSI $\geq$ 700 (see Table 1 for details).

Table 1
Quantification of false positives

Operon name	Operon size	MSI (most significant interval)			
		ANOVA	PLS	PCA	ICA
fliLMY cheY fliZPQR flhBAF ylxH cheBAWCD sigD ylxL	19	2385	2243	1193	2499
yonRSTUVX yopAB	8	61	134	127	251
hemAXCDBL	6	1360	1547		
ruvAB queA tgt yrbF	5			1005	707

For the definition of false positives we exploit the fact that each gene spot had been duplicated on the microarrays and any difference measured for two spots belonging to the same gene cannot have a biological cause. We ranked the genes according to this "duplicate factor", as a function of the differences in their expressions, then identified "detected operons" and calculated the MSI (see Section 2 for details). As there is no biological cause for this detection, we find ourselves with false positives; they are characterized by a large MSI; this leads us to conclude that a "detected operon" is a false positive when MSI $\geq$ 700. One exception is the operon *YonRSTUVXyopAB*, detected by all four tools, with small MSIs. As we cannot give a biological reason, we suspect that its detection is due to a default on the microarray used in the experiments.

2.5.5. Identification of "relevant detected operons"

The definition of "relevant detected operons" follows from the definition of false positives:

"relevant detected operons" have an MSI $<$ 700.

2.5.6. Evaluating the accuracy of a "relevant detected operon"

We define that an operon is detected with good accuracy if its MSI is lower then a given threshold. This threshold was determined such that 80% of the "detected operons" have a MSI below the threshold. Our results lead us to state that: Operons detected with good accuracy have an MSI $<$ 150.

2.5.7. Evaluating the sensitivity of the tools

The sensitivity of the tools is estimated by comparing the number or "relevant detected operons" identified by each tool.

2.6. The comparison of the tools under three typical experimental conditions

We have decided to compare the four statistical tools under three experimental conditions biologists are frequently faced with:

- *The experimental factor is identified and fully controlled*: In the case of the microarray data used in this study, this factor is the sulphur source contained in the growth medium. In one case the sulphur source was methionine, in the other case it was methylthioribose. These two compounds are metabolically closely related. The four statistical tools were tested on these experimental data. The results obtained are displayed in Table 2.
- *The experimental factor is identified but not under control*: In this case it was "day". The experiments were carried out twice, on different days. The protocol followed was the same on these 2 days; however, parameters like "room temperature" were not necessarily the same, thus introducing a factor in the experimental setup that was identified but not under control. The results obtained are displayed in Table 3.
- *The interaction between experimental factors*: The aim of a protocol is to separate completely the different experimental factors. However, the expression of certain genes may be under the control of more than one factor. In this case, one talks of an "interaction between experimental factors". ANOVA and PLS are adapted to the analysis of variations due to a single experimental factor; they are not well suited for the study of interactions between factors; they were not tested under this condition. On the other hand, ICA and PCA are well adapted to cope with possible interactions; these interactions are identified because more than one factor plays a major role in the definition of an axis. The results obtained are displayed in Table 4.

Table 2
Comparison of the statistical tools when the experimental factor is identified and fully controlled

Operon name	Operon size	MSI (most significant interval)			
		ANOVA	PLS	PCA	ICA
yqiXYZ	3	1	1	3	6
argCJBD carAB argF	7	15	28	201	56
argGH ytzD	3	1	1	6	2
ahpCF	2	46	7	11	13
lctEP	2	26		36	8
levDEFG sacC	5	316	220		
sunAT yolIJK	5		**635**		13
ydcPQRST yddABCDEFGHIJ	15			**1313**	116
flgM yvyG flgK yviE yviF csrA hag	8				509
yxbBA yxnB asnH yxaM	5			15	
ytmIJKLM hisP ytmO ytnIJ ribR hipO ytnM	12			92	
fliLMY cheYfliZPQR flhBAF ylxH cheBAWCD sigD ylxL	19				350
Relevant detected operons		6	6	7	9

The identified and controlled experimental factor is the sulphur source (either methionine, or methylthioribose). Genes were ranked as a function of the differences in their expressions, false positives (MSI $\geq$ 700) and "relevant detected operons"(MSI $<$ 700) were identified (see Section 2 for details). The bold entry for PLS, with MSI = 635 is estimated to be a borderline case for a false positive; it has been included for PLS's total of "relevant detected operons". Note that only PCA detects a false positive (shaded entry). ICA is the most sensitive tool under these experimental conditions, identifying the largest number of "relevant detected operons". ANOVA and PLS are the least sensitive.

3. Results and discussion

Microarrays are defined as a tool for analyzing gene expression that consists of a small membrane or glass slide containing samples of many genes arranged in a regular pattern. They are widely used for analyzing the relative transcription level of genes. The number of statistical tools for analyzing the huge amount of data created in the experiments is continuously growing and no-one of these tools has yet emerged as the definitive one.

We have developed a protocol for the comparison of statistical tools applied to the analysis of transcription data. We have applied this method to compare four statistical tools (ANOVA, PLS, ICA and PCA) under three typical experimental conditions. All four tools were compared under two of these conditions (see Tables 2 and 3 for details), whilst only ICA and PCA, which do not need the a priori definition of experimental factors, could be tested under the third condition (see Table 4 for details).

Based on our observations, we have defined threshold values to define "relevant detected operons" (MSI $<$ 700), false positives (MSI $\geq$ 700) and to define a good accuracy (MSI $<$ 150); the sensitivity of the tools is estimated by comparing the number of "relevant detected operons" identified by each tool.

Table 3
Comparison of the statistical tools when the experimental factor is identified but not under control

Operon name	Operon size	MSI (most significant interval)			
		ANOVA	PLS	PCA	ICA
comGABCDEFG yqzE	8	16	26	6	4
comFABC yvyF	4	339		66	19
cotVWXYZ	5		147	315	417
groESL	2		15		
yvaVWXY	4			53	
yqxM sipW cotN	3			79	
comEABC	3				35
Relevant detected operons		2	3	5	4

The experiments were carried out twice, on different days, using the same protocol; however, parameters like "room temperature" were not necessarily the same on the 2 days, introducing an identified but not controlled factor. PCA and ICA are the most sensitive tools, whilst ANOVA is the least sensitive (please refer to the legend of Table 2 for details about the classification procedure).

Table 4
Comparison of the statistical tools to detect possible interactions between the experimental factors

Operon name	Operon size	MSI (most significant interval)	
		PCA	ICA
purMNHD	4	71	57
ybaC rpsJ rplCDWB rpsS rplV rpsC rplP rpmC rpsQ rplNXE rpsNH rplFR rpsE rpmD rplO secY adk map	25	51	56
alsS alsD	2		25
rpsL rpsG fus tufA	4		21
yvaVWXY	4		73
yxbBA yxnB asnH yxaM	5		126
yyaEF rpsF ssb rpsR	5	408	
Relevant detected operons		3	6

The expression of certain genes may be under the control of more than one factor, leading to an interaction between experimental factors. Only ICA and PCA are well adapted to cope with possible interactions; these interactions are identified because more than one factor plays a major role in the definition of an axis. ICA is more sensitive than PCA (please refer to the legend of Table 3 for details about the classification procedure).

Table 5
Overview of the results

	ANOVA	PLS	PCA	ICA
Relevant detected operons				
Tables 2–4	8	9	15	19
Tables 2 and 3	8	9	12	13
Accuracy of detection (%)				
Tables 2–4	75	78	80	84
Tables 2 and 3	75	78	83	77

The table sums up the results obtained in this study. The first part of the table relates to the number of "relevant detected operons" identified and thus to the tools' relative sensitivities. "Tables 2–4": adding the results from Tables 2–4, the total of "relevant detected operons" has been calculated for each tool. The entries for "Tables 2 and 3" have been obtained accordingly. Note that in both cases, ICA has the highest overall sensitivity, identifying the largest number of "relevant detected operons", whilst ANOVA is the least sensitive. The second part of the tables relates to the tools' accuracies: the percentage of "relevant detected operons" identified with a "good accuracy" (MSI < 150) has been calculated for each tool, adding the results from Tables 2–4 ("Tables 2–4"), etc. (see above). Overall, ICA has the highest accuracy, very closely followed by PCA, whilst ANOVA has the lowest accuracy.

Table 5 sums up the results obtained. Overall, we observe that ANOVA has the lowest sensitivity, whilst ICA is the tool with the highest sensitivity. The same observations can be made regarding the accuracies of the tools. It is interesting to note that even under the two experimental conditions for which ANOVA was conceived (Tables 2 and 3), it performs less well than ICA. PLS performs similarly to ANOVA. PCA has an intermediate performance. However, each tool may detect operons not identified by the other tools.

The results obtained by testing the four statistical tools show us that ICA has overall the best performance. This result holds true even if the criteria for "detected operon" are changed (instead of "20; 100" using "10; 50" or "40; 200", results not shown; see http://195.221.65.10:1234/~carpenti/ for details).

In this paper, we have set out to describe a simple and efficient protocol to compare the reliability of different statistical tools available for microarray analyses. The criterion used in our method is based on the expression coherence of genes belonging to the same operon. The method is objective, reflects a biological reality and is not bound to one, particular set of experimental conditions. It allows to compare the sensitivity, the accuracy and the detection of false positives of different statistical tools. As it is a comparative method, any bias linked to the criterion (for example uncertainties about the reality of a predicted operon) will influence in the same way the results obtained for each of the tools tested.

Here, we have used this method to compare statistical tools applied to the analysis of differential gene expression. However, the above protocol can also be applied without modification to compare the statistical tools developed for other types of transcriptome analyses, like the study of gene co-expression.

Acknowledgements

We are grateful to Antoine Danchin and Agnieszka Sekowska for having provided us with their data and to Bruno Torrésani and Pierre Chiappetta for the ICA program that they adapted to gene expression analysis. This work was supported by the French Industry Ministry contract ASG number 01 4 90 6093.

References

Alaiya, A.A., Franzen, B., Hagman, A., Silfversward, C., Moberger, B., Linder, S., Auer, G., 2000. Classification of human ovarian tumors using multivariate data analysis of polypeptide expression patterns. Int. J. Cancer 86 (5), 731–736.

Barrett, J.C., Kawasaki, E.S., 2003. Microarrays: the use of oligonucleotides and cDNA for the analysis of gene expression. Drug Discov. Today 8 (3), 134–141.

Buckhout, T.J., Thimm, O., 2003. Insights into metabolism obtained from microarray analysis. Curr. Opin. Plant Biol. 6 (3), 288–296.

Chiappetta, P., Roubaud, M.C., Torrésani, B., 2002. Blind Source Separation de Sources and the Analysis of Microarray Data, personal communication. http://www.cmi.univ-mrs.fr/~torresan/publi.html.

Callow, M.J., Dudoit, S., Gong, E.L., Speed, T.P., Rubin, E.M., 2000. Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. Genome Res. 10 (12), 2022–2029.

Cho, J.H., Lee, D., Park, J.H., Kim, K., Lee, I.B., 2002. Optimal approach for classification of acute leukemia subtypes based on gene expression data. Biotechnol. Prog. 18 (4), 847–854.

Comon, P., 1994. Independent component analysis—a new concept? Signal Process. 36, 287–314.

Davis, S.J., Millar, A.J., 2001. Watching the hands of the Arabidopsis biological clock. Genome Biol. 2 (3), e-pub.

Didier, G., Brezellec, P., Remy, E., Henaut, A., 2002. GeneANOVA–gene expression analysis of variance. Bioinformatics 18 (3), 490–491.

Draghici, S., 2002. Statistical intelligence: effective analysis of high-density microarray data. Drug Discov. Today 7 (11), S55–S63.

Ekins, R.P., 1989. Multi-analyte immunoassay. J. Pharm. Biomed. Anal. 7 (2), 155–168.

Ekins, R.P., Chu, F., Biggart, E., 1990. Multispot, multianalyte, immunoassay. Ann. Biol. Clin. (Paris) 48 (9), 655–666.

Ekins, R.P., Chu, F.W., 1991. Multianalyte microspot immunoassay—microanalytical "compact disk" of the future. Clin. Chem. 37 (11), 1955–1967.

Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. Science 251 (4995), 767–773.

He, Y.D., Dai, H., Schadt, E.E., Cavet, G., Edwards, S.W., Stepaniants, S.B., Duenwald, S., Kleinhanz, R., Jones, A.R., Shoemaker, D.D., Stoughton, R.B., 2003. Microarray standard data set and figures of merit for comparing data processing methods and experiment designs. Bioinformatics 19 (8), 956–965.

Heyer, L.J., Kruglyak, S., Yooseph, S., 1999. Exploring expression data: identification and analysis of coexpressed genes. Genome Res. 9 (11), 1106–1115.

Hyvarinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Networks 10 (3), 626–634.

Joos, L., Eryuksel, E., Brutsche, M.H., 2003. Functional genomics and gene microarrays—the use in research and clinical medicine. Swiss. Med. Wkly. 133 (3–4), 31–38.

Jordan, B., 2002. Historical background and anticipated developments. Ann. NY Acad. Sci. 975, 24–32.

Kerr, M.K., Churchill, G.A., 2001. Statistical design and the analysis of gene expression microarray data. Genet Res. 77 (2), 123–128.

Kerr, M.K., Martin, M., Churchill, G.A., 2000. Analysis of variance for gene expression microarray data. J. Comput. Biol. 7 (6), 819–837.

Landgrebe, J., Wurst, W., Welzl, G., 2002. Permutation-validated principal components analysis of microarray data. Genome Biol. 3 (4), 0019.0011–0019.0011.

Liebermeister, W., 2002. Linear modes of gene expression determined by independent component analysis. Bioinformatics 18 (1), 51–60.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R., Lockhart, D.J., 1999. High density synthetic oligonucleotide arrays. Nat Genet. 21 (1 Suppl.), 20–24.

McCune, H.J., Donaldson, A.D., 2003. DNA replication: telling time with microarrays. Genome Biol. 4 (2), 204.

Murray, A.E., Lies, D., Li, G., Nealson, K., Zhou, J., Tiedje, J.M., 2001. DNA/DNA hybridization to microarrays reveals gene-specific differences between closely related microbial genomes. Proc. Natl. Acad. Sci. U.S.A. 98 (17), 9853–9858.

Musumarra, G., Condorelli, D.F., Scire, S., Costa, A.S., 2001. Shortcuts in genome-scale cancer pharmacology research from multivariate analysis of the National Cancer Institute gene expression database. Biochem. Pharmacol. 62 (5), 547–553.

Nadon, R., Shoemaker, J., 2002. Statistical issues with microarrays: processing and analysis. Trends Genet. 18 (5), 265–271.

Nguyen, D.V., Rocke, D.M., 2002. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 18 (1), 39–50.

Pan, W., 2002. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 18 (4), 546–554.

Pan, W., 2003. On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. Bioinformatics 19 (11), 1333–1340.

Peterson, L.E., 2003. Partitioning large-sample microarray-based gene expression profiles using principal components analysis. Comput. Methods Programs Biomed. 70 (2), 107–119.

Reiner, A., Yekutieli, D., Benjamini, Y., 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 19 (3), 368–375.

Sabatti, C., Rohlin, L., Oh, M.K., Liao, J.C., 2002. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. Nucleic Acids Res. 30 (13), 2886–2893.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270 (5235), 467–470.

Sekowska, A., Robin, S., Daudin, J.J., Henaut, A., Danchin, A., 2001. Extracting biological information from DNA arrays: an unexpected link between arginine and methionine metabolism in *Bacillus subtilis*. Genome Biol. 2 (6), 0019.0011–0019.0012.

Southern, E.M., Maskos, U., Elder, J.K., 1992. Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. Genomics 13 (4), 1008–1017.

Stanewsky, R., 2003. Genetic analysis of the circadian system in *Drosophila* melanogaster and mammals. J. Neurobiol. 54 (1), 111–147.

Templin, M.F., Stoll, D., Schrenk, M., Traub, P.C., Vohringer, C.F., Joos, T.O., 2002. Protein microarray technology. Trends Biotechnol. 20 (4), 160–166.

Troyanskaya, O.G., Garber, M.E., Brown, P.O., Botstein, D., Altman, R.B., 2002. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 18 (11), 1454–1461.

Vrana, K.E., Freeman, W.M., Aschner, M., 2003. Use of microarray technologies in toxicology research. Neurotoxicology 24 (3), 321–332.

Wei, Y., Lee, J.M., Richmond, C., Blattner, F.R., Rafalski, J.A., LaRossa, R.A., 2001. High-density microarray-mediated gene expression profiling of Escherichia coli. J. Bacteriol. 183 (2), 545–556.

Wold, H., 1973. Nonlinear iterative partial least squares (NIPALS) modelling—some current development. In: Krishnajah, P.R. (Ed.), Multivariate Analysis, Academic Press, New York, pp. 383–407.

Zimmer, D.P., Soupene, E., Lee, H.L., Wendisch, V.F., Khodursky, A.B., Peter, B.J., Bender, R.A., Kustu, S., 2000. Nitrogen regulatory protein C-controlled genes of Escherichia coli: scavenging as a defense against nitrogen limitation. Proc. Natl. Acad. Sci. U.S.A. 97 (26), 14674–14679.