The operons, a criterion to compare the reliability of transcriptome analysis tools

A.-S. Carpentier, A. Riva, G. Didier, J.-L. Risler, and A. Hénaut

Laboratoire Génome et informatique UMR 8116
Tour Evry2, 523 Place des Terrasses
91034 EVRY, France
(e-mail: carpentier@genopole.cnrs.fr)

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. We have compared five statistical tools using *Bacillus subtilis* expression data: ANOVA, PCA, ICA, the *t*-test and the paired *t*-test. Our results show ICA to be the most sensitive and accurate of the tools tested.

Keywords: operon, criterion of comparison, transcriptome, expression analysis.

1 Introduction on microarrays and their analysis

Protein activities are the bases of cell and organism functioning. In order to fit to changes in extern or interne physiological conditions the expression level of some genes and the quantity of the corresponding proteins may vary. As proteins are much harder to analyze than mRNAs, techniques for transcriptome analysis have been more popular up to now. In the last decades a tool has been developed in order to measure the expression levels of many genes (several thousands of genes) at the same time.

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]. This allow the identification of gene functions, regulation and networks.

Different tools have been developed for or adapted to the analysis of the huge amount of data created in microarray experiments. The number of tools is continuously increasing and no one, definitive method has so far emerged. 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 ([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].

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. For an operon a single mRNA corresponds to several genes whereas for isolated genes one mRNA corresponds to one gene. It has been shown that the genes within an operon exhibit the same expression patterns [Sabatti *et al.*, 2002].

That is why, 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. 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.

We have tested this criterion on five 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), the *t*-test and the paired *t*-test. Note: ANOVA and the *t*-tests 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.

2 Methods

The microarray data used in this study stem from experiments on the sulphur metabolism of *Bacillus subtilis* [Sekowska *et al.*, 2001]. The experiments were carried out using *B. subtilis* gene arrays; 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 with 4 factors (sulphur source, day of experiment, amount of RNA used and duplicate of each spot).

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. We have normalized the data (mean equal to 0 and variance equal to 1 for each experimental condition).

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. For PCA 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 model used for ANOVA is the following:

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

where Y_{ijkl} is the gene intensity

μ is 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

ϵ_{ijkl} is the residual error.

We need to know how the genes of *Bacillus 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 *Bacillus subtilis* (Subtilist).

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).
   In order to compare the five 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 and the *t*-tests, the p-value obtained for each gene;
   * for PCA and ICA, the remoteness from the cloud centre 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 order of the genes on the lists obtained may differ from each other.
2. "Detected Operons" are identified based on the ranks (one gene of the operon with rank ≤ 20 and another gene with rank ≤ 100).
   It should be noted that a priori the "Detected Operons" may be different for the various tools tested.
3. The Most Significant Interval (MSI) is determined.

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:

$$MSI_j = median_j - first_j$$

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
$first_j$ is the smallest rank value within "Detected Operon" j

4. False positives are evaluated (MSI≥700).
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. The results of this analysis lead us to conclude that a "Detected Operon" is a false positive when MSI≥700 (see table 1 for details).

Operon name	Operon size	MSI (most significant interval)				
		ANOVA	*t*-test	Paired *t*-test	PCA	ICA
fliLMY cheY fliZPQR flhBAF ylxH cheBAWCD sigD ylxL	19	2385	2242	1613	1193	2499
yonRSTUVX yopAB	8	61	134	124	127	251
hemAXCDBL	6	1360	1547			
ruvAB queA tgt yrbF	5				1005	707

Table 1. Quantification of false positives

[We find ourselves with false positives. 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.]

5. "Relevant Detected Operons" are identified (MSI<700). The definition of "Relevant Detected Operons" follows from the definition of false positives: "Relevant Detected Operons" have an MSI<700.
6. The accuracy of a "Relevant Detected Operon" is evaluated (MSI<150). We define that an operon is detected with good accuracy if its MSI is lower then a given threshold. Our results lead us to state that: Operons detected with good accuracy have an MSI<150.
7. The sensitivity of a tool is evaluated.

The sensitivity of the tools is estimated by comparing the number or "Relevant Detected Operons" identified by each tool.

We have decided to compare the five 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. The five statistical tools were tested on these experimental data. The results obtained are displayed in table 2.

Operon name	Operon size	MSI (most significant interval)				
		ANOVA	*t*-test	Paired *t*-test	PCA	ICA
yqiXYZ	3	1	1	4	3	6
argCJBD carAB argF	7	15	28	29	201	56
argGH ytzD	3	1	1	6	6	2
ahpCF	2	46	7	85	11	13
lctEP	2	26			36	8
levDEFG sacC	5	316	220	287		
sunAT yolIJK	5		634			13
ydcPQRST yddABCDEFGHIJ	15				1313	116
ytmIJKLM hisP ytmO ytnIJ ribR hipO ytnM	12			45	92	
flgM yvyG flgKL yviEF csrA hag	8					509
fliLMY cheY fliZPQR flhBAF ylxH cheBAWCD sigD ylxL	19					350
yxbBA yxnB asnH yxaM	5				15	
yvrPONM	4			494		
ycbCD	2			40		
comGABCDEFG yqzE	8			49		
Relevant detected operons		6	6	9	7	9

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

* 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 two 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

Operon name	Operon	MSI (most significant interval)				
	size	ANOVA	*t*-test	Paired	PCA	ICA
				t-test		
comGABCDEFG yqzE	8	16	26	28	6	4
comFABC yvyF	4	339			66	19
cotVWXYZ	5		148		315	417
groESL	2			37		
yvaVWXY	4				53	
yqxM sipW cotN	3				79	
comEABC	3					35
Relevant detected operons		2	2	2	5	4

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

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 the *t*-tests 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.

Operon name	Operon	MSI	
	size	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

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

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 five statistical tools (ANOVA, *t*-test, paired *t*-test, ICA and PCA) under three typical experimental conditions. All five 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 ($MSI\geq700$) 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.

	ANOVA	*t*-test	Paired *t*-test	PCA	ICA
Relevant detected operons					
Table 2-4	8	8	11	15	19
Table 2-3	8	8	11	12	13
Accuracy of Detection					
Table 2-4	75%	75%	82%	80%	84%
Table 2-3	75%	75%	82%	83%	77%

Table 5. Overview of the results

[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, 3 and 4, the total of "Relevant Detected Operons" has been calculated for each tool. The entries for "Tables 2 - 3" have been obtained accordingly. 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, 3 and 4 ("Tables 2 - 4") etc.]

Table 5 sums up the results obtained. Overall, we observe that ANOVA and *t*-test have 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 and the *t*-test were conceived (tables 2 and 3), it performs less well than ICA. The paired *t*-test has a high accuracy but a lower sensitivity than ICA just like PCA. However, each tool may detect operons not identified by the other tools.

The results obtained by testing the five statistical tools show us that ICA has overall the best performance.

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.

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.

References

[Callow *et al.*, 2000]M. J. Callow, S. Dudoit, E. L. Gong, T. P. Speed, and E. M. Rubin. Microarray expression profiling identifies genes with altered expression in hdl-deficient mice. *Genome Res*, pages 2022–9., 2000.

[He *et al.*, 2003]Y. D. He, H. Dai, E. E. Schadt, G. Cavet, S. W. Edwards, S. B. Stepaniants, S. Duenwald, R. Kleinhanz, A. R. Jones, D. D. Shoemaker, and R. B Stoughton. Microarray standard data set and figures of merit for comparing data processing methods and experiment designs. *Bioinformatics*, pages 956–65., 2003.

[Heyer *et al.*, 1999]L. J. Heyer, S. Kruglyak, and S. Yooseph. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res*, pages 1106–15., 1999.

[Pan, 2002]W. Pan. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatic*, 2002.

[Reiner *et al.*, 2003]A. Reiner, D. Yekutieli, and Y. Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, pages 368–75., 2003.

[Sabatti *et al.*, 2002]C. Sabatti, L. Rohlin, M. K. Oh, and J. C. Liao. Co-expression pattern from dna microarray experiments as a tool for operon prediction. *Nucleic Acids Res*, pages 2886–93., 2002.

[Sekowska *et al.*, 2001]A. Sekowska, S. Robin, J. J. Daudin, A. Henaut, and A. Danchin. Extracting biological information from dna arrays: an unexpected link between arginine and methionine metabolism in bacillus subtilis. *Genome Biol*, pages 0019.1–0019.12, 2001.

[Troyanskaya *et al.*, 2002]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*, pages 1454–61., 2002.