Minireview

Expression profiling in reference bacteria: dreams and reality

Antoine Danchin and Agnieszka Sekowska

Address: Pasteur Research Centre, Hong Kong University, 8 Sassoon Road, Pokfulam, Hong Kong.

Correspondence: Antoine Danchin. E-mail: adanchin@hkucc.hku.hk

Published: 10 October 2000

Genome ***Biology*** 2000, **1(4)**:reviews1024.1–1024.5

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2000/1/4/reviews/1024

Abstract

Profiling of gene expression in bacteria is now being used to uncover unknown genes expressed in particular genetic backgrounds or environmental conditions. Obtaining the best possible information from the expected avalanche of such experiments will require standardization of both experimental approach and statistical analysis. The first such experiments reveal challenges, pitfalls and reasonable solutions.

Biology has a long history of describing and classifying objects, mostly in structural terms using the techniques and language of systematics. Even genetics, which identifies gene linkage, has often studied genes as individual entities. In these earlier approaches, a cell was considered but a bag of genes and gene products: it was not usual to find biologists asking questions about the collective behavior of these genes and proteins. Because selection pressure may act on any type of organization, the study of whole-genome sequences now enables us to consider whether genomes are simply collections of genes, or whether there is indeed something more to be discovered in terms of the structure and dynamics of cells and organisms at the global level.

Functional genomics has emerged as a new discipline that uses innovative technologies for genome-wide analysis supported by information technology. It depends both on experiment and on mathematical and computational methods. High-throughput experimental technologies generate large amounts of data on gene expression, protein structure and protein interactions, for example, and powerful information systems are required to analyze these data efficiently. Transcription expression profiling can be used to investigate either the transcriptome (the totality of genes transcribed) or the proteome (the totality of the proteins produced) of a bacterium. DNA arrays and two-dimensional gel electrophoresis are expected to provide a global, high-throughput approach to revealing which genes are expressed at a detectable level, where they are expressed, and which are over- or under-expressed at a given growth stage or following changes in environmental conditions. The use of different growth conditions, different RNA extraction procedures and different array systems has created problems in comparing results, and highlights the need for benchmarking between different laboratories. Here, we review some recent articles describing expression profiling experiments in bacteria, and try to sort out the potential of this approach from the many pitfalls.

Expression profiling in *Escherichia coli* as the bacterial model

Early in the genome sequencing projects, many scientists advocated the study of the whole set of transcripts in a bacterium. It is therefore useful to take as our beginning point the first paper on this topic by Blattner's group, which came out immediately after the complete sequence of *Escherichia coli* was published. In this paper [1], the transcriptome of *E. coli* is analyzed under different growth conditions, and general conclusions are drawn about global gene expression in this organism. Most, if not all, of the conclusions were as expected. It is, however, important to analyze the technical background of the experiment in order to explore whether more information could have been obtained.

The first observation (which will be valid for all the papers analyzed in the present review) is that protein and RNA preparation is a crucial step, unfortunately not properly described in most of the papers. Indeed, membrane proteins are difficult to extract and mRNA turnover is very fast in bacteria (the half-life is often less than a minute at 37°C). A second difficulty is that mRNA is heavily contaminated with a very large amount of rRNA (usually 95%); this should be taken into account, as it may interfere with cDNA construction and subsequent hybridization. A third problem is the nature of the hybridization process on DNA arrays. It can be performed either directly with RNA (which is in principle the best choice, because there is no need for any intermediate, but is extremely difficult to implement for sensitivity or safety reasons), or by generating a cDNA intermediate. In the latter case, the difficulty is that bacterial mRNA is not polyadenylated, so there is no universal primer that can be used for cDNA construction. (Further difficulties will be discussed separately in the analysis of other work.)

The system used by Tao *et al.* [1] consisted of a DNA array on nylon membranes, containing all the presumed *E. coli* coding sequences in duplicate. (*E. coli* coding sequences are properly known as CDSs, although often wrongly and misleadingly named open reading frames, ORFs.) This array was constructed and proposed as a basis for transcriptome analysis by the GenoSys company (see Figure 1 for a typical example of such a membrane for *Bacillus subtilis*). The primers were designed from the 3′ end of each putative CDS and mixed together in the reverse transcription mixture. Every scientist who has constructed a DNA library knows that, if one wishes to have the most even coverage of the sequences of interest, amplification must be avoided at all cost. In the same way, any amplification step in cDNA preparation will lead to uncontrolled differential amplification of some mRNAs over others and will introduce systematic errors. The reverse transcription system used must therefore be free of further amplification steps. Unfortunately, in this earliest work, it is likely that the enzyme used amplified the cDNAs in an uncontrolled manner. As a result, as stated by the authors [1], "some individual expression ratios may be in error, due to technical problems, including cross-hybridization, PCR failures, misapplied DNA spots on the arrays, or scatter in the data (see [2] for a review of the technical aspects of using *E. coli* DNA arrays). A few of the ratios are in conflict with published results, and it is possible that other ratios will not be validated in subsequent experiments. Thus, these data should not be taken as specific evidence for gene regulation and should be independently verified."

The merit of this first article was that the general trends of the data were clear and were valuable for generating experimental leads. It is, however, important to sound a further note of caution in that the rich medium used (Luria Broth) is not a defined medium, and varies from batch to batch. If one does not take into account these difficulties, ^{33}P hybridization of

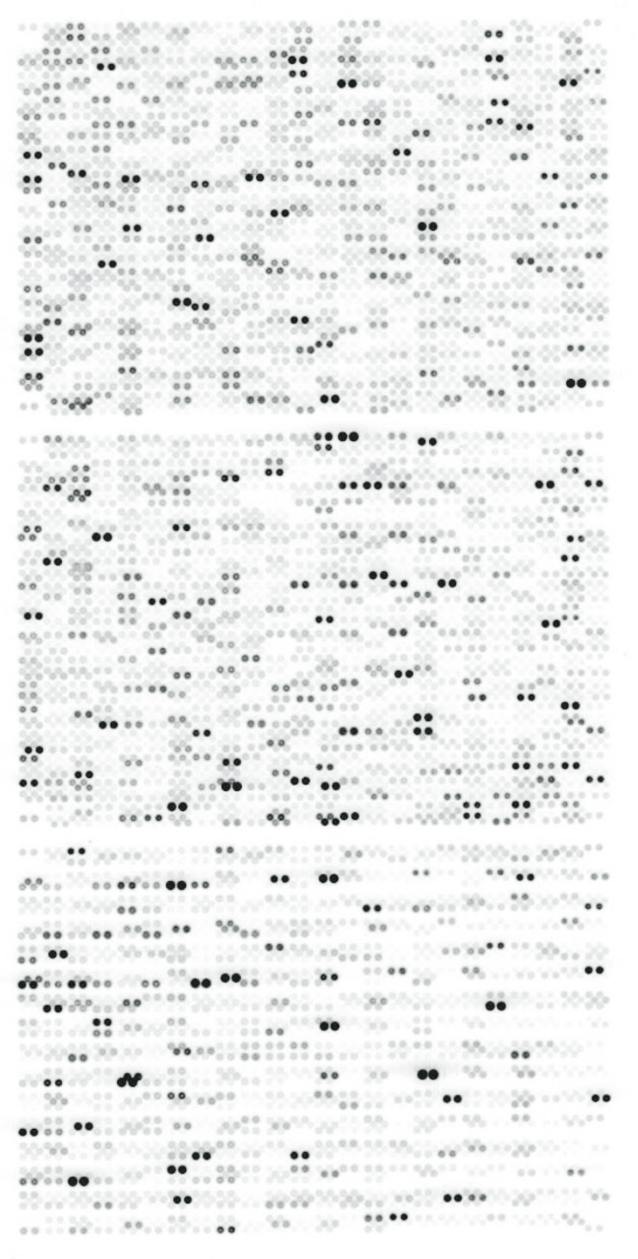

Figure 1
A typical transcriptome analysis of *Bacillus subtilis* grown on methionine as sulfur source. The black dots are genes most strongly expressed, gray less so, and so on.

DNA arrays on nylon filters and fluorescent hybridization of microarrays on glass gave similar results [2].

More recently, an excellent paper by Hatfield and co-workers [3], which was intended to clarify the technical issues posed by expression profiling experiments, has pointed to problems and solutions associated with most, if not all, studies of the bacterial transcriptome. Three different questions were tackled using *E. coli* as a model. First,

emphasis was rightly put on the inescapable consequences of working with a large set of data points. Because a few thousand gene hybridizations have to be observed, a Laplace-Gauss ('normal') distribution would imply the existence of a sizable fraction of apparently significant deviations in hybridization levels from one experiment to another (false positives; typically 250 values in 5,000 would apparently deviate from each other with significance level $p < 0.05$). Incidentally, this implicitly emphasizes the need to use the convenient normal distribution statistics, making it necessary to use the logarithmic value of the hybridization levels (A. Hénaut, A. Sekowska, J.J. Daudin, S. Robin and A. Danchin, unpublished observations), not the value directly measured in the experiment (Figure 2). Fortunately, this is what most automatic statistics software proposes - usually, however, without explicit conceptual justification (or explicit verification).

Thus, it is necessary to interpret data from DNA arrays using statistical methods that can distinguish chance occurrences from biologically meaningful data. This naturally requires repetition of experiments (as should always be done). The second question tackled by Arfin *et al.* [3] was that of the nature of the hybridization preparation needed for the experiment. They showed convincingly, in contrast to what they initially expected (and what is generally expected by scientists performing expression profiling experiments), that random hexanucleotide primers were much more reliable than primers complementary to the expected 3′ ends of the mRNAs for the construction of the labeled cDNA hybridization library. Their third important contribution was to show convincingly that the popular tendency to equate the magnitude of the fold-difference between the expression levels of a gene obtained under two experimental conditions with the accuracy of those measurements was very misleading.

With all these points dealt with, the authors could convincingly show that some new genes were important in the processes controlled by the integration host factor (IHF) in *E. coli*.

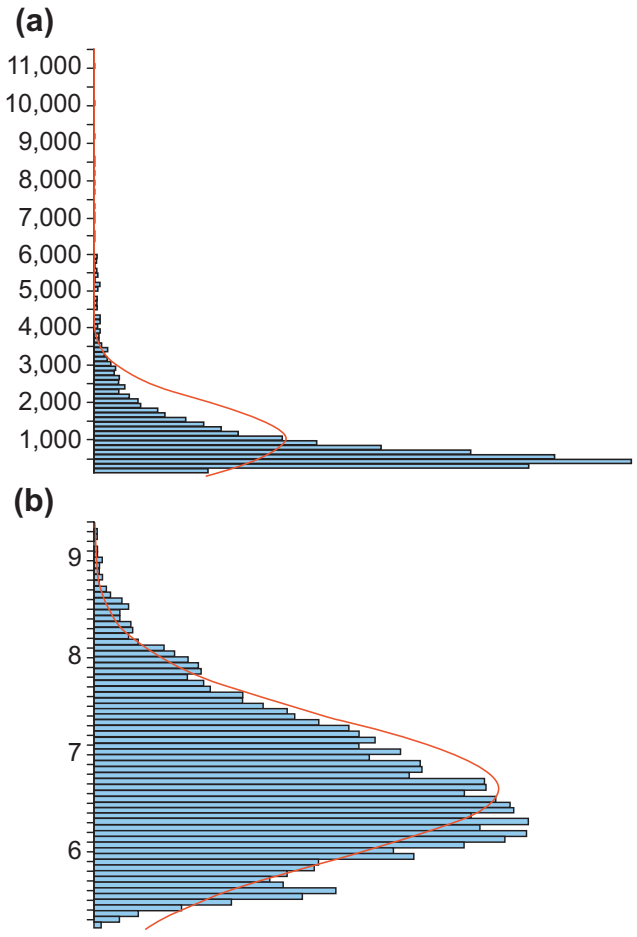

Figure 2
Distribution of hybridization values from the data points taken from Figure 1. **(a)** Histogram of the direct hybridization values. The curve is the best approach to a normal distribution that fits the histogram (note the poor fit). **(b)** Histogram of the logarithm of the hybridization values. Note that the normal distribution exhibits a much better fit (the excess of points at the low end of the curve corresponds to the excess of blank points in Figure 1, which show the background level of the hybridization signal). Note that this behavior should be verified before any interpretation of transcriptome data.

Genome-scale responses in *Bacillus subtilis* and other A+T-rich Gram-positive bacteria

In a series of experiments paralleling those in *E. coli*, Fawcett *et al.* [4] provided a phenomenological analysis of the effect of spo0A and sigma(F) in *B. subtilis*. Their work used the commercial arrays constructed by the company Eurogentech, an RNA preparation prepared using cell centrifugation, and hexanucleotide primers (with a reverse transcriptase presumably resulting in uncontrolled amplification of some templates). There was no investigation of the statistical validity of the approach: indeed, as is popular practice (but not statistically validated), expression-level ratios were calculated with averaged values obtained from replicate filters, and the authors considered only those CDSs showing at least a threefold difference. Superimposed on this basic criterion, a statistical heuristics was used to eliminate CDSs that gave inconsistent hybridization by calculating a 90% confidence interval on the replicate data and asking that ratios calculated at the extremes of the confidence intervals remained greater than threefold. No details are provided about the internal consistency of such heuristics. In view of the thorough study of Arfin *et al.* [3], this type of approach should be reconsidered. Nevertheless, the conservative threefold ratio difference is likely to yield a small number of false positives and the results obtained can therefore be considered as conclusive. Several operons of unknown function

(such as *ybcOPQSTybdAB*) were discovered to be controlled by SpoOA but not sigma(F), others being controlled by both factors. The most interesting outcome of this work was that new genes involved in the process of sporulation were discovered, such as the *yabP* and *yabQ* genes, whose disruption leads to a severe sporulation defect, as observed in the course of the European-Japanese functional analysis program [5].

In a parallel work, Antelmann *et al*. [6] analyzed the phenomenon known as the phosphate-starvation (or stress) response of *B. subtilis*, using mostly proteome analysis under phosphate-limitation conditions (which are similar to the normal conditions of *B. subtilis* growth, between 0.1 and 0.5 mM available phosphate). Phosphate limitation induces both the phosphate regulon (PhoRP) and the sigma(B)-dependent regulon, which is used to allow the cells to survive in frequently encountered environmental conditions where they cannot multiply and have no time to initiate the sporulation process because of transient excess osmolarity, oxygen, heat or pH extremes. Indeed, as revealed by the study of the genome sequence, these bacteria are most probably associated with leaf surfaces (the phylloplane) as their usual environment [7], a biotope that is subject to frequent transient changes. In the work of Antelmann *et al*. [6], the phosphate-limitation conditions allowed monitoring of the time course of protein expression as the cells run out of phosphate. The cytoplasmic and extracellular protein fractions were studied separately; as is usual in current proteome studies, the membrane fraction (25-30% of the protein types) could not be analyzed in this way. Proteins stained with Coomassie blue were identified by mass spectrometry after tryptic digestion in the gel. Known products of the phosphate regulon were used as reference spots on the two dimensional gels. PhoPR- and sigma(B)-dependent proteins were thus identified, including some of unknown function which had not been previously recognized; several new members of the Pho regulon were identified in the supernatant. It should, however, be noted that the medium used contained Tris as the buffer, which can be considered as a mild detergent and thus calls for some caution in interpreting the data. Interestingly, in addition to PhoPR- and sigma(B)-dependent operons (some of which were confirmed individually by northern blot experiments), the authors discovered a few genes (for example, *yjbC*, *yfhM* and *yxiE*) which were expressed independently of these regulators. The authors state that they have unpublished transcriptome experiments substantiating all these results, which are more consistent with the expected operon structure of the gene loci identified than those revealed in the proteome analysis. This is presumably because many genes in these operons code for membrane proteins, which are not present in the two-dimensional gel.

Bacillus subtilis was considered for a long time to be an aerobe barely able to grow in the absence of oxygen, but is now known to be able to grow under anaerobic conditions if appropriate electron acceptors such as nitrate are provided. To study the genes expressed under these conditions, Ye *et al*. [8], from DuPont, constructed DNA microarrays on glass. Out of the predicted 4,100 *B. subtilis* CDSs, 4,020 were present in the array (not all PCRs were successful), and each spot in the array included 2 kb or less of PCR amplification product. The cDNA was produced by random hexamer priming using 10-15 μg RNA purified with a commercial kit using columns for final purification. The kind of reverse transcriptase used in this work is likely to have produced signal amplification, as the authors state that they must degrade RNA after cDNA production. The arrays were hybridized with two sets of cDNAs labeled with fluorescent Cy3 and Cy5 dyes, each corresponding to a different genetic background or growth condition. RNA was normalized by averaging the total amount of fluorescent RNA in each preparation. Each experiment was performed on two glass slides, where cDNAs obtained from the two different conditions were mixed together; the first condition cDNA was labeled with Cy3 and the second with Cy5 on one slide, and vice versa on the second slide. Although Ye *et al*. [8] do not give any detail about the statistical analysis of their data, the experiment gives self-consistent results, showing that a large fraction of the genes they identified are indeed involved in anaerobic metabolism by *B. subtilis*. Most of these genes were known to be involved in the process (in particular, the genes for nitrate and nitrite dissimilation), but several new genes have been identified as putative candidates for anaerobic gene expression: genes controlling electron transfer, iron transport and antibiotic production were found to be involved in the process. Many genes of unknown function were also seen to vary considerably in expression, revealing an intricate network of control when *B. subtilis* grows in the absence of oxygen. At this stage, however, as recognized by the authors themselves, it is important to individually characterize these genes by genetic and biochemical studies. Furthermore, the rich medium used (2YT) is likely to lack reproducibility; similar experiments with better defined conditions will have to be undertaken.

In a similar way, Rimini *et al*. [9] published a global analysis of transcription kinetics during competence development in *Streptococcus pneumoniae* using high-density nylon DNA arrays constructed at Glaxo Wellcome. This study used clones obtained in the *Streptococcus* sequencing project as templates for PCR amplification. They covered most of the genome (1.6-fold coverage on average, with 3,986 overlapping clones, gridded in duplicate). The clones contained inserts of about 800 bp and, to fill the gap, 315 sequenced clones with inserts of about 1,200 bp were used. In this transcriptome experiment, most probes did not contain entire genes, and many contained overlapping genes. The authors claim that this feature ensured the detection of signals from very small CDSs which could be missed by algorithms used for CDS identification. However, clones containing the rDNA

regions were excluded. This meant that some genes in the immediate vicinity were also omitted. RNA was extracted from cultured bacteria grown in diverse conditions and cDNA was produced by random hexamer priming using 25 μg of RNA (with probable signal amplification). As in many other experiments, only twofold differences in signal intensity were retained for further study. As expected, most of the known genes involved in the competence process were expressed upon addition of the competence simulation peptide, and a few other unknown genes were identified. Northern and mutational analyses were used to substantiate some of the results obtained. As in the other cases, no statistical analysis was used to exclude likely false positives. The lack of coverage of the whole genome may also have led to the oversight of some genes involved in competence.

Common challenges

While the studies analyzed in the present article are all interesting, they are mostly descriptive. Nevertheless they show - at least at this stage where scientists are still very careful about the interpretation of their data - that, as a heuristic approach, expression profiling, whether to analyze the proteome or the transcriptome, is an excellent way of making educated guesses about the role of genes of unknown function. The situation will change when more and more data are obtained and the processes will have to be automated. Indeed, the work of Hatfield and co-workers [3] shows that there is a want of an appropriate statistical background. If this could be developed, studies would be much more convincing, and it could show up differences which cannot be seen as statistically meaningful with the present approaches. In particular, it is most important when new or unexpected genes are identified in a process, to exclude the possibility that this is a false-positive result. Also, many important genes, such as regulatory genes, have a general leverage effect on other gene expression; even if they vary little in their expression, this can result in a large change in that of the genes they control. Therefore, using the fold-difference between the expression of a gene in different conditions is misleading, and yields many false-negative results. A second important conclusion is that there is still a need for research into optimal experimental conditions.

As a consequence, there is an urgent need for appropriate benchmarking and standardization, to enable valid exchange of experimental data and comparison of results, as well as to reduce the current considerable variation in the presentation of information from expression profiling, whether of the proteome or the transcriptome. Researchers must be able to combine their efforts by relating experimental results produced in one laboratory to those produced in others. This is especially important when new data libraries are flooded by large tables of expression data, with no means of correlating them with each other. In addition, the general use of expression data by other biologists is only feasible if the experimental procedures, the source materials used and the analysis algorithms are comparable and well defined. Benchmarks for methods in functional genomics are therefore as important as the standardized reference materials on which expression profiling is performed. As can be seen from the papers reviewed here, there is a serious problem of heterogeneity of experimental approaches and in the data generated, and a detrimental lack of traceability of information. This concerns the nature of the growth media (rich media are not reproducible), RNA extraction procedures, and, above all, the methods used for cDNA preparation. It is important, therefore, that collaborative networks be constituted to promote standardization not only of nomenclature, but of growth media, growth conditions and biochemical procedures for two-dimensional gel electrophoresis, DNA array construction and hybridization protocols. Last but not least, standard protocols in statistical analysis must be constructed to take into account the not so far distant time when analysis will have to be almost entirely automated. This will, as a complement to genome sequencing programs, make expression profiling a hugely powerful approach to the study of bacterial physiology.

Acknowledgements

The HKU Pasteur Research Centre is a non profit-making institution. This work was supported by The University of Hong Kong and by private donations from James Kung.

References

1. Tao H, Bausch C, Richmond C, Blattner FR, Conway T: **Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media.** *J Bacteriol* 1999, **181:**6425-6440.
2. Richmond CS, Glasner JD, Mau R, Jin H, Blattner FR: **Genome-wide expression profiling in *Escherichia coli* K-12.** *Nucleic Acids Res* 1999, **27:**3821-3835.
3. Arfin SM, Long AD, Ito ET, Tolleri L, Riehle MM, Paegle ES, Hatfield GW: **Global gene expression profiling in *Escherichia coli* K12: the effects of Integration Host Factor.** *J Biol Chem* 2000, **275:**29672-29684.
4. Fawcett P, Eichenberger P, Losick R, Youngman P: **The transcriptional profile of early to middle sporulation in *Bacillus subtilis*.** *Proc Natl Acad Sci USA* 2000, **97:**8063-8068.
5. **European-Japanese Functional Analysis Program.** [http://bacillus.genome.ad.jp]
6. Antelmann H, Scharf C, Hecker M: **Phosphate starvation-inducible proteins of *Bacillus subtilis*: proteomics and transcriptional analysis.** *J Bacteriol* 2000, **182:**4478-4490.
7. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A, Borchert S, *et al.*: **The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*.** *Nature* 1997, **390:**249-256.
8. Ye RW, Tao W, Bedzyk L, Young T, Chen M, Li L: **Global gene expression profiles of *Bacillus subtilis* grown under anaerobic conditions.** *J Bacteriol* 2000, **182:**4458-4465.
9. Rimini R, Jansson B, Feger G, Roberts TC, de Francesco M, Gozzi A, Faggioni F, Domenici E, Wallace DM, Frandsen N, *et al.*: **Global analysis of transcription kinetics during competence development in *Streptococcus pneumoniae* using high density DNA arrays.** *Mol Microbiol* 2000, **36:**1279-1292.