J Mol Med (1999) 77: 3–7

Bioinformatics: Bits and Bytes

Computational aspects of expression data

Martin Vingron · Jörg Hoheisel

Abstract Several experimental techniques are available nowadays to study the spectrum of genes expressed in a cell at a specific moment. Typically, such methods generate large amounts of expression data that may be hard to interpret. Here we review computational questions and approaches resulting from the various experimental techniques.

Abbreviations *EST* Expressed sequence tags · *SAGE* Sequential analysis of gene expression

Introduction

In order to understand the workings of a living cell, knowledge of the spectrum of genes expressed at a given time, or under certain conditions, as under the influence of specific drugs, should prove instrumental. Such information can aid in the understanding of gene regulation, for example, or in understanding the changes that occur in disease. Further insights from the identification of new genes associated with certain diseases may also be gained. In fact, the study of differences in gene expression has become a standard method in the search for genes that might be of medical or pharmaceutical relevance. Such studies may focus, for example, on the differences in gene expression between normal and cancer cells [1].

Current studies of global mRNA expression levels (transcriptome analysis) as well as the study of the proteome, i.e. the total protein contents of a cell, are undergoing rapid development. While the amount of mRNA in a cell at a given moment does not guarantee the precise prediction of amounts of subsequently produced protein, mRNA can nevertheless serve as an indicator that a certain protein is being produced, especially when sudden changes, as opposed to steady-state conditions, are analyzed. Although studies have shed doubt on a quantitative correlation between mRNA and protein levels [2], they nevertheless underscore the importance of mRNA processing in regulating protein levels. We will choose to ignore such matters for the moment and focus on methods to compare and determine transcript levels so as to consider the computational problems posed by such data.

Typically, analytical techniques for gene expression result in large amounts of data that are hard to interpret without computational methods. However, the kinds of large scale experiments described below require very careful analysis because biological signals may be obscured by experimental noise and other systematic influences arising from experimental methodologies.

Experimental techniques

Several techniques are currently in use to study levels of expressed mRNA. In terms of the data they produce we believe it is important to distinguish from among methods that specifically *compare* two populations of mRNA, methods that *precisely count* how many molecules of a particular species are present in a sample, and methods that *read* the mRNA contents from the strength of a hybridization signal.

Subtractive methods compare two mRNA populations.This class of method includes techniques like representational difference analysis [3] and differential display [4]. The results are rarely quantitative but may allow for the identification of interesting new genes. Tag sequencing of identified mRNA representatives will frequently reveal the same sequence and therefore additional hybridization may be necessary in order to differentiate from among various clones. The microchip hybridization method developed by Pat Brown [5] labels two populations with different fluorescent dyes and then hybridizes them to arrayed clones. Thus, in a sense, one may also categorize it under the heading of subtractive meth-

Martin Vingron (✉) · Jörg Hoheisel
Deutsches Krebsforschungszentrum,
INF 280, D-69120 Heidelberg, Germany
Fax +49 6221 42–4682,
e-mail: m.vingron@dkfz-heidelberg.de

Please send articles to:
Peer Bork
Max-Delbrück-Center
for Molecular Medicine (MDC)
Robert-Rössle-Strasse 10
D-13122 Berlin, Germany
and:
EMBL
Meyerhofstrasse 1
D-69117 Heidelberg, Germany
E-mail: bork@embl-heidelberg.de
http://www.embl-heidelberg.de/~bork/

ods. In principle, any comparison between two mRNA populations can be derived if the composition of each is given, which is what the other two strategies are aiming at by characterizing a given population.

Tagging. Given a method to determine certain tags on a mRNA, like a stretch of sequence at either end of the molecule, one can then count the numbers of molecules carrying the same tag. For example, expressed sequence tag (EST) sequencing of a library will yield the same sequence repeatedly. The multiplicity at which a certain sequence occurs upon sequencing may serve as an indicator of the number of molecules of this species present in the population. Data quality and reliability strongly depend on the method for selection of tags for sequencing and on the chance that identical mRNAs will produce the same tag sequence. Generally, in order to obtain a reliable estimate of the frequency of a certain mRNA, rather large numbers of clones need to by typed. The same holds for the SAGE (Sequence Analysis of Gene Expression) method [6] where short tags are ligated into one long consecutive sequence prior to sequencing. While for EST sequencing and the SAGE method identity between tags is easily determined, this does not hold true for a third approach based on the use of oligonucleotide fingerprints as tags [7, 8]. Hybridization of a clone with a set of oligonucleotides yields a vector of hybridization signals which may, due to experimental noise, not be identical for identical clones. As opposed to the sequencing based methods, however, oligomer fingerprinting produces information from the entire length of a sequence. In all the above techniques, there is a risk of identical tags being derived from different mRNAs.

Hybridization methods. Two hybridization methods provide semi-quantitative data by correlating hybridization signal strength to the amount of mRNA in a population. In one approach, cDNAs or PCR-products are immobilized either on a filter array or a glass chip and hybridized with the mRNA population [5, 9]. In a rather similar approach, differing mainly in the technique of placing the target sequences on the chip, high density arrays of oligonucleotides are used for hybridization [10]. Both methods generally provide good, reproducible correlation between the amount of mRNA present and the signal strength.

Computational profiling

Subtractive methods directly answer the question as to which transcripts occur in one and not in the other of two given libraries. Ensuing questions concerning the biological interpretation of results will require sequencing of a number of clones in consultations with sequence databases. This effort will be the more rewarding the better the accompanying sequence analysis is. Known sequences and motifs will thereby aid in functional predictions, although it is likely that even after careful analysis some sequences will remain for which comparisons provide no clue as to their role.

Tagging methods implicitly suggest that the number of clones carrying the same tag in relation to the total number of clones is representative of the fraction of that species in the population. However, certain qualifications need to be mad e. The situation is similar to one where a fisherman catches 100 fish out of a lake and finds that one of them is a carp. Does this imply that 1% of fish in the lake are carps? Or conversely, if 1% percent of fish in the pond are carps, does this make it highly likely that the fisherman would find exactly one carp among his 100 fish? This is particularly important since there is the danger that the sample may contain no carp at all which does not justify the conclusion that there are no carps in the pond.

The answer can only be given in the language of probability. Suppose the pond holds a very large number of fish of which 1% percent are carps. Although the probability of having some specific number of carps in the sample should be computed according to a hypergeometric distribution [11 Section II.6], in the case of a sample size much below some large number of fish in the pond one may instead use the simpler binomial distribution [11, Sections II.11 and VI.10]. In this case, the probability of finding exactly one carp among 100 fish is 36.97%. However, the probability to observe no carp at all is 36.6%, too. The chances to find 2 or 3 carps in the sample are 18.5% and 6% respectively. Thus the probability not to observe a certain species (of fish or mRNA) may be disturbingly high. For example, for a species that constitutes 0.5% of the population one needs a sample of size 460 before the probability of not observing this species at all decreases to below 10%. For a species that constitutes 0.1% of a population the corresponding minimal sample size is well above 2000. Figure 1 shows a plot of the probabilities of no observation of a certain species in samples varying in size from 1 to 1000. Each curve corresponds to a specific fraction at which the species occurs in the population. This shows how difficult it is to assess the precise meaning of the counts of tags for one species and how dangerous it is to make deductions about low-copy number transcripts.

When an mRNA population is hybridized to a filter, micro array, or chip the ability to detect low-abundance species depends on the detection method rather than on chance. For this reason, miniaturization and the use of support media other than nylon filter are crucial. Because of a small surface area, the probe concentration can be made high, even from limited amounts of sample material, which immediately translates into improved hybridization and thus better performance. Concomitantly, inert support materials produce low background, again leading to increased sensitivity and accuracy. Irrespective of the actual detection device, whether general purpose imaging devise or special equipment for the microchips, image analysis software may become a bottleneck in extracting this information from the experiment.

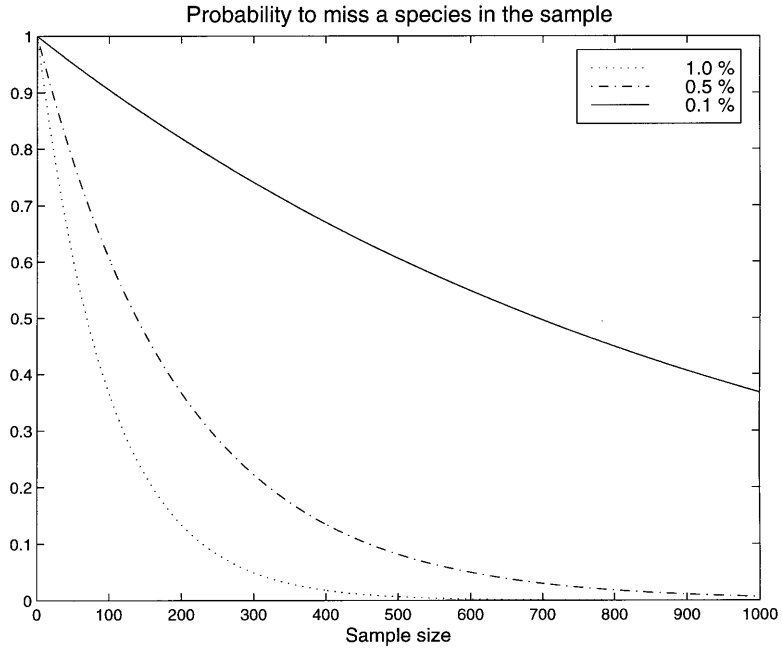

Fig. 1 Probability not to observe a certain species in a sample that is taken out of a large pool of mRNAs. The curves correspond to species constituting 1%, 0.5%, and 0.1% of the entire population. For example, as pointed out in the text, it takes a sample size of 460 for the 0.5%-curve to fall below 10%. If a species constitutes 0.1% of the population in the pool -which is still quite high in terms of mRNA contents – the corresponding curve falls off very slowly, and extremely large numbers need to be sampled. Since the overall number of mRNAs is assumed to be much larger than the sample size, the binomial approximation to the hypergeometric distribution has been used to compute these numbers

Comparing two experiments

Both tag sequencing and hybridization methods will eventually lead to a quantification, however approximate, of the expression level of each mRNA species in one or several populations. The simplest type of question to ask for two populations is, "Which mRNAs are either specific or are increased or decreased in one of the experiments?" This is easily approached by simply plotting each set of values on an axis of a two dimensional coordinate system [9, 12]. The resulting correlation plot gives a good visual impression of the relationship between the two expression profiles. Due to the inaccuracies of the data generation and due to real fluctuations of mRNA levels it will, however, only be possible to detect fairly large changes. Subtle changes will hardly be distinguishable from random noise.

While the correlation plot is easy to interpret visually, it is not obvious how to assess the statistical significance of a change in expression level of a particular gene. For the tagging methods which are able to count molecules from a sample, there is some statistical theory to apply here. Audic and Claverie [13] derive a significance formula for an observed change in the number of mRNAs of a certain species in two libraries. A simpler approach, and according to Audic and Claverie a conservative one, is the Fisher exact test, originally designed for two-way contingency tables [14].

Analysis of several experiments

It is far less obvious how to study several expression profiles simultaneously. Several questions arise with regards to such data: are there groups of experiments where certain genes can be expected to undergo the same changes? Given prior knowledge as to the clustering of experiments into groups, is there an indication that the expression profiles mirror this classification? Are there predictive factors in the expression profiles? Can expression profiles be monitored in a time-dependent manner?

In general, multiple expression profiles may be summarized as a matrix, with rows being experiments (cells, tissues, time points) and the columns corresponding to the genes. An entry in the matrix reflects the expression level of the particular gene in the particular cell. Thus, one row is represented by its transcript profile which is a vector with as many dimensions as there were genes or clones assessed. This number may be fairly high, like some 6,000 for the yeast genes, for example, or the approximately ten times larger number of genes in mammalians.

The fundamental problem here is that one would need to visualize many dimensions simultaneously in order to comprehend all information. Several statistical techniques have been developed for the analysis and visualization of high-dimensional data and the hope is that some of these will prove fruitful in the context of expression data. For this purpose, we proceed to sketch a few approaches. Of course, any technique chosen needs to reflect the underlying biological question. For example, different methods will be used for the analysis of profiles that reflect a development in time and for the study of changes in a diseased organ.

In one generic method of visualizing high dimensional data one represents each dimension as one of many vertical lines in a plane [15], such that each vertical line corresponds to a gene and the expression level of that gene corresponds to a certain height on that vertical line. The visual impression is created by linking all those marks on adjacent lines that come from the same row (experiment) of the data matrix. One might also link from left to right those marks that belong to the same experiment in a specific color and thus distinguish the rows of the data matrix by color. The expectation is that one will be able to pick out

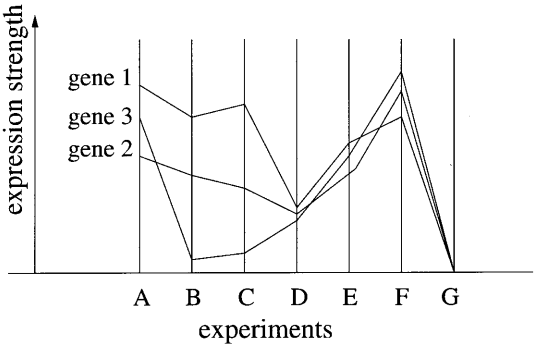

Fig. 2 Example for parallel lines depiction of expression strength based on fictitious data. The expression levels for three genes in seven experiments A to H are shown. The obvious interpretation would be that all three genes behave very similarly in experiments D, E, F, G, and possibly also in A, while they differ in experiments B and C

groups of experiments displaying similar profiles by eye or that outliers might be spotted.

Any visual impression will be dependent on the order of the lines. For an experiment representing a time course the order is naturally given, while without this information different arrangements need to be tested. Certainly, there are clear quantitative limits to this method. While it may be helpful for the study of, say, 20 genes, 20,000 parallel lines with expression levels on them will be hard to interpret.

A classical data analysis tool is multidimensional scaling. In this approach, a low dimensional (generally a 2- or 3-dimensional) projection of the original higher-dimensional data is computed. This projection is chosen such that clustering and spread in the original data are maintained as far as possible. Typically, the image generated by this projection gives a good impression of the rough clustering among the data points. Thus, one would expect to find experiments near to each other where the expression levels of the genes tend to be very similar to each other. The work by Spanakis and Brouty-Boye [16] uses such an approach. Projection methods may also be used to derive classifiers, i.e. decision criteria as to which predetermined group a new transcript profile belongs to. In some instances even trees may constitute a meaningful way of summarizing and interpreting the data. Carr et al. [17, 18] apply trees to represent a clustering of expression profiles. This is in line with classical data analysis where trees are used to represent a hierarchic classification. In the study of evolution, trees not only represent grouping but also summarize development. This aspect is stressed in the work of Gawantka et al. [19] who fit a tree structure to expression data taken from in situ hybridization in Xenopus embryos. The resulting tree describes to a certain extent the differentiation process by which the tissues develop from a small number of embryonic cells.

The representation of several transcription profiles as a data matrix allows one to interchange rows and columns. Indeed, interesting questions may be posed when the data matrix is used to cluster genes instead of experiments. When sequences or functions are available for the genes, one may study the degree to which profile-based clustering agrees with a sequence or function-based clustering. For example, Gawantka et al. [19] found interesting functional correlations among genes that clustered together according to their expression levels. This duality between rows and columns of the data matrix promises to open many interesting routes of theoretical inquiry.

Discussion

Expression profiling analysis is an emerging field that exemplifies the growing interdependancy between experimental techniques and data analysis. The tighter the feedback loop beween bioinformatics and experiment, the more fruitful this approach will be. It is very unlikely that we have mentioned all the conceivable applications and options in analyzing expression profiles, especially since actual applications vary so widely. For instance, algorithms appropriate for observations of steady-state cultures might not be the most useful for the evaluation of data resulting from a sudden change in culture conditions, and vice versa. Another example is the difference, and the consequences thereof, between studies that aim at merely matching hybridization patterns, as for the identification of similarly acting drugs, and really quantitative analyses. In any case, the amount of data will be such that new and yet unthought means need to be developed to deal with them and, even more importantly, to allow appropriate use of the information. The full potential of this type of analysis, especially with regard to the unraveling of regulatory pathways, will only become available with the ability to distinguish relevant from irrelevant data, in a development that looks not too different to the advance from the initial analysis of monogenic diseases to the identification and isolation of all genes involved in multigenic treats.

For all its merits, one should keep in mind that the mere transcriptional information will permit what soon will be perceived as a rather limited analysis, providing pieces to the puzzle. Only when merged with appropriately matched data on promoter activity and actual protein expression, a more comprehensive analysis of the complex and interacting regulative factors will become possible, even on this relatively low molecular level. Other important effectors such as the consequences of post-translational activity or the influence of cell compartmentalization, for example, will still not be included.

References

1. Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR,

Vogelsten B, Kinzler KW (1997) Gene expression profiles in normal and cancer cells. Science 276:1268–127
2. Anderson L, Seilhammer J (1997) A comparison of selected mRNA and protein a bundances in human liver. Electrophoresis 18:533–537
3. Hubank M, Schatz DG (1994) Identifying differences in mRNA expression by representational difference analysis of cDNA . Nucleic Acids Res 22:5640–5648
4. Liang P, Pardee AB (1995) Recent advances in differential display. Curr Opin Immunol 7:274–280
5. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci USA 93:10614–10619
6. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270:484–487
7. Drmanac S, Stavropoulos NA, Labat I, Vonau J, Hauser B, Soares MB, Drmanac R. (1996) Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. Genomics 37:29–40
8. Meier-Ewert S, Lange J, Gerst H, Herwig R, Schmitt A, Freund J, Elge T, Mott R, Hermann B, Lehrach H (1998) Comparative gene expression profiling by oligonucleotide fingerprinting. Nucleic Acids Res 26:2216–2223
9. Hauser NC, Vingron M, Schneideler M, Krems B, Hellmuth K, Entian K-D, Hoheisel JD (1998) Transcriptional profiling on all open reading frames of Saccharomyces crevisiae. Yeast 14:1209–1221
10. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology 14:1675–1680
11. Feller W (1950) An Introduction to Probability Theory and Its Applications. John Wiley and Sons, New York
12. Pietu G, Alibert O, Guichard V, Lamy B, Bois F, Leroy E, Mariage-Samson R, Houlgatte R, Soularue P, and Auffray Ch (1996) Novel gene transcripts prefentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. Genome-Research 6:492–503
13. Audic St and Claverie J-M (1997) The significance of digital gene expression profiles. Genome Research 7:986–995
14. See http:/www.ncbi.nlm.nih.gov/ncigap/fisher.html
15. Inselberg A, and Dimsdale B (1987) Parallel coordinates for visualizing multi-dimensional geometry. Computer Graphics 1987 (Proceedings of CG International), 25–44
16. Spanakis E, Brouty-Boye D (1997) Discrimination of Fibroblast subtypes by multivariate analysis of gene expression. Int J Cancer 71:402–409
17. Carr DB, Somogyi R, Michaels G (1997) Templates for looking at gene expression clustering. Statistical computing and statistical graphics newsletter 8:20–29
18. Wen X, Fuhrmann S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R(1998) Large-scale temporal gene expression mapping of central nervous system development. Proc Natl Acad Sci USA 95:334–339
19. Gawantka V, Pollet N, Delius H, Vingron M, Pfister R, Nitsch R, Blumenstock C, Niehrs C (1998) Large scale gene expression screening in Xenopus identifies molecular pathways, predicts gene function and provides a global view of embryonic gene expression. Mechanisms of Development (in press)