Computational methods for the identification of differential and coordinated gene expression

Jean-Michel Claverie[+]

Structural and Genetic Information Laboratory, CNRS UMR 1889, 31 Chemin Joseph Aiguier,
13402 Marseille cedex 20, France

Received May 10, 1999; Accepted June 1, 1999

With the first complete 'draft' of the human genome sequence expected for Spring 2000, the three basic challenges for today's bioinformatics are more than ever: (i) finding the genes; (ii) locating their coding regions; and (iii) predicting their functions. However, our capacity for interpreting vertebrate genomic and transcript (cDNA) sequences using experimental or computational means very much lags behind our raw sequencing power. If the performances of current programs in identifying internal coding exons are good, the precise 5′→3′ delineation of transcription units (and promoters) still requires additional experiments. Similarly, functional predictions made with reference to previously characterized homologues are leaving >50% of human genes unannotated or classified in uninformative categories ('kinase', 'ATP-binding', etc.). In the context of functional genomics, large-scale gene expression studies using massive cDNA tag sequencing, two-dimensional gel proteome analysis or microarray technologies are the only approaches providing genome-scale experimental information at a pace consistent with the progress of sequencing. Given the difficulty and cost of characterizing genes one by one, academic and industrial researchers are increasingly relying on those methods to prioritize their studies and choose their targets. The study of expression patterns can also provide some insight into the function, reveal regulatory pathways, indicate side effects of drugs or serve as a diagnostic tool. In this article, I review the theoretical and computational approaches used to: (i) identify genes differentially expressed (across cell types, developmental stages, pathological conditions, etc.); (ii) identify genes expressed in a coordinated manner across a set of conditions; and (iii) delineate clusters of genes sharing coherent expression features, eventually defining global biological pathways.

INTRODUCTION

For a long time, the common view has been that the deciphering of genomic sequence information would mostly be accomplished by means of automated computational methods, implementing a set of rules describing the architecture of genes and a finite catalogue of regulatory elements and functional signatures. With the first complete genome of a multicellular organism in hand (1), several others to follow rapidly (2,3) and a 'draft' of the human genome to be completed by Spring 2000 (4), we know that this will not happen. The accomplishments of bioinformatics in the context of higher eukaryote genomes have been humbling and basic analyses such as precisely identifying the intron/exon architecture of genes and the precise boundaries of their transcript(s) are still performed with unacceptable uncertainties (5). The prediction of promoter locations (and properties) is also a notable failure (6–8).

On the other hand, bioinformatics methods become immediately more useful if they can be supplemented with some experimental knowledge (i.e. transcript maps, homologous sequences, etc.). Thus, the present successes of bioinformatics are truly in the realm of 'reverse engineering', i.e. decoding the genetic information using some associated experimental insight.

In the context of functional genomics, computational methods were also expected to significantly contribute to the prediction of gene function. Here again, the results have been poor. If rich catalogues of recurrent protein motifs have been designed (9–14), their association with a precise function is often too vague to identify the precise biological pathway in which they operate. Moreover, close to 50% of all newly identified genes do not exhibit a significant similarity with a previously well-characterized homologue or fall into uninformative categories such as 'protein kinases' or 'transcription factors'.

As genomic sequencing was picking up, methods to monitor the expression of many genes simultaneously were also designed and progressively scaled up to allow genome-wide studies. The technique of 'differential display' (15,16) and the generation of expressed sequence tags (ESTs) (17–23) have first been used for the identification of genes exhibiting marked differential expressions across tissues, development stages or normal versus pathological conditions. The original EST approach was then improved by the use of smaller, concatenated and more numerous cDNA tags (24–26). As an alternative to sequencing, cDNAs can also be identified by oligonucleotide fingerprinting (27). More recently, microarrays capable of providing hybridization-based expression monitoring of tens of thousands of genes in parallel

[+]Tel: +33 4 91 16 45 48; Fax: +33 4 91 16 45 49; Email: jmc@igs.cnrs-mrs.fr

have become available, with two main technologies competing: oligonucleotide-grafted chips (28–31) and cDNA-printed glass slides (32–34). The latter are most often used in conjunction with a two-colour fluorescence competitive assay (35,36). Numerous recent review articles have commented on the enormous potential of microarray-based transcriptome studies (see, for example, refs 37–42). However, tag-counting methods are still very popular and important results have been obtained with the EST (43–53) or SAGE (54–59) protocols.

The analysis of gene expression patterns derived from normal and pathological situations is a valuable tool in the discovery of therapeutics targets and diagnostic markers. The recognition of coordinated expression profiles between characterized or anonymous genes also enables inferences about biological pathways and gene functions to be made.

At the moment, the measurement of gene expression using microarrays or cDNA tag sampling appears to be the sole approach to gene characterization capable of matching the speed of sequencing and the scale requirement of functional genomics. In consequence, expression profiles for many genes and from multiple experimental conditions often constitute the main information (besides the sequences themselves) with which to guide the 'reverse engineering' process of functional genomics. Thus, a general understanding of the various ways such data can be used becomes central. In this article, I review the concepts and methodologies involved in the interpretation of gene expression profiling experiments. Following the pioneering work of Anderson *et al.* (60), several recent articles have discussed the bioinformatics of large-scale expression monitoring, emphasizing computational (61–69) or data management (70–77) aspects.

METHODS

Differential expression studies: pairwise condition comparison

In the simplest experimental situation, gene expression is compared between two conditions such as normal versus pathological or control versus drug-treated. The general form of the expression data table is then:

	Condition A	Condition B
Gene 1	g_{1A}	g_{1B}
Gene 2	g_{2A}	g_{2B}
Gene 3	g_{3A}	g_{3B}
Gene…	$g_{...A}$	$g_{...B}$
Gene *n*	g_{nA}	g_{nB}

In the context of EST (17) or SAGE (24) studies, the expression 'intensities' g_A and g_B are cDNA tag counts and the genes listed in the first column are those for which at least one occurrence was observed from at least one of the libraries (A or B). The sampling sizes N_A and N_B (i.e. the total numbers of sequenced tags) may vary from a few hundred to several tens of thousands depending on the laboratory (or company) sequencing capacity. Accordingly, the number *n* of observed genes also varies from a few hundred to a few thousand.

In RT–PCR- (64) or hybridization-based (30,32) studies, the expression intensities are derived from absorbency or fluorescence 'analog' signals, often normalized to a number of mRNA molecules using a known quantity of exogenous control mRNA. Here again, the total number of simultaneously studied genes ranges from a hundred to several thousands (or tens of thousands).

The lay-out of the above data table recalls a setting for which biologists might think it appropriate to use the traditional χ^2 2×*k* significance test. However, this would be incorrect. The purpose of the χ^2 computation is to test whether conditions A and B significantly differ as a whole, using the entire A and B columns of expression intensities. The question asked through differential expression experiments is different; it is to identify the peculiar genes, the expression of which significantly varies between the two conditions. At the two extremes, ubiquitous genes will exhibit no variation, while 'condition-specific' (e.g. tissue-, developmental stage- or disease-specific) genes will only be detected in A or B. In this section I review and discuss the different statistical methods required to mine the expression intensity tables generated from tag sampling or microarray experiments.

Detecting differential expression in tag sampling experiments

Large tag sampling experiments are usually not replicated. This implies that the standard errors associated with each expression measurement cannot be estimated from its dispersion and that none of the standard tests requiring variance estimates (such as Student's *t*-test) can be used. Fortunately, the result of randomly sampling tags from a large set of genes is very well approximated by the Poisson distribution, which implicitly provides a built-in estimate of the standard error. In this context, Audic and Claverie (61) have derived a new significance test specifically adapted to the analysis of tag sampling data. Their basic result is quite simple: for two sampling experiments A and B, involving the same total number of tags, the probability of observing g_A and g_B tags for a gene equally expressed in both conditions is given by:

$$p(g_B|g_A) = \frac{(g_A + g_B)!}{g_A!g_B!2^{(g_A + g_B + 1)}} \qquad \mathbf{1}$$

Small values for *p* correspond to large differences between g_A and g_B, unlikely to arise by chance if the gene under scrutiny is expressed at the same level in conditions A and B. Provided that all experimental factors are well replicated, statistically significant discrepancies (such as $p << 1\%$) between the values of g_A and g_B can thus be used to point out the gene most likely to be differentially expressed.

It is worth noting that the sampling size (i.e. the total number *N* of generated tags) does not appear in **1** and has no direct influence on the *p* value. The statistical significance of the differences observed in tag counts only depends on the absolute values g_A and g_B. This apparent paradox is discussed in Audic and Claverie (61). The form of **1** also indicates that analyzing expression measurements in terms of g_A/g_B ratios (as is customary in most published works) is not appropriate, as it cannot be related to a confidence estimate. Equation **1** provides a quantitative test of our intuition that a 2-fold increase computed from a $g_A = 10$ versus $g_B = 20$ change in tag counts [$p(20|10) = 0.014$] is more robust and significant than the same ratio computed from a $g_A = 1$ versus $g_B = 2$ transition [$p(2|1) = 0.19$]. In fact, a more rigorous significance test requires the use of the cumulative form of **1**. A table of [g_A, g_B] couples corresponding to the usual 5 and 1% significance thresholds is provided in Audic and Claverie (61). In the same work, Audic and Claverie also extended **1** to the more practical case of A versus B comparisons involving different total numbers of tags N_A and N_B. This significance test was successfully validated using computer simulation on real EST data. As expected, the frequency of genes falsely identified as differentially expressed was found to be less than or equal to the selected significance threshold (i.e. 5% of false identifications when using a significance *p* value of 5%). The general form of the significance test can be used interactively on a web site at http://igs-server.cnrs-mrs.fr or the source code obtained from the authors.

Fisher's 2 × 2 exact test (78) is also being used to analyse tag sampling experimental data, most notably the Cancer Genome Anatomy project (43). This test is traditionally used for the analysis of 2 × 2 contingency tables arising from treatment versus control experiments. To fit this test, the data corresponding to each gene in the original two-condition expression matrix must be, quite artificially, rewritten as:

	Condition A	Condition B
Gene 1	g_{1A}	g_{1B}
All others	$N_A - \mathrm{g}_{1A}$	$N_B - g_{1B}$

where g_{1A} and g_{1B} are the tag counts associated with gene 1 and N_A and N_B are the total numbers of tags generated in experiments A and B, respectively. On theoretical grounds, the validity of using Fisher's 2 × 2 exact test in such a setting is not clear. Rigorously, the test requires the sums of columns and rows to be known prior to the experiment. Also, the definition of the 'all others' aggregated gene category is logically inconsistent, as it implies that the genes expressed and observed in conditions A and B are the same, which might be a largely incorrect assumption. In practice, however, probability values computed according to Audic and Claverie (61) or from Fisher's test are close, with the latter being slightly too conservative (i.e. a larger expression bias is required to reach a given significance threshold) (61). As for **1**, the setting for Fisher's test again emphasizes that the significance of expression changes must be assessed from the tag count values themselves and not from their ratio.

Fisher's exact test is more appropriate when studying the distribution of alternative transcript forms in two different conditions. The data setting then becomes:

	Condition A	Condition B
Short form of gene 1	g_{SA}	g_{SB}
Long form of gene 1	g_{LA}	g_{LB}

now corresponding to a traditional 'association' experiment (i.e. is a gene form preferentially associated with one of the conditions?) for which Fisher's 2 × 2 exact test is well suited. Such a design has been applied to a large-scale analysis of alternative polyadenylation in human mRNAs (63) using the Merck-sponsored EST (22,23) data set.

Detecting differential expression in hybridization-based experiments

Hybridization-based experiments [or quantitative RT–PCR protocols (79,80)] produce 'analog' expression intensities in contrast to the 'digital' nature of the tag counting protocols discussed above. The expression data matrix thus consists of real numbers such as:

	Condition A	Condition B
Gene 1	$g_{1A}, g'_{1A}, g''_{1A}$	$g_{1B}, g'_{1B}, g''_{1B}$
Gene 2	$g_{2A}, g'_{2A}, g''_{2A}$	$g_{2B}, g'_{2B}, g''_{2B}$
Gene 3	$g_{3A}, g'_{3A}, g''_{3A}$	$g_{3B}, g'_{3B}, g''_{3B}$
Gene…	$g_{\ldots A}, g'_{\ldots A}, \mathrm{g}''_{\ldots A}$	$g_{\ldots B}, g'_{\ldots B}, g''_{\ldots B}$
Gene *n*	$g_{nA}, g'_{nA}, g''_{nA}$	$g_{nB}, g'_{nB}, g''_{nB}$

where g, g' and g'' denote replicated measurements (here in triplicate) of expression intensities in A versus B conditions, normalized to a common internal control (e.g. exogenous mRNA). By definition, the genes deemed to exhibit significant expression changes will be those for which the absolute difference in the average expression intensities $|\overline{g}_B - \overline{g}_A|$ is much larger than the estimated standard error $\hat{\sigma}_A$ or $\hat{\sigma}_B$ computed from the dispersion of g, g' and g'' measurements. Multiple independent experiments are thus essential to the assessment of significance with traditional statistical procedures such as the unrelated *t*-test. For a given confidence level, smaller differences will be required as the number of replicate measurements increases. For experiments performed in duplicate, $|\overline{g}_B - \overline{g}_A|$ has to be larger than $4.3\hat{\sigma}$ ($\hat{\sigma}$ is the estimated standard error) to reach the 5% significance threshold, and larger than $22.3\hat{\sigma}$ for the 1% significance level. For experiments performed in triplicate, the requirements are $|\overline{g}_B - \overline{g}_A| > 2.8\hat{\sigma}$ and $|\overline{g}_B - \overline{g}_A| > 5.2\hat{\sigma}$ for the 5 and 1% levels, respectively. Most published large-scale studies are quite elusive about measurement reproducibility and the confidence levels of the observed changes in expression are rarely assessed using standard methods. When the information is available, experiments have been done in triplicate (64), in duplicate (32,36) or only partially duplicated (http://cmgm.stanford.edu/~kimlab/). Some redundancy (e.g. 20 probes/gene, some probe sets duplicated) is built into the Affimetrix oligonucleotide array technology and directly used by the data acquisition software (GENECHIP; Affymetrix, Santa Clara, CA). However, this does not alleviate the need to assess the dispersion of expression intensities obtained from different chips and different complex mRNA probes.

In the above studies, the traditional methods to assess the statistical significance of the observed differences are not used. Instead, *ad hoc* thresholding procedures are used, resulting in the elimination of subsets of genes and expression measurements. An all-or-none 'reliable' versus 'unreliable' classification is thus used in place of a progressive ranking of expression changes according to *p* values. In the rare cases where the filtering procedure is described in enough detail, it can then be compared with a more traditional significance assessment.

For instance, Schena *et al.* (32), in their pioneering work on large-scale cDNA microarrays using two-colour fluorescence hybridization, adopted the rule of only retaining expression intensities for which the difference of duplicate measurements did not exceed half their average. Translated into classical statistical terms, this constraint corresponds to:

$$\hat{\sigma}_{\mathrm{A}} \le \frac{\bar{g}_{\mathrm{A}}}{\sqrt{2}} \text{ and } \hat{\sigma}_{\mathrm{B}} \le \frac{\bar{g}_{\mathrm{B}}}{\sqrt{2}} \tag{2}$$

Then, they classified 'differentially expressed' genes as those exhibiting at least a 2-fold change in expression, i.e.:

$$\frac{\bar{g}_{\mathrm{B}}}{\bar{g}_{\mathrm{A}}} \ge 2 \text{ or } \frac{\bar{g}_{\mathrm{B}}}{\bar{g}_{\mathrm{A}}} \le \frac{1}{2} \tag{3}$$

A straightforward combination of the two previous equations shows that expression changes considered to be significant may include cases for which:

$$|\bar{g}_{\mathrm{B}} - \bar{g}_{\mathrm{A}}| \ge \sqrt{2}\hat{\sigma}_{\mathrm{A}} \tag{4}$$

Expression differences of the order of $1.5\hat{\sigma}$ do not reach the 5% significance threshold (which is $4.3\hat{\sigma}$ for duplicate experiments). Thus, the filtering procedure used in this work is not conservative enough. In their simultaneous monitoring of 1000 genes, Schena *et al.* (32) found 15 genes exhibiting expression ratios of ~2. For purely statistical reasons, we expect that a fraction of those genes might be 'false positive' rather than *bona fide* differentially expressed genes. I noticed that, in their first study with 46 *Arabidopsis* cDNAs (36), the same authors adopted a more conservative $g_{\mathrm{B}}/g_{\mathrm{A}} \ge 5$ ratio as their threshold for differential expression. Combined with their reproducibility constraint (2), this higher ratio is valid and does ensure that (on average) <2.5% of the calls for differentially expressed genes will be due to random fluctuations.

The role of significance testing and the Bonferroni correction

Large-scale hybridization experiments require numerous manipulations to produce the final expression data matrix. Various calibration steps are for instance needed to ensure the linearity of the fluorescence measurements and internal controls are used to transform fluorescence intensities into number of mRNA molecules. Various controls (e.g. 'housekeeping' genes) are also used to verify the consistency of expression intensities obtained from different mRNA pools and different microarrays. At the end, mathematical conversions (using offsets, logarithms, ratios, etc.) are used in the production of the final expression data. These many calibrations and normalization procedures might convey a false sense of security, in particular for protocols such as the elegant two-colour competitive hybridization assays in which error correction mechanisms may appear to be built in. Yet, if one wishes to associate a confidence level with the measured changes, it has to be clear that no data processing or elegant protocol can substitute for the requirement of multiple (at least two) independent determinations of the expression intensities.

Confidence levels offer a rational way to output and interpret the results of large-scale differential expression experiments. As discussed in Audic and Claverie (61), *p* values constitute an objective measure of the quality of the evidence (that a gene is differentially expressed) and can be used to rank the candidate genes (as in the output of database similarity searches) and to prioritize further analyses. For instance, the confidence levels associated with couples of expression intensities (in number of tags) such as [$g_{\mathrm{A}} = 10$, $g_{\mathrm{B}} = 26$, ratio 2.6] and [$g_{\mathrm{A}} = 1$, $g_{\mathrm{B}} = 5$, ratio 5] point out the former as far better evidence for differential expression. Similarly, the application of significance testing to microarray data would sort out the best candidate genes among confusing combinations of ratios and expression levels. Lowly expressed genes with expression ratio $g_{\mathrm{B}}/g_{\mathrm{A}} \approx 5$ might then become less promising than highly expressed genes with $g_{\mathrm{B}}/g_{\mathrm{A}} \approx 1.5$.

The use of confidence levels is also relative to the number of genes simultaneously tested. Given a (small) probability *p* that a result will occur by chance (i.e. its significance threshold), there is a probability:

$$P = 1 - (1 - p)^{\mathrm{N}} \approx Np \tag{5}$$

for this result to occur at least once in *N* independent trials. Thus, if candidate genes are selected on the basis of expression changes significant at the 5% level, a false prediction rate equal to 5% of the total number of assayed genes is expected. For a 1000 gene array, this is 50. Choosing a *p* value threshold thus corresponds to fixing the level of acceptable risk (i.e. the fraction of false leads an experimenter is willing to tolerate).

Conversely, significance testing can be used in the traditional way, i.e. to 'demonstrate' the reality of an observation. In this case, the experimenter will have to apply the so-called 'Bonferroni' correction when fixing its significance threshold. This simply consists of imposing a *p* value such as:

$$Np << 1 \tag{6}$$

to ensure the absence of 'false positives'. Given the large (and increasing) number of genes tested in microarray or in tag experiments this corresponds to very small *p* values (e.g. 5%/10 000 = 5×10^{-6}). Unfortunately, the constraints on the magnitude of expression changes and on the measurement accuracy required to achieve such a high confidence level might not be experimentally feasible. Strict application of the Bonferroni correction could discard many biologically significant changes.

Multi-conditional gene expression analysis

The various technologies allowing parallel expression monitoring of large sets of genes are now being applied to the study of development and differentiation (43,49,54,64,68,81; http://cmgm.stanford.edu/~kimlab/) and of the transcriptional response to various factors in yeast (31,33,56,65,82–85) and mammalian (32,65,86) cells. In this context, the data take the form of a multi-conditional expression intensity matrix:

	Condition A	Condition B	Condition C	Condition…	Condition Z
Gene 1	$g_{1\mathrm{A}}$	$g_{1\mathrm{B}}$	$g_{1\mathrm{C}}$	$g_{1\ldots}$	$g_{1\mathrm{Z}}$
Gene 2	$g_{2\mathrm{A}}$	$g_{2\mathrm{B}}$	$g_{2\mathrm{C}}$	$g_{2\ldots}$	$g_{2\mathrm{Z}}$
Gene 3	$g_{3\mathrm{A}}$	$g_{3\mathrm{B}}$	$g_{3\mathrm{C}}$	$g_{3\ldots}$	$g_{3\mathrm{Z}}$
Gene…	$g_{\ldots\mathrm{A}}$	$g_{\ldots\mathrm{B}}$	$g_{\ldots\mathrm{C}}$	$g_{\ldots}$	$g_{\ldots\mathrm{Z}}$
Gene *n*	$g_{n\mathrm{A}}$	$g_{n\mathrm{B}}$	$g_{n\mathrm{C}}$	$g_{n\ldots}$	$g_{n\mathrm{Z}}$

where the various A–Z conditions might correspond to a time series (i.e. after stimulation), successive stages of differentiation, various stages of a disease process (cancer), growth conditions or tissue or cell types. The same analysis might also be used to investigate the transcriptional effects of drugs or gene transfer.

As before, the expression intensities *g* may consist of cDNA tag counts (i.e. each condition corresponds to the sampling of a different library) or analog values obtained from quantitative RT–PCR, microarray-based protocols or even protein two-dimensional gel electrophoresis.

Detecting differential expression from multi-conditional expression data

Obviously, the results of a multi-conditional expression experiment can still be used to identify differentially expressed genes by comparing gene expression levels between any pair of conditions. However, the proper Bonferroni correction will have to be applied to assess the statistical significance of the results. For an expression intensity table with *M* conditions, the correction factor (i.e. multiplying the probability for a result occurring by chance) is $M(M - 1)/2$.

Greller and Tobin (67) have proposed a new and robust computational method for the identification of ‘selective expression’ (i.e. a pattern in which the expression is markedly high or markedly low in a single particular condition) from multiple condition expression data. The method combines assessments of the reliability of expression measurements with a statistical test of expression profiles. They consider that measurements in at least 10 different conditions are required to make a reliable assessment of exceptional gene expression intensities.

Beyond the detection of differential expression, however, two new types of analyses can be performed using multi-conditional expression data, namely: (i) the identification of pairs of genes exhibiting coordinated expression; and (ii) the clustering of genes according to their expression profiles.

The identification of coordinated gene expression: pairwise analysis

Each row of a multi-conditional expression matrix corresponds to a gene expression profile, technically a vector in a space of *M* dimensions (where *M* is the number of assayed conditions). A given gene is thus represented as:

$$\text{gene i} = \{\, g_{iA}, g_{iB}, g_{iC}, g_{i...}, \ldots, g_{iZ}\}$$

Two genes can exhibit various forms of ‘coordinated expression’ (Fig. 1). At the qualitative level, they might tend to be expressed together (genes 1 and 3) or exclude each other (gene 2 versus genes 1 and 3). At the quantitative level, their abundance might follow a linear dependency or a more complex relationship (quadratic, sigmoidal, exponential, etc.). A simplified statistical procedure to identify pairs of genes exhibiting correlated expression is described in the Appendix. The basic principle behind all methods is that coordinated expression will be suspected when the expression profiles of two genes are more similar (or more dissimilar) than expected by chance. Coordinated expression is thus inferred through pairwise comparisons of all rows (gene profile vectors) in the expression data matrix.

Various methods can be used to assess the pairwise similarity of gene expression profiles. For tag sampling experiments, for instance, 2 × 2 co-expression contingency tables can be computed from the multi-conditional expression data, such as:

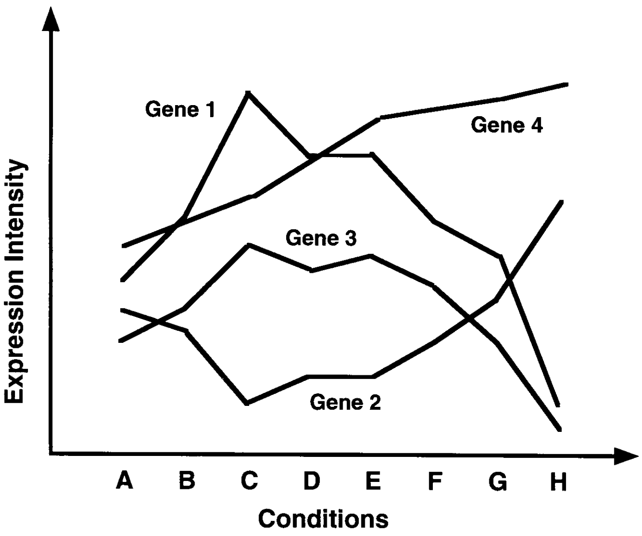

Figure 1. Example of expression profiles (fictitious data). Gene 1 = $\{g_{1A}, g_{1B}, g_{1C}, g_{1D}, \ldots, g_{1H}\}$, gene 2, gene 3 and gene 4 vectors are represented as profiles using their expression intensities as coordinates for the various conditions (or time points) A–H. The profiles for genes 1 and 3 have a similar overall shape, suggesting a correlated expression. The profile for gene 2 exhibits the opposite variation, suggesting an anti-correlation with genes 1 and 3. Thus, genes 1–3 illustrate coordinated expression patterns. The profile for gene 4 indicates an expression pattern independent of the other three genes.

	Gene 2 detected	Gene 2 not detected
Gene 1 detected	In 10 libraries	In 2 libraries
Gene 1 not detected	In 2 libraries	In 8 libraries

For the above example, the application of Fisher’s 2 × 2 exact test would indicate that a significant association exists between the occurrence of genes 1 and 2 in the panel of the sampled 22 libraries. The same design is capable of detecting anti-association as well. However, reducing the tag counts to a binary scale (detected versus not detected) is only advisable when the quantitative data are unreliable, such as in the case of normalized libraries.

For both tag sampling and ‘analog’ expression data, Spearman’s rank correlation test provides a way to assess the overall shape similarity of two expression profiles. For each gene, the conditions are ranked according to expression intensities. For instance, gene 1 follows the decreasing order C, D, E, B, F, G, A, H (Fig. 1). Genes associated with ‘parallel’ profiles, for instance gene 3, correspond to a similar order: C, E, D, F, B, A, G, H. Spearman’s rank correlation r_s is simply a linear correlation computed on the ranked expression intensities of genes 1 and 3, as in the table:

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8
Gene 1	g_{1C}	g_{1D}	g_{1E}	g_{1B}	g_{1F}	g_{1G}	g_{1A}	g_{1H}
Gene 2	g_{3C}	g_{3E}	g_{3D}	g_{3F}	g_{3B}	g_{3A}	g_{3G}	g_{3H}

Kendall’s significance test can also be used for the purpose of assessing a correlation in rank order. While the ranking procedure may cause important information loss in the data, it is well suited to the detection of strongly non-linear correlations between two genes.

Nevertheless, the traditional Pearson’s linear correlation coefficient gives good results in most cases and has been used in many studies (32,65,84,86–88). This test can detect pairs of genes with similar (gene i = gene j), proportional (gene i $\propto$ gene j) or opposite (gene i = $-\propto$ gene j) expression profiles. Pearson's linear correlation coefficient is also associated with a p value which assesses the confidence level for suspected coordinated expression.

The results of a complete pairwise correlation analysis can be summarized in a matrix of 'expression similarity' across all genes, such as:

	Gene 1	Gene 2	Gene 3	Gene…	Gene n
Gene 1	r_{11}	r_{12}	r_{13}	$r_{1\ldots}$	r_{1n}
Gene 2	r_{21}	r_{22}	r_{23}	$r_{2\ldots}$	r_{2n}
Gene 3	r_{31}	r_{32}	r_{33}	$r_{3\ldots}$	r_{3n}
Gene…	$r_{\ldots 1}$	$r_{\ldots 2}$	$r_{\ldots 3}$	$r_{\ldots}$	$r_{\ldots n}$
Gene n	r_{n1}	r_{n2}	r_{n3}	$r_{n\ldots}$	r_{nn}

There are a number of ways of visualizing these results. They involve different methods of classifying the genes that exhibit correlated expression patterns into 'similarity clusters'. The simplest procedure will use the property of transitivity: if gene i is significantly correlated with gene j and gene j significantly correlated with gene h, then gene i, j and h are put in the same cluster. Unfortunately, this method is very sensitive to the choice of an arbitrary threshold and to the uncertainty of pairwise gene correlation assessment. A better procedure consists of using the concept of Euclidean distance (square root of the sum of the squared differences in each dimension) to transform the above gene correlation table into a matrix of *bona fide* pairwise distances (88). For instance, the distance between genes 1 and 2 is computed as:

$$d_{1,\,2} = \sqrt{(r_{11}-r_{21})^2 + (r_{12}-r_{22})^2 + (r_{13}-r_{23})^2 + \ldots + (r_{1n}-r_{2n})^2} \tag{7}$$

An important point is that the distance $d_{1,2}$ between gene 1 and gene 2 now takes into account their similarity with all other genes and is no longer computed from a single pairwise comparison. Such a distance is then less sensitive to the random fluctuation of expression measurements. Using Euclidean distances, two genes exhibiting a poor pairwise correlation might still appear close by virtue of their correlation patterns with other genes. Indeed, other types of Euclidean distance can be computed from the multi-conditional expression matrix, for instance by directly using the expression intensities for each condition as coordinates (64).

Once a set of *bona fide* pairwise distances is available, a number of clustering methods can be applied to reveal subsets of genes obeying similar patterns of expression. These methods are discussed in the next section.

Identifying gene expression clusters

Pairwise gene distance matrices computed from expression analysis or from sequence alignments are similar mathematical objects. Methods traditionally used for molecular phylogeny can thus be used to identify clusters of genes sharing a similar expression pattern. Wen *et al.* (64), for instance, used the Fitch algorithm (89) in the Phylip package (90,91) to interpret the temporal gene expression of 112 genes during the development of the rat central nervous system. The resulting tree clearly identified four major subsets of genes sharing four types of expression profile. A hierarchical (i.e. tree building) method adapted to the direct clustering of the correlation matrices mentioned above has been used by Eisen *et al.* (65) to analyse a 12 point time course of the serum response of 8600 human genes and a 75 condition expression study of the whole yeast genome. For other examples of the use of hierarchical clustering see Schena *et al.* (32) Lashkari *et al.* (33) and Khan *et al.* (81).

As pointed out by Tamayo *et al*. (68), hierarchical methods have their shortcomings and are best suited to situations of actual hierarchical descents (such as in molecular evolution). Fortunately, the design of clustering methods is a well-established field of research and a large number of alternative procedures are possible and can be used.

Principal component analysis (PCA), for instance, can be directly applied to a matrix of multiple condition expression intensities to compute the relative positions of genes and project them into the most discriminant three- or two-dimensional space (64,92). Visual inspection or pattern recognition software can then be used to delineate expression clusters.

Computer scientists are also beginning to specifically adapt classical graph theory-based clustering techniques (93) to the problem of analysing 'noisy' expression data. For instance, the *corrupted clique-graph* model and the cluster affinity search technique (CAST) proposed by Ben-Dor and Yakhini (94) were successfully tested on two large multiple experiment data sets. The HCS algorithm proposed by Hartuv *et al.* (95) to cluster collections of cDNAs on the basis of their oligonucleotide fingerprints is also a good example of a graph theoretical technique designed to tolerate large stochastic perturbations.

The traditional hierarchical approaches, principal component analyses and the graph theoretical techniques cited above, are all examples of clustering procedures not requiring any assumption of the number of clusters sought.

Other methods, such as the self-organizing map (SOM) procedure adopted by Tamayo *et al.* (68) require fixation of the number of 'nodes' (which will serve as nucleation points for the genes clusters), as well as their initial geometry in the space of the multi-conditional expression intensities. However, SOMs can be computed very quickly, even on a large data set. An iterative procedure can thus be implemented to explore the underlying cluster structure and converge to an optimal partition.

Finally, the most sophisticated clustering methods are those aiming at inferring causal relationships and regulatory mechanisms from multi-conditional expression measurements: the genes are still partitioned into clusters, but the partition now has an internal structure involving inhibitory or activating interactions. Genes belonging to the same cluster are now integrated into a coherent pathway. Initially developed for the analysis of chemical reactions (96) or the interpretation of complex genetic networks (97), such clustering approaches are now adapted to the analysis of large-scale expression experiments and to the modelling (or 'reverse engineering') of transcriptional regulatory pathways (98,99). In a recent work, Chen *et al.* (100) addressed the problem of identifying a small set of candidate regulatory genes from multiple time series of expression measurements. Using Boolean circuits to model biological pathways, Karp *et al.* (101) are tackling the problem from the other end, by designing algorithms for choosing the most appropriate conditional expression experiments (e.g. gene

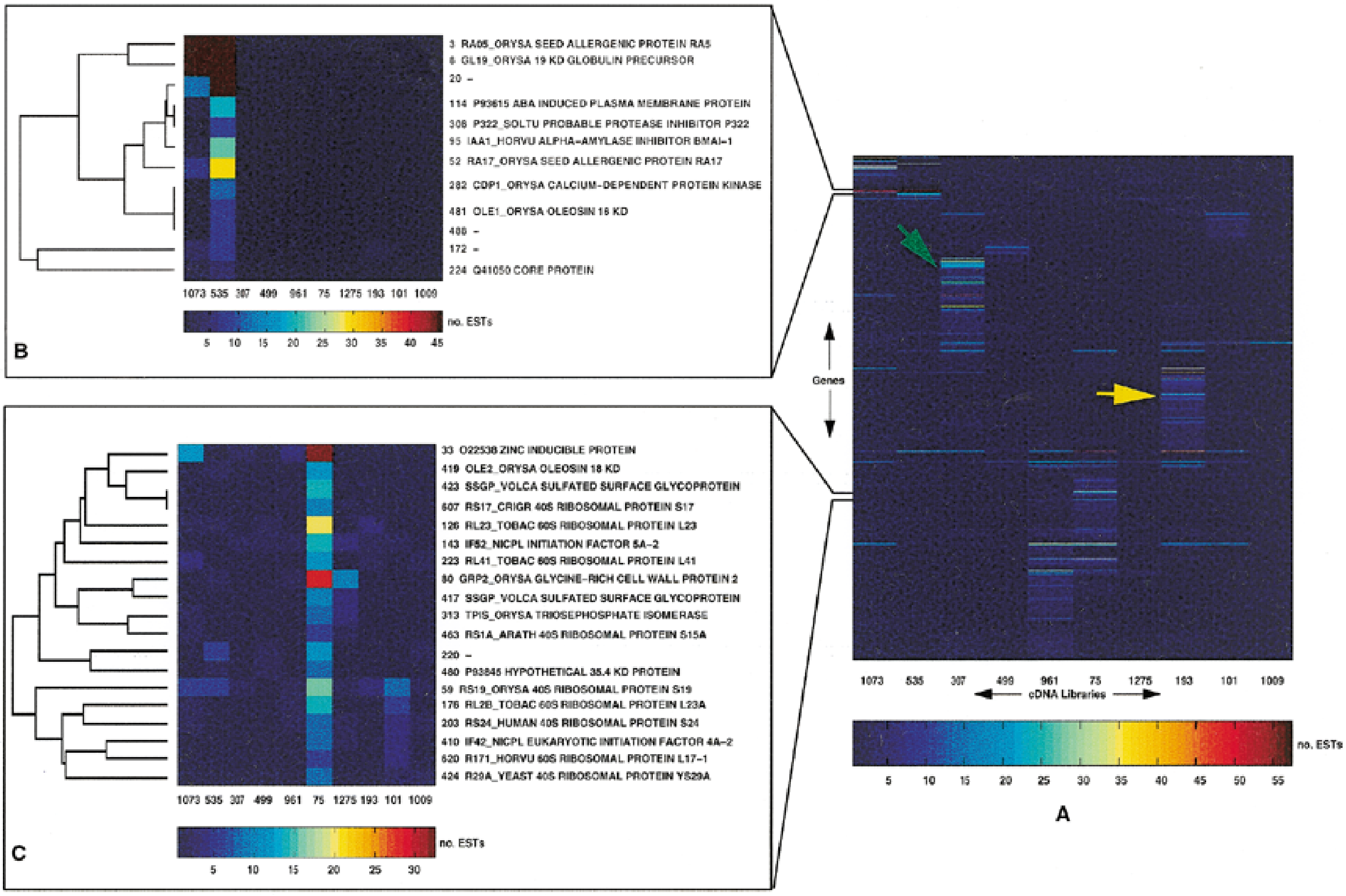

Figure 2. Example of a colour map derived from rice EST data. (**A**) Colour map generated from a gene expression correlation analysis (88) of publicly available rice EST data (707 genes in 10 cDNA libraries). EST counts are represented according to the colour scale shown below the map. The colour scale has been chosen so as to optimally represent the [5–55] range of EST counts (cells with a value >55 are assigned the colour red). The green and yellow arrows point to groups of genes with specific expression patterns. The green arrow points to a set of genes mostly expressed in library 307 (green shoot, 8 days old) and the yellow arrow to genes mostly expressed in library 193 (etiolated shoot, 8 days old). (**B** and **C**) Different magnified regions of the colour map. To the right of each region, genes are shown with their putative identification, if available. To the left of each region, the relevant portion of the tree used to re-order the original data table is shown. Different colour scales were used in (A), (B) and (C), to provide optimal contrast in the display of the EST counts.

disruption or external stimulus) that will reveal underlying regulatory networks.

Clustering methods and graphical displays are two closely related aspects of the interpretation of multi-conditional expression experiments. After hierarchical clustering, for instance, trees and colour maps are the most natural representations (64,65,92). A colour map is designed as follows (Fig. 2). Given the gene hierarchy, the rows (i.e. the genes) of the primary data matrix can be re-ordered by placing the genes sharing similar expression profiles next to each other. With the exception of time series, a similar re-ordering procedure can also be applied to the columns (e.g. tissue type, growth condition or pathological samples) most similar in terms of gene expression. The re-ordered primary data table can then be displayed by colouring each cell on the basis of intensity, variability, gene function, etc. Visual inspection of the resulting colour map will often reveal domains of similar or contrasting colours or remarkable shapes, eventually suggesting new regulatory pathways as well as disease or differentiation mechanisms. Standard image processing techniques (e.g. contrast enhancing, boundary detection, etc.) may be used to supplement human natural talent for pattern recognition.

DISCUSSION

Coordinated gene expression analysis in functional genomics

The elegant two-hybrid system assay (102) is one of the most popular techniques in functional genomics. Given the cDNA of a protein of interest, this technique allows the identification of other proteins capable of interacting with it directly from a pool of target cDNAs. In this assay, the specific physical interaction between the probe and target proteins is directly used to trigger a reporter gene.

The computational analysis of a multi-conditional gene expression experiment can be seen as an extension of this technique, using the statistical interaction between the expression data of genes, rather than the physical interaction between their products.

The network of interaction revealed by the computational technique may encompass genes involved in the same biological pathway in a non-contiguous manner, as well as genes negatively interfering with each other.

In practice, the detection of correlated/coordinated gene expression nicely complements sequence-based bioinformatics methods in three main ways:

- in assigning a precise biological pathway to a gene of 'generic' function (such as transcription factor or kinase);
- in relating an anonymous gene to better characterized genes;
- in revealing unexpected relationship between previously known genes or pathways.

Gene expression correlation analyses might also considerably help sequence-based bioinformatics approaches in the study of eukaryotic promoters. Among genes exhibiting correlated expression patterns across a large panel of biological conditions, a significant fraction is expected to be co-regulated, i.e. responsive to a common set of expression factors. The promoters of these genes should then contain common regulatory elements. Thus, the identification of gene expression clusters constitutes precious accessory information for the 'reverse engineering' of the very elusive architecture of eukaryotic promoters. This opportunity did not escape the attention of Brazma *et al.* (103), who have systematically analysed the upstream regions of yeast genes exhibiting similar expression profiles. Completion of the genomic sequence of the nematode *Caenorhabditis elegans* should allow a similar study to be run on a multicellular organism, using the large gene expression data set (~150 conditions) generated by Kim and co-workers (http://cmgm.stanford.edu/~kimlab/).

It is worth noting that the computational approaches reviewed here can be applied at the protein level. In the search for correlations, cDNA microarrays and gene expression levels are simply replaced by two-dimensional electrophoresis and protein spot intensities (60,104). In the few comparative studies that have been performed, important discrepancies have been noted between expression measurements made at the transcriptome versus the proteome level (104,105). It has been suggested that protein spot signatures correlate better with phenotype than gene expression intensity (104).

Finally, the same computational techniques are also being used in the information intensive, massively parallel, drug screening protocols (106–108). Euclidean distances, hierarchical clustering and colour maps are very familiar concepts in the interpretation of large-scale (e.g. 49 000 compounds tested against 60 cell lines) molecular pharmacology studies (108).

Future directions

I have distinguished three levels in the interpretation of multi-condition gene expression data:

(i) the identification of differentially expressed genes;
(ii) the identification of pairwise gene expression correlations;
(iii)the delineation of gene clusters according to gene expression patterns.

The computational methods to accomplish step 1 are well worked out and the complexity of the task is only of the order of N, the total number of genes analysed in the multi-conditional expression experiment.

Methods to perform step 2 are also well defined. Given the *a priori* unknown mathematical form of the correlation, the best approach would certainly involve the use of a variety of tests, each of them best suited to the recognition of a specific type of dependency (Pearson's, Spearman's, mutual information, etc.) (87). Provided a large number of conditions are tested, it is also conceivable that a relationship more subtle than a simple monotonic dependency (i.e. pairs of genes always going up or down together, or the opposite) might become detectable (e.g. pairs of genes positively correlated in some conditions and negatively correlated in some others). In any case, the identification of pairwise correlation has a complexity of the order of $N(N-1)/2$ and is not a computational difficulty for modern computers.

The real challenge remains in the clustering step, for which algorithmic approaches abound, but the best choice is not clear to biologists. For most clustering methods, the complexity is of the order of $N\log(N)$, and again is not a real computational difficulty. However, experimental errors and the complexity of the underlying regulatory network structure require that arbitrary similarity thresholds or minimal graph connectivity rules are incorporated in the practical implementation of all algorithms. The difficulty with clustering is thus not in the design, or the choice, of a perfect method, but rather in the fact that all algorithms will fail for an unknown fraction of the cases and that there is no simple way to decide which will perform best for an arbitrary (experimental) data set. Bioinformatics research is thus very active in this area, but the prospect is poor that alternative clustering protocols will produce vastly different biological results from the same multi-conditional expression data.

Given the complexity of regulatory networks, it is also not clear whether clustering by traditional methods (PCA, hierarchical clustering, etc.) is adding more to our understanding of biological pathways than the simple knowledge of all pairwise gene expression correlations. In practice, most academic or industrial biologists will be mining multi-conditional gene expression data with an idea in mind and look for correlations with previously defined genes, such as specific tumor suppressors, cytokines or membrane receptors, in a search for surrogate markers or alternative targets. Improving clustering might thus be both difficult and biologically pointless.

The true, more biologically relevant, exciting future of gene expression clustering might thus lie in the more abstract researches attempting a complete reverse engineering of transcriptional regulation networks. Only such approaches have the potential to produce an integrated view of the cell pathways from the intricate combination of individual gene expression patterns. Detailed modelling of signalling pathways and the establishment of causal relationships are required to model developmental mechanisms and elucidate, for example, how commonly used signalling pathways are able to elicit tissue-specific responses in multicellular organisms. However, current works (101,109) in this area are only at the preliminary stage of defining which constraints must be fulfilled to allow a complex regulatory network architecture to be inferred from gene expression patterns. It is thus too early to tell whether this goal will ever be reachable from a reasonable number of experiments.

ACKNOWLEDGEMENTS

I wish to thank C. Abergel, S. Audic and F. Gosse for reading the manuscript and R. Ewing and A. Ben Kahla for allowing the use of their results in Figure 2 prior to publication.

REFERENCES

1. The C. elegans Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.
2. Dickson, D. (1998) *Drosophila* set for fast-track sequencing. *Nature*, **393**, 296.
3. Ecker, J.R. (1998) Genome sequencing: genes blossom from a weed. *Nature*, **391**, 438–439.
4. Wadman, M. (1999) Human genome project aims to finish 'working draft' next year. *Nature*, **398**, 177.
5. Claverie, J.-M. (1997) Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.*, **6**, 1735–1744.
6. Werner, T. (1999) Models for prediction and recognition of eukaryotic promoters. *Mamm. Genome*, **10**, 168–175.
7. Audic, S. and Claverie, J.-M. (1998) Visualizing the competitive recognition of TATA-boxes in vertebrate promoters. *Trends Genet.*, **14**, 10–11.
8. Fickett, J.W. and Hatzigeorgiou, A.G. (1997) Eukaryotic promoter recognition. *Genome Res.*, **7**, 861–878.
9. Hofmann, K., Bucher, P., Falquet, L. and Bairoch, A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
10. Corpet, F., Gouzy, J. and Kahn, D. (1999) Recent improvements of the ProDom database of protein domain families. *Nucleic Acids Res.*, **27**, 263–267.
11. Attwood, T.K., Flower, D.R., Lewis, A.P., Mabey, J.E., Morgan, S.R., Scordis, P., Selley, J.N. and Wright, W. (1999) PRINTS prepares for the new millennium. *Nucleic Acids Res.*, **27**, 220–225.
12. Ponting, C.P., Schultz, J., Milpetz, F. and Bork, P. (1999) SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res.*, **27**, 229–232.
13. Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D. and Sonnhammer, E.L. (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.*, **27**, 260–262.
14. Henikoff, J.G., Henikoff, S. and Pietrokovski, S. (1999) New features of the Blocks Database servers. *Nucleic Acids Res.*, **27**, 226–228.
15. Liang, P. and Pardee, A.B. (1992) Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science*, **257**, 967–971.
16. Liang, P. and Pardee, A.B. (1998) Differential display. A general protocol. *Mol. Biotechnol.*, **10**, 261–267.
17. Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y. and Matsubara, K. (1992) Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genet.*, **2**, 173–179.
18. Adams, M., Dubnick, M., Kerlavage, A., Moreno, R., Kelley, J., Utterback, T., Nagle, J., Fields, C. and Venter, J. (1992) Sequence identification of 2 375 human brain genes. *Nature*, **355**, 632–634.
19. Matsubara, K. and Okubo, K. (1993) Identification of new genes by systematic analysis of cDNAs and database construction. *Curr. Opin. Biotechnol.*, **4**, 672–677.
20. Lee, N.H., Weinstock, K.G., Kirkness, E.F., Earle-Hughes, J.A., Fuldner, R.A., Marmaros, S., Glodek, A., Gocayne, J.D., Adams, M.D., Kerlavage, A.R. *et al.* (1995) Comparative expressed-sequence-tag analysis of differential gene expression profiles in PC-12 cells before and after nerve growth factor treatment. *Proc. Natl Acad. Sci. USA*, **92**, 8303–8307.
21. Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., White, O. *et al.* (1995) Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*, **377** (suppl. 6547), 3–174.
22. Aaronson, J., Eckman, B., Blevins, R., Borkowski, J., Myerson, J., Imran, S. and Elliston, K. (1996) Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. *Genome Res.*, **6**, 829–845.
23. Hillier, L.D., Lennon, G., Becker, M., Bonaldo, M.F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., Hawkins, M., Hultman, M., Kucaba, T., Lacy, M., Le, M., Le, N., Mardis, E., Moore, B., Morris, M., Parsons, J., Prange, C., Rifkin, L., Rohlfing, T., Schellenberg, K., Marra, M. *et al.* (1996) Generation and analysis of 280 000 human expressed sequence tags. *Genome Res.*, **6**, 807–828.
24. Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. *Science*, **270**, 484–487.
25. Ryo, A., Kondoh, N., Wakatsuki, T., Hada, A., Yamamoto, N. and Yamamoto, M. (1998) A method for analyzing the qualitative and quantitative aspects of gene expression: a transcriptional profile revealed for HeLa cells. *Nucleic Acids Res.*, **26**, 2586–2592.
26. Datson, N.A., van der Perk-de Jong, J., van den Berg, M.P., de Kloet, E.R. and Vreugdenhil, E. (1999) MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue. *Nucleic Acids Res.*, **27**, 1300–1307.
27. Meier-Ewert, S., Lange, J., Gerst, H., Herwig, R., Schmitt, A., Freund, J., Elge, T., Mott, R., Herrmann, B. and Lehrach, H. (1998) Comparative gene expression profiling by oligonucleotide fingerprinting. *Nucleic Acids Res.*, **26**, 2216–2223.
28. Lipshutz, R.J., Morris, D., Chee, M., Hubbell, E., Kozal, M.J., Shah, N., Shen, N., Yang, R. and Fodor, S.P. (1995) Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques*, **19**, 442–447.
29. Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S. and Fodor, S.P. (1996) Accessing genetic information with high-density DNA arrays. *Science*, **274**, 610–614.
30. Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.*, **14**, 1675–1680.
31. Wodicka, L., Dong, H., Mittmann, M., Ho, M.H. and Lockhart, D.J. (1997) Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnol.*, **15**, 1359–1367.
32. Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. and Davis, R. (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl Acad. Sci. USA*, **93**, 10614–10619.
33. Lashkari, D.A., DeRisi, J.L., McCusker, J.H., Namath, A.F., Gentile, C., Hwang, S.Y., Brown, P.O. and Davis, R.W. (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl Acad. Sci. USA*, **94**, 13057–13062.
34. Schena, M. (1996) Genome analysis with gene expression microarrays. *BioEssays*, **18**, 427–431.
35. Shalon, D., Smith, S.J. and Brown, P.O. (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.*, **6**, 639–645.
36. Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
37. Watson, A., Mazumder, A., Stewart, M. and Balasubramanian, S. (1998) Technology for microarray analysis of gene expression. *Curr. Opin. Biotechnol.*, **9**, 609–614.
38. Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. and Trent, J.M. (1999) Expression profiling using cDNA microarrays. *Nature Genet.*, **21** (suppl. 1), 10–14.
39. Cheung, V.G., Morley, M., Aguilar, F., Massimi, A., Kucherlapati, R. and Childs, G. (1999) Making and reading microarrays. *Nature Genet.*, **21** (suppl. 1), 15–19.
40. Brown, P.O. and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays. *Nature Genet.*, **21** (suppl. 1), 33–37.
41. Debouck, C. and Goodfellow, P.N. (1999) DNA microarrays in drug discovery and development. *Nature Genet.*, **21** (suppl. 1), 48–50.
42. Schena, M., Heller, R.A., Theriault, T.P., Konrad, K., Lachenmeier, E. and Davis, R.W. (1998) Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol.*, **16**, 301–306.
43. Strausberg, R.L., Dahl, C.A. and Klausner, R.D. (1997) New opportunities for uncovering the molecular basis of cancer. *Nature Genet.*, **16** (suppl.), 415–516.
44. Hwang, D.M., Dempsey, A.A., Wang, R.X., Rezvani, M., Barrans, J.D., Dai, K.S., Wang, H.Y., Ma, H., Cukerman, E., Liu, Y.Q., Gu, J.R., Zhang, J.H., Tsui, S.K., Waye, M.M., Fung, K.P., Lee, C.Y. and Liew, C.C. (1997) A genome-based resource for molecular cardiovascular medicine: toward a compendium of cardiovascular genes. *Circulation*, **96**, 4146–4203.
45. Ji, H., Liu, Y.E., Jia, T., Wang, M., Liu, J., Xiao, G., Joseph, B.K., Rosen, C. and Shi, Y.E. (1997) Identification of a breast cancer-specific gene, BCSG1, by direct differential cDNA sequencing. *Cancer Res.*, **57**, 759–764.
46. Adjaye, J., Daniels, R., Bolton, V. and Monk, M. (1997) cDNA libraries from single human preimplantation embryos. *Genomics*, **46**, 337–344.
47. Shimizu-Matsumoto, A., Adachi, W., Mizuno, K., Inazawa, J., Nishida, K., Kinoshita, S., Matsubara, K. and Okubo, K. (1997) An expression profile of genes in human retina and isolation of a complementary DNA for a

novel rod photoreceptor protein. *Invest. Ophthalmol. Vis. Sci.*, **38**, 2576–2585.

48. Mao, M., Fu, G., Wu, J.S., Zhang, Q.H., Zhou, J., Kan, L.X., Huang, Q.H., He, K.L., Gu, B.W., Han, Z.G., Shen, Y., Gu, J., Yu, Y.P., Xu, S.H., Wang, Y.X., Chen, S.J. and Chen, Z. (1998) Identification of genes expressed in human CD34 (+) hematopoietic stem/progenitor cells by expressed sequence tags and efficient full-length cDNA cloning. *Proc. Natl Acad. Sci. USA*, **95**, 8175–8180.
49. Gubbay, J., Doyle, J.P., Skinner, M. and Heintz, N. (1998) Changing patterns of gene expression identify multiple steps during regression of rat prostate *in vivo*. *Endocrinology*, **139**, 2935–2943.
50. Vasmatzis, G., Essand, M., Brinkmann, U., Lee, B. and Pastan, I. (1998) Discovery of three genes specifically expressed in human prostate by expressed sequence tag database analysis. *Proc. Natl Acad. Sci. USA*, **95**, 300–304.
51. Nelson, P.S., Ng, W.L., Schummer, M., Tru, L.D., Liu, A.Y., Bumgarner, R.E., Ferguson, C., Dimak, A. and Hood, L. (1998) An expressed-sequence-tag database of the human prostate: sequence analysis of 1168 cDNA clones. *Genomics*, **47**, 12–25.
52. Tanabe, K., Nakagomi, S., Kiryu-Seo, S., Namikawa, K., Imai, Y., Ochi, T., Tohyama, M. and Kiyama, H. (1999) Expressed-sequence-tag approach to identify differentially expressed genes following peripheral nerve axotomy. *Brain Res. Mol. Brain Res.*, **64**, 34–40.
53. Skvorak, A.B., Weng, Z., Yee, A.J., Robertson, N.G. and Morton, C.C. (1999) Human cochlear expressed sequence tags provide insight into cochlear gene expression and identify candidate genes for deafness. *Hum. Mol. Genet.*, **8**, 439–452.
54. Zhang, L., Zhou, W., Velculescu, V., Kern, S., Hruban, R., Hamilton, S., Vogelstein, B. and Kinzler, K. (1997) Gene expression profiles in normal and cancer cells. *Science*, **276**, 1268–1272.
55. Madden, S.L., Galella, E.A., Zhu, J., Bertelsen, A.H. and Beaudry, G.A. (1997) SAGE transcript profiles for p53-dependent growth regulation. *Oncogene*, **15**, 1079–1085.
56. Velculescu, V., Zhang, L., Zhou, W., Vogelstein, B., Basrai, M.D.B.Jr, Hieter, P., Vogelstein, B. and Kinzler, K. (1997) Characterization of the yeast transcriptome. *Cell*, **88**, 243–251.
57. He, T.C., Sparks, A.B., Rago, C., Hermeking, H., Zawel, L., da Costa, L.T., Morin, P.J., Vogelstein, B. and Kinzler, K.W. (1998) Identification of c-MYC as a target of the APC pathway. *Science*, **281**, 1509–1512.
58. Hibi, K., Liu, Q., Beaudry, G.A., Madden, S.L., Westra, W.H., Wehage, S.L., Yang, S.C., Heitmiller, R.F., Bertelsen, A.H., Sidransky, D. and Jen, J. (1998) Serial analysis of gene expression in non-small cell lung cancer. *Cancer Res.*, **58**, 5690–5694.
59. de Waard, V., van den Berg, B.M.M., Veken, J., Schultz-Heienbrok, R., Pannekoek, H. and van Zonneveld, A.J. (1999) Serial analysis of gene expression to assess the endothelial cell response to an atherogenic stimulus. *Gene*, **226**, 1–8.
60. Anderson, N.L., Hofmann, J.P., Gemmell, A. and Taylor, J. (1984) Global approaches to quantitative analysis of gene-expression patterns observed by use of two-dimensional gel electrophoresis. *Clin. Chem.*, **30**, 2031–2036.
61. Audic, S. and Claverie, J.-M. (1997) The significance of digital gene expression profiles. *Genome Res.*, **7**, 986–995.
62. Burke, J., Wang, H., Hide, W. and Davison, D.B. (1998) Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.*, **8**, 276–290.
63. Gautheret, D., Poirot, O., Lopez, F., Audic, S. and Claverie, J.-M. (1998) Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res.*, **8**, 524–530.
64. Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L. and Somogyi, R. (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl Acad. Sci. USA*, **95**, 334–339.
65. Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
66. Michaels, G.S., Carr, D.B., Askenazi, M., Fuhrman, S., Wen, X. and Somogyi, R. (1998) Cluster analysis and data visualization of large-scale gene expression data. *Pac. Symp. Biocomput.*, 42–53.
67. Greller, L.D. and Tobin, F.L. (1999) Detecting selective expression of genes and proteins. *Genome Res.*, **9**, 282–296.
68. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S. and Golub, T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.
69. Vingron, M. and Hoheisel, J. (1999) Computational aspects of expression data. *J. Mol. Med.*, **77**, 3–7.
70. Ringwald, M., Davis, G.L., Smith, A.G., Trepanier, L.E., Begley, D.A., Richardson, J.E. and Eppig, J.T. (1997) The mouse gene expression database GXD. *Semin. Cell Dev. Biol.*, **8**, 489–497.
71. Davidson, D., Bard, J., Brune, R., Burger, A., Dubreuil, C., Hill, W., Kaufman, M., Quinn, J., Stark, M. and Baldock, R. (1997) The mouse atlas and graphical gene-expression database. *Semin. Cell Dev. Biol.*, **8**, 509–517.
72. Miller, G., Fuchs, R. and Lai, E. (1997) IMAGE cDNA clones, UniGene clustering and ACeDB: an integrated resource for expressed sequence information. *Genome Res.*, **7**, 1027–1032.
73. Bailey, L.C.Jr, Searls, D.B. and Overton, G.C. (1998) Analysis of EST-driven gene annotation in human genomic sequence. *Genome Res.*, **8**, 362–376.
74. Ermolaeva, O., Rastogi, M., Pruitt, K.D., Schuler, G.D., Bittner, M.L., Chen, Y., Simon, R., Meltzer, P., Trent, J.M. and Boguski, M.S. (1998) Data management and analysis for gene expression arrays. *Nature Genet.*, **20**, 19–23.
75. Bassett, D.E.Jr, Eisen, M.B. and Boguski, M.S. (1999) Gene expression informatics—it's all in your mine. *Nature Genet.*, **21** (Suppl. 1), 51–55.
76. Stoeckert, C.J.Jr, Salas, F., Brunk, B. and Overton, G.C. (1999) EpoDB: a prototype database for the analysis of genes expressed during vertebrate erythropoiesis. *Nucleic Acids Res.*, **27**, 200–203.
77. Hawkins, V., Doll, D., Bumgarner, R., Smith, T., Abajian, C., Hood, L. and Nelson, P.S. (1999) PEDB: the Prostate Expression Database. *Nucleic Acids Res.*, **27**, 204–208.
78. Siegel, S. (1956) *Nonparametric Methods for the Behavioral Sciences*. McGraw-Hill, New York, NY.
79. Livak, K.J., Flood, S.J., Marmaro, J., Giusti, W. and Deetz, K. (1995) Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *PCR Methods Appl.*, **4**, 357–362.
80. Somogyi, R., Wen, X., Ma, W. and Barker, J.L. (1995) Developmental kinetics of GAD family mRNAs parallel neurogenesis in the rat spinal cord. *J. Neurosci.*, **15**, 2575–2591.
81. Khan, J., Simon, R., Bittner, M., Chen, Y., Leighton, S.B., Pohida, T., Smith, P.D., Jiang, Y., Gooden, G.C., Trent, J.M. and Meltzer, P.S. (1998) Gene expression profiling of alveolar Rhaddomyosarcoma with cDNA microarrays. *Cancer Res.*, **58**, 5009–5013.
82. DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, **278**, 680–686.
83. Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodika, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D. and Davis, R. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, **2**, 65–73.
84. Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *S. cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
85. Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O. and Herskowitz, I. (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699–705.
86. Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson, J., Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D. and Brown, P.O. (1999) The transcriptional program in the response of human fibroblast to serum. *Science*, **283**, 83–87.
87. D'Haeseleer, P., Wen, X., Fuhrman, S. and Somogyi, R. (1998) Mining the gene expression matrix: inferring gene relationships from large scale gene expression data (HTML). In Paton, R.C. and Holcombe, M. (eds), *Information Processing in Cells and Tissues*. Plenum Publishing, New York, NY, pp. 203–212.
88. Ewing, R.M., Ben Kahla, A., Poirot, O., Lopez, F., Audic, S. and Claverie, J.-M. (1999) Large-scale, statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res.*, **9**, in press.
89. Fitch, W.M. and Margoliash, E. (1967) Construction of phylogenic trees. *Science*, **155**, 279–284.
90. Felsenstein, J. (1993) *PHYLIP (Phylogeny Inference Package) Version 3.5c*. J. Felsenstein, Department of Genetics, University of Washington, Seattle, WA.
91. Felsenstein, J. (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). *Cladistics*, **5**, 164–166.

92. Carr, D.B., Somogyi, R. and Michaels, G. (1997) Templates for looking at gene expression clustering. *Stat. Comput. Stat. Graphics Newslett. (April)*, 20–29.
93. Matula, D.W. (1977) Graph theoretic techniques for cluster analysis algorithms. In van Ryzin, J. (ed.), *Classification and Clustering*. Academic Press, London, UK, pp. 95–129.
94. Ben-Dor, A. and Yakhini, Z. (1999) Clustering gene expression patterns. In Istrail, S., Pevzner, P. and Waterman, M. (eds), *Recomb 99*. ACM Press, Washington, DC, pp. 33–42.
95. Hartuv, E., Schmitt, A., Lange, J., Meirer-Ewert, S., Lehrach, H. and Shamir, R. (1999) An algorithm for clustering cDNAs for gene expression analysis. In Istrail, S., Pevzner, P. and Waterman, M. (eds), *Recomb 99*. ACM Press, Washington, DC, pp. 188–197.
96. Arkin, A., Shen, P. and Ross, J. (1997) A test case of correlation metric construction of a reaction pathway from measurements. *Science*, **277**, 1275–1279.
97. Thieffry, D. and Thomas, R. (1998) Qualitative analysis of gene network. *Pac. Symp. Biocomput.*, 77–87.
98. Michaels, G.S., Carr, D.B., Askenazi, M., Fuhrman, S., Wen, X. and Somogyi, R. (1998) Cluster analysis and data visualization of large-scale gene expression data. *Pac. Symp. Biocomput.*, 42–53.
99. Somogyi, R. and Sniegoski, C.A. (1996) Modeling the complexity of genetic networks: understanding multigenic and pleiotropic regulation. *Complexity*, **1**, 45–63.
100. Chen, T., Filkov, V. and Skiena, S.S. (1999) Identifying gene regulatory networks from experimental data. In Istrail, S., Pevzner, P. and Waterman, M. (eds), *Recomb 99*. ACM Press, Washington, DC, pp. 94–103.
101. Karp, R.M., Stoughton, R. and Yeung, K.Y. (1999) Algorithms for choosing differential gene expression experiments. In Istrail, S., Pevzner, P. and Waterman, M. (eds), *Recomb 99*. ACM Press, Washington, DC, pp. 208–217.
102. Chien, C.T., Bartel, P.L., Sternglanz, R. and Fields, S. (1991) The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc. Natl Acad. Sci. USA*, **88**, 9578–9582.
103. Brazma, A., Jonassen, I., Vilo, J. and Ukkonen, E. (1998) Predicting gene regulatory elements in silico on a genomic scale. *Genome Res.*, **8**, 1202–1215.
104. Myers, T.G., Anderson, N.L., Waltham, M., Li, G., Buolamwini, J.K., Scudiero, D.A., Paull, K.D., Sausville, E.A. and Weinstein, J.N. (1997) A protein expression database for the molecular pharmacology of cancer. *Electrophoresis*, **18**, 647–653.
105. Anderson, L. and Seilhamer, J. (1997) A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis*, **18**, 533–537.
106. Weinstein, J.N., Myers, T.G., O'Connor, P.M., Friend, S.H., Fornace, A.J.Jr, Kohn, K.W., Fojo, T., Bates, S.E., Rubinstein, L.V., Anderson, N.L., Buolamwini, J.K., van Osdol, W.W., Monks, A.P., Scudiero, D.A., Sausville, E.A., Zaharevitz, D.W., Bunow, B., Viswanadhan, V.N., Johnson, G.S., Wittes, R.E. and Paull, K.D. (1997) An information-intensive approach to the molecular pharmacology of cancer. *Science*, **275**, 343–349.
107. O'Connor, P.M., Jackman, J., Bae, I., Myers, T.G., Fan, S., Mutoh, M., Scudiero, D.A., Monks, A., Sausville, E.A., Weinstein, J.N., Friend, S., Fornace, A.J.Jr and Kohn, K.W. (1997) Characterization of the p53 tumor suppressor pathway in cell lines of the National Cancer Institute anticancer drug screen and correlations with the growth-inhibitory potency of 123 anticancer agents. *Cancer Res.*, **57**, 4285–4300.
108. Wosikowski, K., Schuurhuis, D., Johnson, K., Paull, K.D., Myers, T.G., Weinstein, J.N. and Bates, S.E. (1997) Identification of epidermal growth factor receptor and c-erbB2 pathway inhibitors by correlation with gene expression patterns. *J. Natl Cancer Inst.*, **89**, 1505–1515.
109. Liang, S., Fuhrman, S. and Somogyi, R. (1998) Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac. Symp. Biocomput.*, 18–29.

APPENDIX

Identifying coordinated gene expression from perfect binary data

We start from a binary (0/+) multi-conditional gene expression matrix such as:

	lib0	lib1	lib2	lib3	lib4	lib5	lib6	lib7	lib8	lib9
Gene 0	+	+	+	+	+	0	0	0	0	0
Gene 1	0	0	0	0	0	+	+	+	+	+
Gene 2	0	+	0	+	0	0	+	0	+	0
Gene 3	+	0	+	0	+	0	0	+	0	0
Gene 4	0	0	+	0	0	+	0	+	0	+
Gene 5	+	+	+	+	+	0	0	0	0	0
Gene 6	0	0	0	0	0	+	+	+	+	+
Gene 7	+	0	+	0	+	0	0	+	0	0
Gene 8	+	+	0	+	+	+	+	0	+	+
Gene 9	0	+	0	+	0	+	+	0	+	+

From this matrix we can compute the frequency $p_{\mathrm{g,L}}$ for each gene to be detected or absent in the libraries lib0–lib9; here we have:

$$p_{0+} = p_{1+} = p_{2+} = \ldots = p_{9+} = p_{+} = 1/2$$
$$p_{00} = p_{10} = p_{20} = \ldots = p_{90} = p_{0} = 1/2 \qquad \mathbf{A1}$$

Thus, each gene exhibits a maximal variation in expression across the libraries and the data set is optimal for the detection of coordinated gene expression. We now proceed to the pairwise comparison of all rows (gene expression profiles). What distinguishes the following two pairs of expression profiles:

Gene 0	+	+	+	+	+	0	0	0	0	0
Gene 5	+	+	+	+	+	0	0	0	0	0

Gene 0	+	+	+	+	+	0	0	0	0	0
Gene 9	0	+	0	+	0	+	+	0	+	+

is the 'unexpected' high number (10) of coincidences of scores in the first case, compared with only three coincidences observed in the second case. Given **A1**, five random coincidences are expected by chance, on average. A number of coincidences >5 is a sign of correlated expression (<5 is a sign of anti-correlation).

The statistical significance of observing 1, 2, 3, …, 10 coincidences can be computed from a binomial model. Given p_{+}, the probability for any gene being detected in any library, and p_{0}, the probability for any gene being absent in any library, the probability p_{c} of chance coincidence (+/+ or 0/0) per library is:

$$p_{\mathrm{c}} = p_{+} \cdot p_{+} + p_{0} \cdot p_{0}$$

and the complementary probability p_{nc} of not observing a coincidence is:

$$p_{nc} = (1 - p_{\mathrm{c}})$$

For any pair of genes, the random probability of getting exactly:

10 coincidences is	$P_{10} = p_{\mathrm{c}}^{10}$
9 coincidences is	$P_{9} = {}_{10}C_{9}\, p_{\mathrm{c}}^{9} \cdot p_{nc}$, with ${}_{10}C_{9} = 10!/[9!(10 - 9)!]$
8 coincidences is	$P_{8} = {}_{10}C_{8}\, p_{\mathrm{c}}^{8} \cdot p_{nc}^{2}$
7 coincidences is	$P_{7} = {}_{10}C_{7}\, p_{\mathrm{c}}^{7} \cdot p_{nc}^{3}$
6 coincidences is	$P_{6} = {}_{10}C_{6}\, p_{\mathrm{c}}^{6} \cdot p_{nc}^{4}$
5 coincidences is	$P_{5} = {}_{10}C_{5}\, p_{\mathrm{c}}^{5} \cdot p_{nc}^{5}$
etc.	

From these, we can compute the *cumulative* probabilities of observing:

10 coincidences:	$P_{10,}$
at least 9 coincidences:	$P_9 + P_{10}$,
at least 8 coincidences:	$P_8 + P_9 + P_{10},\ldots$

and so forth to the probability of observing at least 0 coincidence: $P_0 + P_1 + \ldots + P_{10} = 1$.

Given our perfect data set ($p_c = p_{nc} = 1/2$), the numerical values are:

$P_0 = P_{10} = (1/2)^{10} = 1/1024 =$	0.00097
$P_9 = P_1 = (1+ 10)/1024) =$	0.0107
$P_8 = P_2 = (45 + 10 + 1)/1024 =$	0.055
$P_7 = P_3 = (120 + 45 + 10 + 1)/1024 =$	0.17
$P_6 = P_4 = (210 + 120 + 45 + 10 + 1)/1024 =$	0.377
$P_5 = (252 + 210 + 120 + 45 + 10 + 1)/1024 =$	0.62

At the 5% significance level, observing >8 coincidences between the expression profiles of two genes is thus indicative of correlated behaviour, while observing <2 coincidences is a sign of an anti-correlation.

The feasibility of a whole genome expression correlation analysis

The most significant evidence for correlated expression is achieved for 10 coincidences and is associated with a p value of $(1/2)^{10}$. In general, for any binary (+/0) multi-conditional gene expression experiment involving L (independent) conditions and equal proportions of + and 0 scores, the most significant pairwise correlation will be associated with a p value of $p \approx (1/2)^L$. However, we are looking for any possible association among N different genes and thus embarking on what statisticians call a 'fishing trip'. On this fishing trip we will be trying to hook a significant result $N(N-1)/2$ times. We must thus expect a number of false positive pairwise 'correlations' of the order of:

$$N(N-1)/2\ 1/2^L$$

To ensure a reliable identification of coordinated expressed genes among the N tested, we must thus impose the constraint:

$$N(N-1)/2\ 1/2^L << 1 \qquad \textbf{A2}$$

This equation establishes a direct relationship between the maximal number N of genes one can analyse simultaneously and the minimal number L of independent expression conditions required to design a reliable study. From **A2** it follows that the parallel monitoring of 100 000 human genes using a binary detection system and a perfectly balanced data set would require expression measurements for ~35 conditions.

The situation is more favourable if our detection system can discriminate c (>2) expression levels. If, for the sake of simplicity, we consider such levels equiprobable, the probability of a score coincidence becomes $p_c = 1/c$. We then can rewrite the previous constraint as a three-way relationship between the number N of genes, the number L of conditions and the parameter c characterizing both the dynamic range and the accuracy of gene expression measurement:

$$N(N-1)/2\ 1/c^L << 1 \text{ or, approximately, } N << \sqrt{2c^L}$$

Although this simple formula has been obtained using drastic assumptions, it already indicates that, given the dynamic range and accuracy of the detection technologies at hand, simultaneous parallel monitoring for pairwise correlations of all human genes (N ≈100 000) is indeed feasible using a relatively small number of independent expression experiments. This result is reminiscent of the rather surprisingly small number of radiation hybrids required to map all human genes.