Minireview

Gene expression data analysis

Alvis Brazma*, Jaak Vilo

European Molecular Biology Laboratory, Outstation Hinxton – the European Bioinformatics Institute, Cambridge CB10 1SD, UK

Received 5 June 2000

Edited by Gianni Cesareni

Abstract Microarrays are one of the latest breakthroughs in experimental molecular biology, which allow monitoring of gene expression for tens of thousands of genes in parallel and are already producing huge amounts of valuable data. Analysis and handling of such data is becoming one of the major bottlenecks in the utilization of the technology. The raw microarray data are images, which have to be transformed into gene expression matrices – tables where rows represent genes, columns represent various samples such as tissues or experimental conditions, and numbers in each cell characterize the expression level of the particular gene in the particular sample. These matrices have to be analyzed further, if any knowledge about the underlying biological processes is to be extracted. In this paper we concentrate on discussing bioinformatics methods used for such analysis. We briefly discuss supervised and unsupervised data analysis and its applications, such as predicting gene function classes and cancer classification. Then we discuss how the gene expression matrix can be used to predict putative regulatory signals in the genome sequences. In conclusion we discuss some possible future directions. © 2000 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.

Key words: Microarray; Gene expression; Promoter sequence; Pattern discovery; Clustering

1. Introduction

With several eukaryotic genomes completed and the draft human genome published, we are now entering the postgenomic age. The main focus in genomic research is switching from sequencing to using the genome sequences in order to understand how genomes are functioning. Some questions we would like to ask are

- what are the functional roles of different genes and in what cellular processes do they participate;
- how are genes regulated, how do genes and gene products interact, what are these interaction networks;
- how does gene expression level differ in various cell types and states, how is gene expression changed by various diseases or compound treatments.

Knowing the gene transcript abundance in various tissues, developmental stages and under various conditions is important for attacking these questions. Although mRNA is not the ultimate product of a gene, transcription is the first step in gene regulation, and information about the transcript levels is needed for understanding gene regulatory networks. Moreover, the measurement of mRNA levels currently is considerably cheaper and can be done in a more high-throughput way than direct measurements of the protein levels. The correlation between the mRNA and protein abundance in the cell may not be straightforward, nevertheless the absence of mRNA in a cell is likely to imply a not very high level of the respective protein and thus at least qualitative estimates about the proteome can be based on the transcriptome information. The mRNA and protein level correlation studies are under way (see [1]).

The ability to monitor gene expression at the transcript level has become possible due to the advent of DNA microarray technologies (see [2]). A microarray is a glass slide, onto which single-stranded DNA molecules are attached at fixed locations (spots). There may be tens of thousands of spots on an array, each related to a single gene. Microarrays exploit the preferential binding of complementary single-stranded nucleic acid sequences. There are several variations of microarray technologies each used in a specific way.

One of the most popular experimental platforms is used for comparing mRNA abundance in two different samples (or a sample and a control). RNA from the sample and control cells are extracted and labeled with two different fluorescent labels, e.g. a red dye for the RNA from the sample population and a green dye for that from the control population. Both extracts are washed over the microarray. Gene sequences from the extracts hybridize to their complementary sequences in the spots.

To measure the relative abundance of the hybridized RNA the array is excited by a laser. If the RNA from the sample population is in abundance, the spot will be red, if the RNA from the control population is in abundance, it will be green. If sample and control bind equally, the spot will be yellow, while if neither binds, it will not fluoresce and appear black. Thus, from the fluorescence intensities and colors for each spot, the relative expression levels of the genes in the sample and control populations can be estimated.

By measuring transcription levels of genes in an organism under various conditions, at different developmental stages and in different tissues, we can build up 'gene expression profiles' which characterize the dynamic functioning of each gene in the genome. We can imagine the expression data represented in a matrix with rows representing genes, columns representing samples (e.g. various tissues, developmental stages and treatments), and each cell containing a number characterizing the expression level of the particular gene in the particular sample. We will call such a table a *gene expres-*

*Corresponding author. Fax: +44 1223 494468.
E-mail: brazma@ebi.ac.uk; vilo@ebi.ac.uk

0014-5793/00/$20.00 © 2000 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.
PII: S0014-5793(00)01772-5

sion matrix. Building up a database of such matrices will help us to understand gene regulation, metabolic and signaling pathways, the genetic mechanisms of disease, and the response to drug treatments. For instance, if overexpression of certain genes is correlated with a certain cancer, we can explore which other conditions affect the expression of these genes and which other genes have similar expression profiles. We can also investigate which compounds (potential drugs) lower the expression level of these genes.

2. From raw data to gene expression matrix

Like many experimental technologies, microarrays measure the target quantity (i.e. relative or absolute mRNA abundance) indirectly by measuring another physical quantity – the intensity of the fluorescence of the spots on the array for each fluorescent dye, i.e. for each optical wavelength (so-called channel). Therefore the raw data produced by microarrays are in fact monochrome images (Fig. 1). Transforming these images into the gene expression matrix is a non-trivial process: the spots corresponding to genes on the microarray should be identified, their boundaries determined, the fluorescence intensity from each spot measured and compared to the background intensity and to these intensities for other channels. The software for this initial image processing is often provided with the image scanner, since it will depend on particular properties of the hardware. Often laborious manual adjustment of the grid for spots is used. We will not discuss the raw data processing in detail in this paper, some survey of image analysis software can be found on http://cmpteam4.unil.ch/biocomputing/array/software/MicroArray_Software.html.

In any physical experiment it is important to know not only the value of the measurement, but also the standard error or

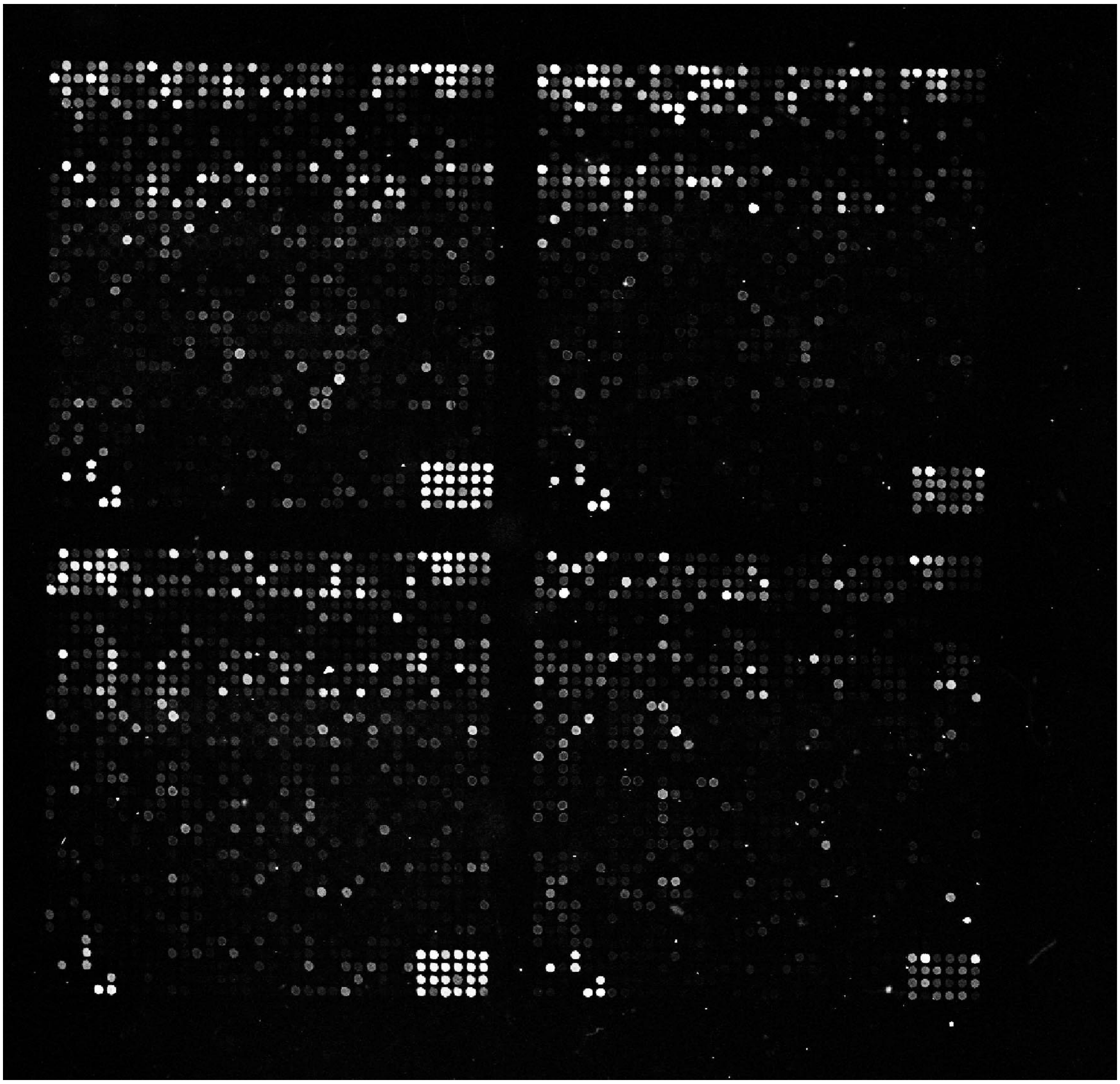

Fig. 1. A sample image from scanning a hybridized rat microarray containing over 5000 genes. Each spot features a pool of identical single-stranded DNA molecules representing a single gene. The brightness of the spot is proportional to the amount of fluorescent mRNA hybridized to the DNA of the spot. Automated image analysis software should identify these fluorescence spots, determine their boundaries, and the fluorescence intensity from each spot should be measured and compared to the background fluorescence. Moreover, the image should be compared to a similar image obtained from the control measurements and the ratio of background-subtracted intensities calculated. In this way images are transformed into a gene expression matrix, which can be analyzed further by numerical methods. The image was kindly provided by Tom Freeman (Sanger Centre, Cambridge, UK).

some other indicator of reliability for each data point. For most microarray technology platforms only the ratio of the background-subtracted signals of the given sample and the control is meaningful. If the spot intensity is low, the ratio of these numbers may be high, but the measurement may not be reliable. The spot quality can be assessed not only by the absolute intensity in each channel, but also by many other factors, such as uniformity of the individual pixel intensities, or the shape of the spot. Unfortunately there is currently no standard way of assessing the spot measurement reliability. If experiments have been done in replicates, they can be used to assess the standard errors in addition to the single measurement quality assessments. Little has been published yet on how to use the reliability of gene expression measurements by combining the information about the spot image in each channel and the replicate images.

Another difficulty in creating a gene expression matrix comes from the necessity to identify each spot with the respective gene. This is not always possible, since spots are typically based on EST sequences, and linking the EST to the respective gene may be non-trivial. Typically it is done through EST clustering. Additionally, the same gene may be represented by several spots on the array, either by exactly the same or by a different sequence. What expression level to attribute to the gene, if measurements from these different spots differ?

Microarray-based gene expression measurements are still far from giving estimates of mRNA counts per cell in the sample. The measurements are relative by nature: essentially we can compare the expression level either of the same gene in different samples, or of different genes in the same sample. Moreover, appropriate normalization should be applied to enable any data comparisons. Typically it is assumed that abundance ratios of 1.5–2 are indicative of a change in gene expression, but such estimates are very crude. The reliability of ratios depends on the absolute intensity values, as well as varying from spot to spot due to specificity of the sequence and cross-hybridization of homologous sequences (for instance see [3]). This should be kept in mind while analyzing the gene expression matrix. The value of microarray-based gene expression measurements would be considerably higher if reliability and limitations of particular microarray platforms for particular kinds of measurements, as well as cross-platform comparison and normalization, were studied and published.

After we have processed the raw image data into the gene expression matrix, the next task is to analyze this matrix and to try to extract from it some knowledge about the underlying biological processes.

3. Gene expression matrix analysis

There are two straightforward ways how gene expression matrix can be studied:

1. comparing expression profiles of genes by comparing rows in the expression matrix;
2. comparing expression profiles of samples by comparing columns in the matrix.

Additionally both methods can be combined (provided that the data normalization allows it). When comparing rows or columns, we can look either for similarities or for differences. If we find that two rows are similar, we can hypothesize that the respective genes are co-regulated and possibly functionally related. By comparing samples, we can find which genes are differentially expressed and, for instance, study effects of various compounds.

Before we can perform any comparisons, we need a way to measure the similarity (or distance) between the objects we are comparing. We can regard these objects (rows or columns in the matrix) as points in n-dimensional space or as n-dimensional vectors, where n is the number of samples for gene comparison, or number of genes for sample comparison. The natural, so-called Euclidean distance (for definition see [4]) between these points in the n-dimensional space may be the most obvious, but not necessarily the best choice. It is intuitively appealing to use the correlation coefficient calculated by treating the two n-dimensional vectors as series of random variables. In fact this distance is related to the angle between the two n-dimensional vectors. Euclidean and correlation distance measures are related, if we normalize the length of the n-dimensional vectors to 1. This makes it possible to use correlation distance even in the cases when Euclidean properties are important. Some other distance measures, including rank correlation coefficient and mutual information-based measure, are proposed in D'haesleer et al. [5]. Currently, to the best of our knowledge, there is no theory how to choose the best distance measure. Possibly one 'right' distance measure in the expression profile space does not exist, and the choice should depend on the questions that we are asking. Standard sets of known co-regulated genes in various organisms and gene regulatory network modeling can potentially help in finding theoretically substantiated similarity measures.

After having chosen the similarity measure in the expression profile space we can study the expression matrix in either a supervised or an unsupervised manner. The supervised approach assumes that for some (or all) profiles we have additional information, such as functional classes for the genes, or diseased/normal states attributed to the samples. We can view this additional information as labels attached to the rows or columns. Having this information, a typical task is to build a classifier able to predict the labels from the expression profile. A typical example of unsupervised data analysis is expression profile clustering to find groups of co-regulated genes or related samples. For conceptual illustration of unsupervised and supervised analysis see Fig. 2. First we discuss the clustering approach.

3.1. Unsupervised analysis

The goal of clustering is to group together objects (genes or samples) with similar properties. This can also be viewed as the reduction of the dimensionality of the system. Clustering is not a new technique, many algorithms have been developed for it and many of these algorithms have been applied to analyze expression data. The hierarchical [6] and K-mean clustering algorithms [7,8] as well as self-organizing maps [9] have all been used for clustering expression profiles. Even a simple clustering algorithm based on binning (i.e. discretizing the expression profile space and clustering together the profiles that map into the same bin) has been shown to be useful for clustering genes and subsequent discovering of transcription factor binding sites [10]. More recently new algorithms

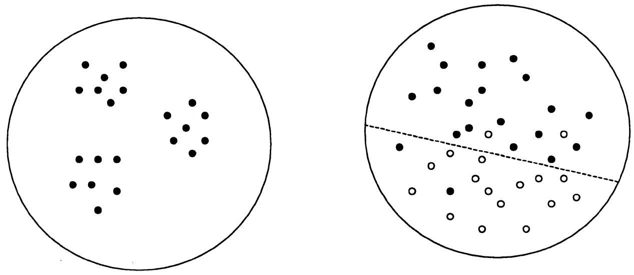

Fig. 2. Supervised and unsupervised data analysis. In the unsupervised case (left) we are given data points in n-dimensional space ($n=2$ in the example) and we are trying to find ways how to groups together points with similar features. For instance, there are three natural clusters in the example, each consisting of data points close to each other in a sense of Euclidean distance. A clustering algorithm should identify these clusters. In the supervised case (right), the objects are labelled (e.g. we have filled and unfilled points in the example), and the task is to find a set of classification rules allowing us to discriminate between these points as precisely as possible. For instance, the dotted line in the drawing discriminates most of the points correctly, allowing us to predict their 'labels' – filled or unfilled – by their position above or below the dotted line.

have been developed specifically for gene expression profile clustering, for instance based on finding approximate cliques in graphs [11].

Hierarchical clustering works by iteratively joining the two closest clusters starting from singleton clusters [6] or iteratively partitioning clusters starting with the complete set [12], see Fig. 3. After each joining of two clusters, the distances between all the other clusters and a new joined cluster are recalculated. The complete linkage, average linkage, and single linkage methods use maximum, average, and minimum distances between the members of two clusters respectively. Note that to obtain a particular partitioning into clusters, the threshold distance should be chosen by independent means (typically by the user himself).

The K-means clustering algorithm typically uses the Euclidean properties of the vector space. The desired number of clusters K has to be chosen a priori. After the initial partitioning of the vector space into K parts, the algorithm calculates the center points in each subspace and adjusts the partition so that each vector is assigned to the cluster the center of which is the closest. This is repeated iteratively until either the partitioning stabilizes or the given number of iterations is exceeded. The approaches for the initial selection of the first set of K cluster centers can vary.

Clustering of expression profiles has been used for grouping genes as well as samples. The clustering of genes for finding co-regulated and functionally related groups is particularly interesting in the cases when we have complete sets of an organism's genes. In a frequently quoted paper DeRisi et al. [13] used a DNA array containing a complete set of yeast genes to study the diauxic shift time course. They selected small groups of genes with similar expression profiles and showed that these genes are functionally related and contain relevant transcription factor binding sites upstream of their open reading frames (ORFs). More systematic studies of this dataset for regulatory elements were done by Brazma et al. [10] and van Helden et al. [14].

Later more expression studies of yeast under various conditions were carried out, including sporulation [15], cell cycle [16] and yeast gene regulatory machinery [17]. Clustering has been applied to the obtained gene expression matrices, and groups of functionally related and co-regulated genes have been revealed. Tavazoie et al. [8] clustered expression profiles of 3000 most variable yeast genes during the cell cycle (15 time points, data from Cho et al. [18]) into 30 clusters by the K-means algorithm. They found that for half of these clusters, strong sequence patterns are present in the gene upstream sequences. Note that expression profiles of cell cycle-dependent genes are periodic and Fourier analysis has been used to discover these genes [16].

Eisen et al. [6] have developed a hierarchical clustering-based algorithm and visualization software package, which is currently one of the most frequently used tools for expression profile clustering and data visualization. They applied their software to gene expression matrices obtained by combining 80 different yeast samples (experimental conditions) studied in various hybridization experiments at Stanford University (including the ones mentioned above).

Gene expression profile clustering does not necessarily require the full genome. For instance Iyer et al. [19] studied 8600 genes in human fibroblasts and obtained 10 distinct gene clusters each associated with genes with particular functional roles, such as signal transduction, coagulation, hemostasis, inflammation etc.

A simple method for finding sets of interesting genes is comparing expression profiles of two or more samples for differentially expressed genes. For instance, Lee et al. [20] used this method to find genes that are differentially expressed in skeletal muscle of adult (5 months) and old (30 months) mice. Of over 6347 mouse genes surveyed by a microarray, 58 displayed a greater than two-fold increase, whereas 55 displayed a greater than two-fold decrease in expression in the skeletal muscles of the old mice. Of the genes that increased in expression, 16% were mediators or stress response genes and 9% were involved in neuronal growth. Of genes that decreased in expression, 13% were participating in energy metabolism. In the same study gene expression profiles from 30 month old mice with restricted caloric intake (76% of that of a control population) were compared to the 30 month old control population, and it was shown that the expression profile of restricted caloric intake mice was closer to that of younger mice.

Hierarchical clustering [6] has also been used for sample clustering. An interesting application of this approach is the clustering of tumors to find new possible tumor subclasses. In a recent paper by Alizadeh et al. [21], diffuse large B-cell lymphoma (DLBCL) was studied using 96 samples of normal and malignant lymphocytes. Applying a hierarchical clustering algorithm to these samples they showed that there is a diversity in gene expression among the tumors of DLBCL patients. They identified two molecularly distinct forms of DLBCL, which had gene expression patterns indicative of different stages of B-cell differentiation. Interestingly, these two groups correlated well with patient survival rates, thus confirming that the clusters are meaningful.

Sample clustering has been combined with gene clustering to identify which genes are the most important for sample clustering [12,21]. Alon et al. [12] have applied a partitioning-based clustering algorithm to study 6500 genes of 40 tumor and 22 normal colon tissues for clustering both genes and samples. They call this method two-way clustering.

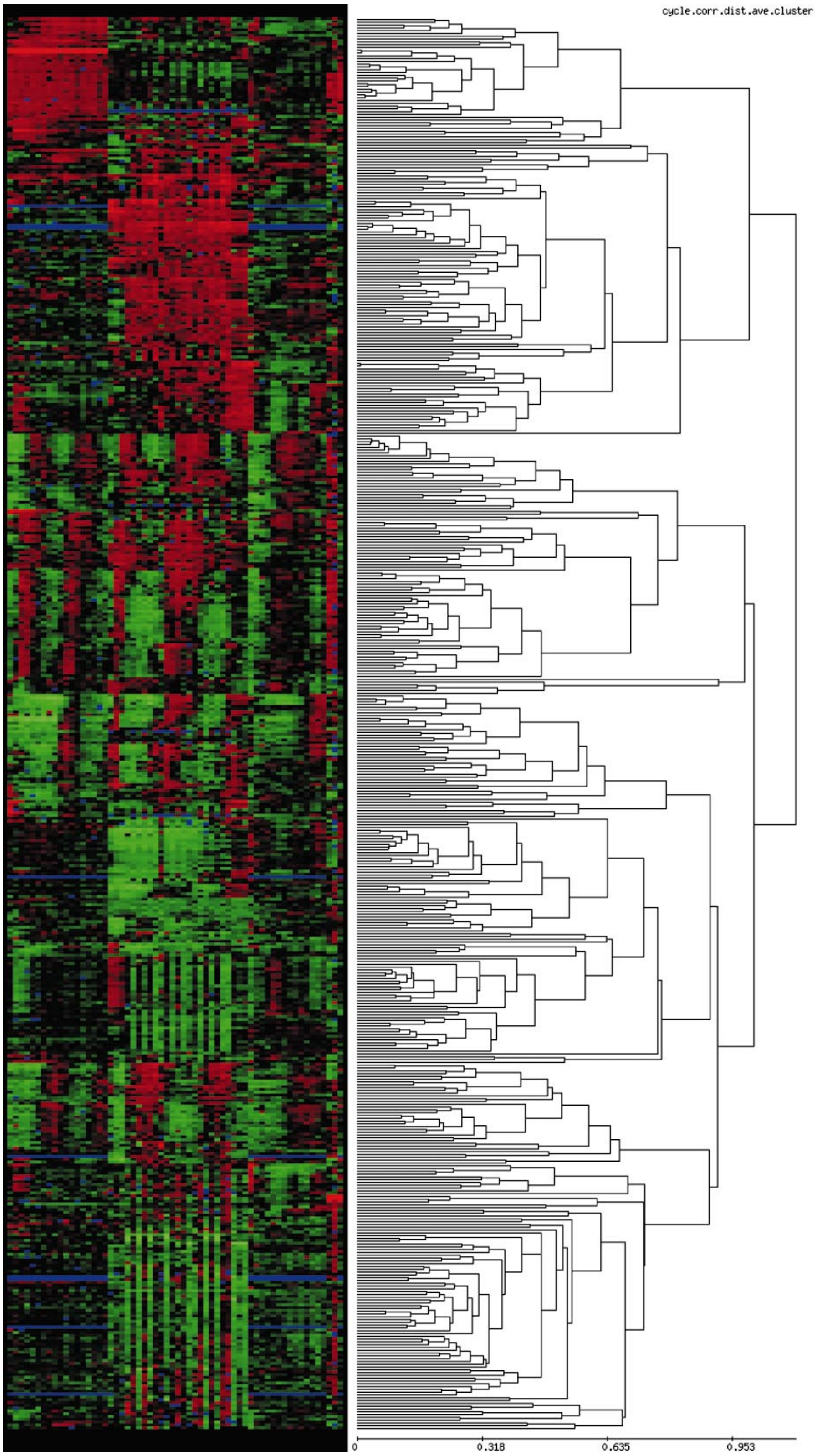

Fig. 3. Hierarchical clustering of gene expression matrices. The image shows an average linkage (UPGMA) clustering of 505 yeast genes during three different cell cycle studies with a total of 60 different time points analyzed. The color image on the left shows the numerical values encoded by color according to the method introduced by Mike Eisen. Red is used to represent the positive values and green the negative values. Blue shows the missing values in the respective experiments. The clustering and the image are produced using WWW-based tools in Expression Profiler (http://www.ebi.ac.uk/microarray/). The interface is interactive and further information about the genes in each subtree is available by clicking on the respective nodes in the tree.

3.2. Supervised analysis

One of the goals of supervised expression data analysis is to construct classifiers, such as linear discriminants, decision trees or support vector machines (SVM), which assign predefined classes to a given expression profile. For instance, if a classifier can be constructed based on gene expression profiles that is able to distinguish between two different, but morphologically closely related tumor tissues, such a classifier can be used for diagnostics. Moreover, if such a classifier is based on a set of relatively simple rules, it can help to understand what the mechanisms involved in each tumor are. Typically, such classifiers are trained on a subset of data with a priori given classification and tested on another subset with known classification. After assessing the quality of the prediction they can be applied to data the classification of which is unknown.

Brown et al. [22] have applied various supervised learning algorithms to six functional classes of yeast genes using gene expression matrices from 79 samples [6]. Genes from some of the classes, such as ribosomal proteins and histones, are expected to be co-expressed. For these classes a good classification accuracy was achieved. Some other functional classes, such as protein kinases, are not expected to have distinct gene expression profiles. It was shown that SVM provides the best prediction accuracy for the functional classes that are expected to be co-regulated.

Golub et al. [23] applied neighborhood analysis to construct class predictors for samples, concretely for leukemias. They were looking for genes the expression of which is best correlated with two known classes of leukemias, acute myeloid leukemia and acute lymphoblastic leukemia. They constructed a classifier based on 50 genes (out of 6817) using 38 samples and applied it to a collection of 34 new samples. The classifier correctly predicted 29 of these 34 samples.

Note that when classifying samples, we are confronted with a problem that there are many more attributes (genes) than objects (samples) that we are trying to classify. This makes it always possible to find a perfect discriminator if we are not careful in restricting the complexity of the permitted classifiers. To avoid this problem we must look for very simple classifiers, compromising between simplicity and classification accuracy. Ben-Dor et al. [24] applied a new clustering algorithm for classification of colon and ovarian cancer data sets. They used unsupervised clustering to find a hierarchical structure in the expression profile space, and supervised learning to find the best threshold to correlate the clustering structure with the known cancer classes.

Whether we use supervised or unsupervised expression profile analysis, they are only the first steps in expression data analysis. It is a long way from finding gene clusters to finding functional roles of the respective genes, and moreover, understanding the underlying biological processes. A natural step downstream of expression profile clustering is the usage of putative promoter sequences of similarly expressed genes for finding regulatory sequence elements in genomes. This is easier for yeast, since typically yeast promoters are relatively close to ORFs. In the next section we describe an approach which uses gene expression data to find regulatory sequence elements in yeast.

4. Identification of putative regulatory signals

It seems reasonable to hypothesize that genes with similar expression profiles, i.e. genes that are co-expressed, may have something in common in their regulatory mechanisms, i.e. may be co-regulated. Therefore by clustering together genes with similar expression profiles one can find groups of potentially co-regulated genes and search for putative regulatory signals. The outline of such a discovery method is as follows:

1. cluster the genes based on a selection of expression measurements;
2. extract putative promoter sequences for the genes in the clusters;
3. search for sequence patterns overrepresented in these clusters;
4. assess the quality of discovered patterns using some statistical significance criteria.

A systematic application of this approach has been reported for the yeast *Saccharomyces cerevisiae* using a public data set from Stanford University [6] combining various yeast expression experiments with a total of 80 conditions for 6221 genes (http://rana.stanford.edu/). The computational analysis consisted of the following steps [25].

1. Clustering the expression data. In the absence of theoretically 'correct' similarity measures and clustering algorithms, the simplest measure was selected and different clusterings carried out. All genes were clustered based on their expression profiles by the *K*-means clustering algorithm using Euclidean distances. Instead of fixing the number of clusters *K* it was varied between 2 and 1000. For each *K* the clustering was repeated 10 times with different random initial cluster centers. In total over 900 separate clusterings were made and clusters of size between 20 and 100 genes were selected, totaling over 52 100 different clusters.
2. Sequence pattern discovery. For each cluster the set of gene upstream sequences of length 600 bp was taken for analysis. All substring patterns of unrestricted length occurring in at least 10 sequences in a cluster were scored according to the binomial probability of their occurrence in the cluster. The background probability was estimated based on the number of occurrences of each pattern in upstream sequences of all 6221 genes.
3. Finding the significance threshold by control experiment. To determine the statistical significance threshold for the patterns, step 2 was repeated on randomized data by replacing the cluster contents by upstream sequences from random sets of genes. A threshold probability of 10^{-8} was chosen as patterns with higher probability were also observable from random clusters.
4. Pattern selection. Of the over 6000 significant patterns many were observed to occur in clusters of genes with high homology in the respective upstream sequences. These clusters, totaling 169 genes, were easily identifiable and they were removed. The remaining clusters of genes with non-homologous upstream sequences contained 3727 ORFs and together they produced 1498 significant patterns.
5. Grouping the patterns. As 1498 substring patterns is still too many for human study, they were clustered using a similarity measure based on common information content [26]. This produced 62 clusters of similar patterns. For each

cluster of patterns an approximate alignment and a consensus pattern were calculated.

6. Evaluation of discovered patterns against known transcription factor binding sites. All 1498 interesting patterns were matched against experimentally verified DNA binding sites of yeast as given in SCPD ([27], http://cgsigma.cshl.org/jian/).

Of the 62 clusters of patterns 48 had matches in SCPD and 14 were such that they did not have a match in any site reported in the SCPD database. Table 1 shows the partial consensus patterns that were calculated from pattern alignments for these 14 clusters. The nucleotide groups (IUPAC groups represented here using a regular expression notation) were introduced when the frequency of the less frequent nucleotide in the respective column was over 25% of the frequency of the more frequent nucleotide. Inside the groups nucleotides are ordered based on their frequency. Lowercase letters are used when the majority of the patterns do not have any nucleotide in that position, i.e. when the most frequent nucleotide in the respective alignment column is a dash.

The fact that 48 out of 62 pattern classes have matches in experimentally verified yeast transcription factor binding sites indicates the validity of the described computational discovery method. Potentially the most interesting patterns, however, are the ones that do not have matches in the known binding sites, and they can be targets for further research (see Table 1). In this way, the described computational experiment has come up with targets for further research by more conventional methods. Automatic or semiautomatic generation of such hypotheses is one of the main tasks of bioinformatics and data mining approaches.

The tools used for the experiments outlined above, as well as the complete results of the experiments, are available online (http://www.ebi.ac.uk/microarray). All the tools, including the clustering and visualization methods for expression data analysis and the regulatory region extraction for the yeast, have a web interface. The individual tools are interconnected so that similar analyses can be carried out over the web for any expression and sequence data.

5. Conclusions

Expression data analysis methods are currently only in their infancy. Even the rather obvious approaches, such as cluster analysis and finding differentially expressed genes, have been used only rather crudely. For instance, the appropriateness of similarity measures has not been systematically explored and these measures are used on an ad-hoc basis. The information characterizing the measurement quality of different data points is typically not used. Advances in this area are hindered by the lack of systematic research in ways of assessing the measurement quality and comparing data from various technology platforms. These shortcomings can be overcome only if the journals encourage publications exploring the gene expression measurement technologies themselves, rather than always concentrating on the biological subject. In the long run the advancement of biological knowledge will be accelerated by technology-centric studies, with biology becoming more quantitative science.

Table 1
Consensus sequences of the pattern clusters that do not have matches in the SCPD database

Cluster	Consensus pattern
2	aaTCTTCATGt
5	cgTACCTCTa
8	gACAGCTAc
17	tAT[TAC]GTTAAgc
20	ACTTTATTT
21	[ag]TAACTT[AT]Ca
26	TATCGAG (singleton)
29	t[ta]CGAATA[AG]aaaa
42	[ta]TGCATGAAc
43	a[TG][GC]GTATAc
45	[ag][ga][AG]ATATG[TG][ga][ag]g
46	tag[AG]TAGA[TA]A[ga]aaaa
50	ATCCAAGAg
59	tTTTTCTG[CT][TA]c

See text for explanations.

Gene expression data analysis methods will develop similarly as sequence analysis methods have developed over the past decades. The amounts of gene expression data will continue growing and the data will become more systematic. Currently the gene expression profiling is similar to gene sequencing before the era of genome sequencing: the measurements are carried out to attack particular questions or sometimes just to demonstrate the concept.

With the technology becoming more reliable, with the introduction of standard controls in experiments and developing generally accepted data normalization and quality control methods, it will become possible to systematically profile genes in various organisms, tissues, developmental stages and conditions. Various chemical compounds will be profiled for their possible toxicity and other effects on organisms, and various signatures will be associated with various toxicity mechanisms or cellular processes. This approach will resemble systematic genome sequencing. Algorithms for reliable searching of similar expression profiles, or analyzing sets of related profiles to discover common signatures, will be needed, just as searching and pattern discovery algorithms are needed to explore sequences.

However, there is a major difference between gene sequence and expression data. Even if eventually we are able to overcome various technological limitations, and even if we are able to measure gene expression in terms of absolute units such as mRNA counts, the gene expression profiles are meaningful only in the context of the experimental conditions in which they have been measured. This requires detailed and systematic annotation of samples and experimental conditions. For this to become a reality, agreed ontologies and controlled vocabularies for tissues, cell types, and treatments, as well as for array designs, image analyses and hybridization protocols, have to be developed. Systematic building up of gene expression matrices for various organisms would be facilitated by establishing a public repository for gene expression data [28].

Like genome sequencing, the systematic gene expression profile is not an end in itself. It is a long way from having detailed gene expression profiles to real understanding of underlying cellular processes. Bioinformatics methods and tools will be needed to cope with the huge amounts of data, but they will not bring any deep understanding by themselves. On the other hand, the traditional 'gene by gene' methods will not be sufficient to understand gene regulatory networks consisting of thousands or tens of thousands of genes. One of the most challenging downstream goals of gene expression profiling and data analysis is the reverse engineering and modeling of gene regulatory networks (see for instance [29–31]). With biology becoming more quantitative science, modeling approaches will become more and more usual.

References

[1] Celis, J.E., Kruhøffer, M., Gromova, I., Frederiksen, C., Østergaard, M., Thykjaer, T., Gromov, P., Yu, Y., Pálsdóttir, H. and Ørntoft, T.F. (2000) FEBS Lett. 480, 2–16.

[2] The Chipping Forecast (1999) Nature Genet. 21, Suppl.

[3] Claverie, J.-M. (1999) Hum. Mol. Genet. 8, 1821–1832.

[4] Legendre, P. and Legendre, L. (1998) Numerical Ecology. Developments in Environmental Modelling, Elsevier, Amsterdam.

[5] D'haesleer, P., Wen, X., Fuhrman, S. and Somogyi, R. (1998) in: Information Processing in Cells and Tissues, Plenum Press, New York.

[6] Eisen, M., Spellman, P.T., Botstein, D. and Brown, P.O. (1998) Proc. Natl. Acad. Sci. USA 95, 14863–14867.

[7] Hartigan, J.A. (1975) Clustering Algorithms, John Wiley and Sons, New York.

[8] Tavazoie, S., Hughes, D., Campbell, M.J., Cho, R.J. and Church, G.M. (1999) Nature Genet. 22, 281–285.

[9] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. and Golub, T. (1999) Proc. Natl. Acad. Sci. USA 96, 2907–2912.

[10] Brazma, A., Jonassen, I., Vilo, J. and Ukkonen, E. (1998) Genome Res. 8, 1202–1215.

[11] Ben-Dor, A. and Yakhini, Z. (1999) Proceedings of the Third Annual International Conference on Computational Molecular Biology RECOMB-1999, pp. 33–42. ACM Press, Lyon.

[12] Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D. and Levine, A.J. (1999) Proc. Natl. Acad. Sci. USA 96, 6745–6750.

[13] DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997) Science 278, 680–686.

[14] van Helden, J., André, B. and Collado-Vides, J. (1998) J. Mol. Biol. 281, 827–842.

[15] Chu, S., DeRisi, J.L., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O. and Herskowitz, I. (1998) Science 282, 699–705.

[16] Spellman, P.T., Sherlock, G., Zhang, M., Iyer, V.R., Anders, K., Eisen, M., Brown, P.O., Botstein, D. and Futcher, B. (1998) Mol. Biol. Cell 9, 3273.

[17] Holstege, F., Jennings, E., Wyrick, J., Lee, T., Hengartner, C., Green, M., Golub, T., Lander, E. and Young, R. (1998) Cell 95, 717–728.

[18] Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J. and Davis, R.W. (1998) Mol. Cell 2, 65–73.

[19] Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C.F., Trent, J.M., Staudt, L.M., Hudson Jr., J., Boguski, M.S., Lashkari, D., Shalon, D., Botstein, D. and Brown, P.O. (1999) Science 283, 83–87.

[20] Lee, C., Klopp, R.G., Weindruch, R. and Prolla, T.A. (1999) Science 285, 1390–1393.

[21] Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson Jr., J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O. and Staudt, L.M. (2000) Nature 403, 503–511.

[22] Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M.J. and Haussler, D. (2000) Proc. Natl. Acad. Sci. USA 97, 262–267.

[23] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D. and Lander, E.S. (1999) Science 286, 531–537.

[24] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, Z. (2000) The Fourth Annual International Conference on Computational Molecular Biology RECOMB-2000, pp. 54–64, ACM Press, Tokyo.

[25] Vilo, J., Brazma, A., Jonassen, I., Robinson, A. and Ukkonen, E. (2000) The Eighth International Conference on Intelligent Systems for Molecular Biology, AAAI Press, La Jolla, CA, in press.

[26] Hertz, G.Z. and Stormo, G.D. (1995) in: Proceedings of the Third International Conference on Bioinformatics and Genome Research, pp. 201–216, World Scientific Publishing, Singapore.

[27] Zhu, J. and Zhang, M.Q. (1999) Bioinformatics 15, 607–611.

[28] Brazma, A., Robinson, A., Cameron, G. and Ashburner, M. (2000) Nature 403, 699–700.

[29] Akutsu, T., Miyano, S. and Kuhara, S. (1999) The Pacific Symposium on Biocomputing '99 (PSB'99), pp. 17–28, World Scientific, Hawaii.

[30] Liang, S., Fuhrman, S. and Somogyi, R. (1998) The Pacific Symposium on Biocomputing, Vol. 3, pp. 18–29, World Scientific, Hawaii.

[31] Thieffry, D., Colet, M. and Thomas, R. (1993) Math. Model. Sci. Comput. 55, 144–151.