Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 14863–14868, December 1998
Genetics

Cluster analysis and display of genome-wide expression patterns

MICHAEL B. EISEN*, PAUL T. SPELLMAN*, PATRICK O. BROWN†, AND DAVID BOTSTEIN*‡

*Department of Genetics and †Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, 300 Pasteur Avenue, Stanford, CA 94305

Contributed by David Botstein, October 13, 1998

ABSTRACT **A system of cluster analysis for genome-wide expression data from DNA microarray hybridization is described that uses standard statistical algorithms to arrange genes according to similarity in pattern of gene expression. The output is displayed graphically, conveying the clustering and the underlying expression data simultaneously in a form intuitive for biologists. We have found in the budding yeast *Saccharomyces cerevisiae* that clustering gene expression data groups together efficiently genes of known similar function, and we find a similar tendency in human data. Thus patterns seen in genome-wide expression experiments can be interpreted as indications of the status of cellular processes. Also, coexpression of genes of known function with poorly characterized or novel genes may provide a simple means of gaining leads to the functions of many genes for which information is not available currently.**

The rapid advance of genome-scale sequencing has driven the development of methods to exploit this information by characterizing biological processes in new ways. The knowledge of the coding sequences of virtually every gene in an organism, for instance, invites development of technology to study the expression of all of them at once, because the study of gene expression of genes one by one has already provided a wealth of biological insight. To this end, a variety of techniques has evolved to monitor, rapidly and efficiently, transcript abundance for all of an organism's genes (1–3). Within the mass of numbers produced by these techniques, which amount to hundreds of data points for thousands or tens of thousands of genes, is an immense amount of biological information. In this paper we address the problem of analyzing and presenting information on this genomic scale.

A natural first step in extracting this information is to examine the extremes, e.g., genes with significant differential expression in two individual samples or in a time series after a given treatment. This simple technique can be extremely efficient, for example, in screens for potential tumor markers or drug targets. However, such analyses do not address the full potential of genome-scale experiments to alter our understanding of cellular biology by providing, through an inclusive analysis of the entire repertoire of transcripts, a continuing comprehensive window into the state of a cell as it goes through a biological process. What is needed instead is a holistic approach to analysis of genomic data that focuses on illuminating order in the entire set of observations, allowing biologists to develop an integrated understanding of the process being studied.

A natural basis for organizing gene expression data is to group together genes with similar patterns of expression. The first step to this end is to adopt a mathematical description of similarity. For any series of measurements, a number of sensible measures of similarity in the behavior of two genes can be used, such as the Euclidean distance, angle, or dot products of the two n-dimensional vectors representing a series of n measurements. We have found that the standard correlation coefficient (i.e., the dot product of two normalized vectors) conforms well to the intuitive biological notion of what it means for two genes to be "coexpressed;" this may be because this statistic captures similarity in "shape" but places no emphasis on the magnitude of the two series of measurements.

It is not the purpose of this paper to survey the various methods available to cluster genes on the basis of their expression patterns, but rather to illustrate how such methods can be useful to biologists in the analysis of gene expression data. We aim to use these methods to organize, but not to alter, tables containing primary data; we have thus used methods that can be reduced, in the end, to a reordering of lists of genes. Clustering methods can be divided into two general classes, designated supervised and unsupervised clustering (4). In supervised clustering, vectors are classified with respect to known reference vectors. In unsupervised clustering, no predefined reference vectors are used. As we have little *a priori* knowledge of the complete repertoire of expected gene expression patterns for any condition, we have favored unsupervised methods or hybrid (unsupervised followed by supervised) approaches.

Although various clustering methods can usefully organize tables of gene expression measurements, the resulting ordered but still massive collection of numbers remains difficult to assimilate. Therefore, we always combine clustering methods with a graphical representation of the primary data by representing each data point with a color that quantitatively and qualitatively reflects the original experimental observations. The end product is a representation of complex gene expression data that, through statistical organization and graphical display, allows biologists to assimilate and explore the data in a natural intuitive manner.

To illustrate this approach, we have applied pairwise average-linkage cluster analysis (5) to gene expression data collected in our laboratories. This method is a form of hierarchical clustering, familiar to most biologists through its application in sequence and phylogenetic analysis. Relationships among objects (genes) are represented by a tree whose branch lengths reflect the degree of similarity between the objects, as assessed by a pairwise similarity function such as that described above. In sequence comparison, these methods are used to infer the evolutionary history of sequences being compared. Whereas no such underlying tree exists for expression patterns of genes, such methods are useful in their ability to represent varying degrees of similarity and more distant relationships among groups of closely related genes, as well as in requiring few assumptions about the nature of the data. The computed trees can be used to order genes in the original data table, so that genes or groups of genes with similar expression patterns are adjacent. The ordered table can then be displayed graphically, as above, with a representation of the tree to indicate the relationships among genes.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

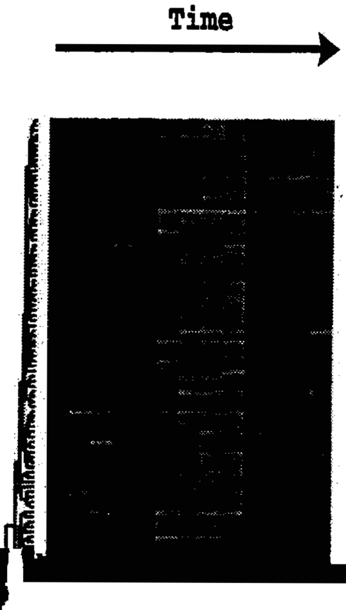
[redacted]

MATERIALS AND METHODS

Sources of Experimental Data. Data analyzed here were collected on spotted DNA microarrays (6, 7). Gene expression

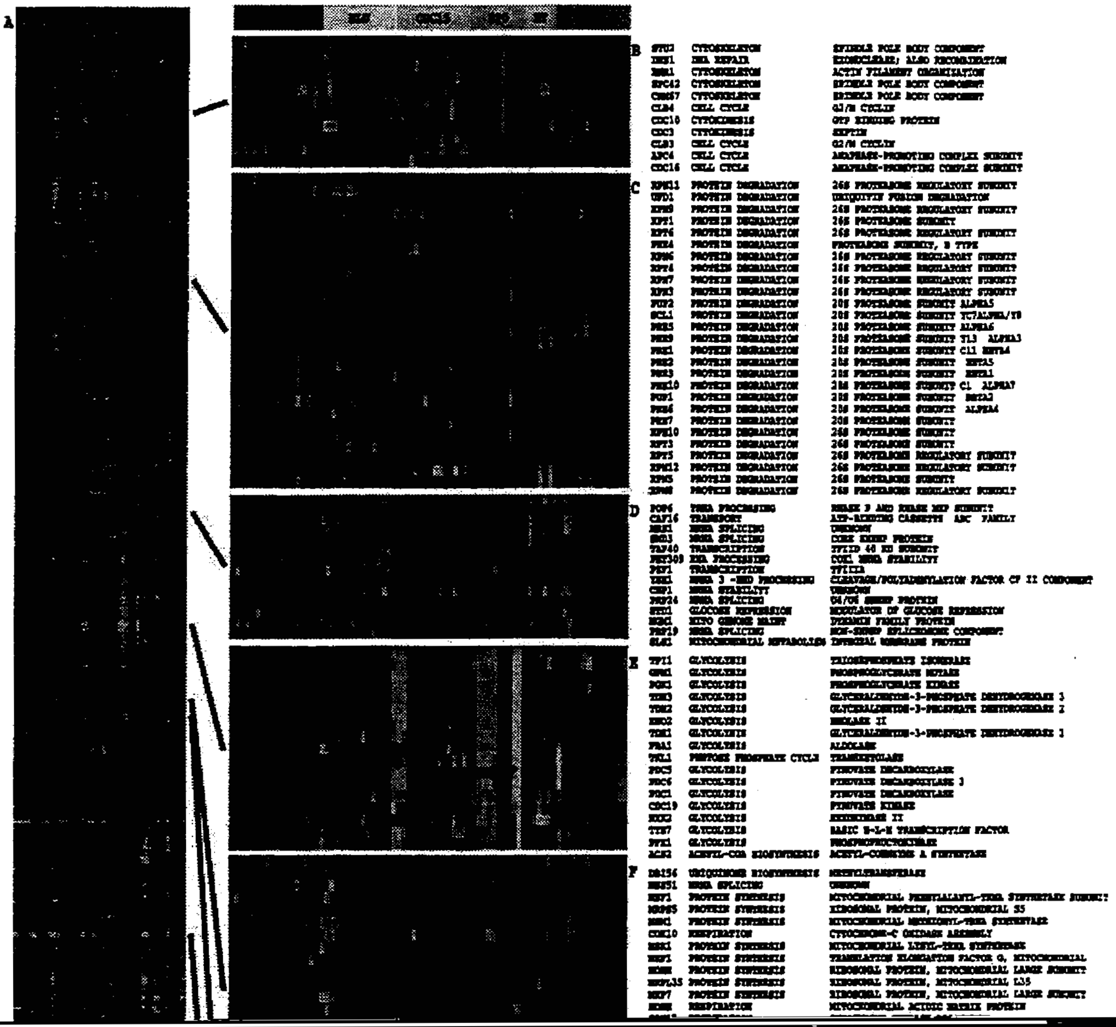
[redacted]

pairs of genes. The matrix is scanned to identify the highest value (representing the most similar pair of genes). A node is created joining these two genes, and a gene expression profile



regulation. Finally, this result also indicates that noise present in single observations does not contribute significantly when genes are compared across even a relatively small number of nonidentical conditions. Therefore, when designing experiments, it may be more valuable to sample a wide variety of conditions than to make repeat observations on identical conditions.

Genes of Similar Function Cluster Together. A far more striking result is found when larger groups of clustered genes are examined, where we observe a strong tendency for these genes to share common roles in cellular processes. This relationship is clearest in data from experiments on the budding yeast *S. cerevisiae*, where arrays representing essentially all of the genes from this organism are available (8) and for which a large fraction of the identified genes (more than 35%) have been studied in some detail. Fig. 2*A* represents a clustering analysis of 2,467 genes, all the genes that currently have a functional annotation in the *Saccharomyces* Genome Database (12). As can be seen in Fig. 2 *B–K*, numerous groups of coexpressed genes representing diverse expression patterns across the sampled conditions are involved in common cellular


A
CDC15
B
STU2 CYTOSKELETON SPINDLE POLE BODY COMPONENT
DIN7 DNA REPAIR EXONUCLEASE; ALSO RECOMBINATION
BNI1 CYTOSKELETON ACTIN FILAMENT ORGANIZATION
SPC42 CYTOSKELETON SPINDLE POLE BODY COMPONENT
CNM67 CYTOSKELETON SPINDLE POLE BODY COMPONENT
CLB4 CELL CYCLE G2/M CYCLIN
CDC10 CYTOKINESIS GTP BINDING PROTEIN
CDC3 CYTOKINESIS SEPTIN
CLB3 CELL CYCLE G2/M CYCLIN
APC4 CELL CYCLE ANAPHASE-PROMOTING COMPLEX SUBUNIT
CDC16 CELL CYCLE ANAPHASE-PROMOTING COMPLEX SUBUNIT
C
RPN11 PROTEIN DEGRADATION 26S PROTEASOME REGULATORY SUBUNIT
UFD1 PROTEIN DEGRADATION UBIQUITIN FUSION DEGRADATION
RPN9 PROTEIN DEGRADATION 26S PROTEASOME REGULATORY SUBUNIT
RPT1 PROTEIN DEGRADATION 26S PROTEASOME SUBUNIT
RPT6 PROTEIN DEGRADATION 26S PROTEASOME REGULATORY SUBUNIT
PRE4 PROTEIN DEGRADATION PROTEASOME SUBUNIT, B TYPE
RPN6 PROTEIN DEGRADATION 26S PROTEASOME REGULATORY SUBUNIT
RPT4 PROTEIN DEGRADATION 26S PROTEASOME REGULATORY SUBUNIT
RPN7 PROTEIN DEGRADATION 26S PROTEASOME REGULATORY SUBUNIT
RPN3 PROTEIN DEGRADATION 26S PROTEASOME REGULATORY SUBUNIT
PUP2 PROTEIN DEGRADATION 20S PROTEASOME SUBUNIT ALPHA5
SCL1 PROTEIN DEGRADATION 20S PROTEASOME SUBUNIT YC7ALPHA/Y8
PRE5 PROTEIN DEGRADATION 20S PROTEASOME SUBUNIT ALPHA6
PRE9 PROTEIN DEGRADATION 20S PROTEASOME SUBUNIT Y13 ALPHA3
PRE1 PROTEIN DEGRADATION 20S PROTEASOME SUBUNIT C11 BETA4
PRE2 PROTEIN DEGRADATION 20S PROTEASOME SUBUNIT BETA5
PRE3 PROTEIN DEGRADATION 20S PROTEASOME SUBUNIT BETA1
PRE10 PROTEIN DEGRADATION 20S PROTEASOME SUBUNIT C1 ALPHA7
PUP1 PROTEIN DEGRADATION 20S PROTEASOME SUBUNIT BETA2
PRE6 PROTEIN DEGRADATION 20S PROTEASOME SUBUNIT ALPHA4
PRE7 PROTEIN DEGRADATION 20S PROTEASOME SUBUNIT
RPN10 PROTEIN DEGRADATION 26S PROTEASOME SUBUNIT
RPT3 PROTEIN DEGRADATION 26S PROTEASOME SUBUNIT
RPT5 PROTEIN DEGRADATION 26S PROTEASOME REGULATORY SUBUNIT
RPN12 PROTEIN DEGRADATION 26S PROTEASOME REGULATORY SUBUNIT
RPN5 PROTEIN DEGRADATION 26S PROTEASOME SUBUNIT
RPN8 PROTEIN DEGRADATION 26S PROTEASOME REGULATORY SUBUNIT
D
POP6 RNA PROCESSING RNASE P AND RNASE MRP SUBUNIT
CAF16 TRANSPORT ATP-BINDING CASSETTE ABC FAMILY
MRE1 MRNA SPLICING UNKNOWN
SMD3 MRNA SPLICING CORE SNRNP PROTEIN
TAF40 TRANSCRIPTION TFIID 40 KD SUBUNIT
PET309 RNA PROCESSING COX1 MRNA STABILITY
PZF1 TRANSCRIPTION TFIIIA
YTH1 MRNA 3 -END PROCESSING CLEAVAGE/POLYADENYLATION FACTOR CF II COMPONENT
CBP1 MRNA STABILITY UNKNOWN
PRP24 MRNA SPLICING U4/U6 SNRNP PROTEIN
STD1 GLUCOSE REPRESSION REGULATOR OF GLUCOSE REPRESSION
MGM1 MITO GENOME MAINT DYNAMIN FAMILY PROTEIN
PRP19 MRNA SPLICING NON-SNRNP SPLICEOSOME COMPONENT
SLS1 MITOCHONDRIAL METABOLISM INTEGRAL MEMBRANE PROTEIN
E
TPI1 GLYCOLYSIS TRIOSEPHOSPHATE ISOMERASE
GPM1 GLYCOLYSIS PHOSPHOGLYCERATE MUTASE
PGK1 GLYCOLYSIS PHOSPHOGLYCERATE KINASE
TDH3 GLYCOLYSIS GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE 3
TDH2 GLYCOLYSIS GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE 2
ENO2 GLYCOLYSIS ENOLASE II
TDH1 GLYCOLYSIS GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE 1
FBA1 GLYCOLYSIS ALDOLASE
TKL1 PENTOSE PHOSPHATE CYCLE TRANSKETOLASE
PDC5 GLYCOLYSIS PYRUVATE DECARBOXYLASE
PDC6 GLYCOLYSIS PYRUVATE DECARBOXYLASE 3
PDC1 GLYCOLYSIS PYRUVATE DECARBOXYLASE
CDC19 GLYCOLYSIS PYRUVATE KINASE
HXK2 GLYCOLYSIS HEXOKINASE II
TYE7 GLYCOLYSIS BASIC H-L-H TRANSCRIPTION FACTOR
PFK1 GLYCOLYSIS PHOSPHOFRUCTOKINASE
ACS2 ACETYL-COA BIOSYNTHESIS ACETYL-COENZYME A SYNTHETASE
F
DB156 UBIQUINONE BIOSYNTHESIS METHYLTRANSFERASE
MSS51 MRNA SPLICING UNKNOWN
MSF1 PROTEIN SYNTHESIS MITOCHONDRIAL PHENYLALANYL-TRNA SYNTHETASE SUBUNIT
MRPS5 PROTEIN SYNTHESIS RIBOSOMAL PROTEIN, MITOCHONDRIAL S5
MSM1 PROTEIN SYNTHESIS MITOCHONDRIAL METHIONYL-TRNA SYNTHETASE
COX10 RESPIRATION CYTOCHROME-C OXIDASE ASSEMBLY
MSK1 PROTEIN SYNTHESIS MITOCHONDRIAL LYSYL-TRNA SYNTHETASE
MEF1 PROTEIN SYNTHESIS TRANSLATION ELONGATION FACTOR G, MITOCHONDRIAL
[illegible] PROTEIN SYNTHESIS RIBOSOMAL PROTEIN, MITOCHONDRIAL LARGE SUBUNIT
MRPL35 PROTEIN SYNTHESIS RIBOSOMAL PROTEIN, MITOCHONDRIAL L35
MRP7 PROTEIN SYNTHESIS RIBOSOMAL PROTEIN, MITOCHONDRIAL LARGE SUBUNIT
MAM33 RESPIRATION MITOCHONDRIAL ACIDIC MATRIX PROTEIN

applied to all of the approximately 6,200 genes of *S. cerevisiae*, the clusters of functionally related genes are maintained, but are usually expanded with the addition of uncharacterized genes (the results of this analysis will be the subject of a subsequent report).

2. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995) *Science* **270**, 484–487.

3. Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M.,