1 The current excitement in bioinformatics – [Au: OK?] analysis of whole-genome expression data: how does it relate to protein structure and function?

Mark Gerstein* and Ronald Jansen

Whole-genome expression profiles provide a rich new datatrove for bioinformatics. Initial analyses of the profiles have included clustering and cross-referencing to 'external' information on protein structure and function. Expression profile clusters do relate to protein function, but the correlation is not perfect, with the discrepancies partially resulting from the difficulty in consistently defining function. Other attributes of proteins can also be related to expression – in particular, structure and localization – and sometimes show a clearer relationship than function.

Addresses
Department of Molecular Biophysics and Biochemistry, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA
*e-mail: Mark.Gerstein@yale.edu

Current Opinion in Structural Biology 2000, **10**:XXX–XXX

0959-440X/00/$ – see front matter

Abbreviations
[Au: Please check the abbreviations carefully.]

EST	expressed sequence tag
PCA	principal component analysis
PCR	polymerase chain reaction
SAGE	serial analysis of gene expression
SOM	self-organizing map
SVM	support vector machine

Introduction

Bioinformatics has traditionally involved the computational analysis of large molecular biology datasets. Initially, these were drawn from the world of protein structure. In 1995, the field changed with the advent of complete genome sequences, which represented a new type of large-scale data. Now, whole-genome expression experiments are providing further sources of large-scale data and transforming bioinformatics yet again. Expression experiments can generate a quantity of information that potentially
2 dwarfs that provided by **[Au: OK?]** genome sequences and protein structures. Whereas it is sufficient, for many practical purposes, to view genome sequencing as a one-time process for each organism (except for the analysis of individual genetic variations), expression experiments can be repeated an arbitrary number of times to monitor the expression of different cell types and states (diseased or healthy), or the same cells at different times or in different individuals. The number of potential experiments is only limited by cost and imagination. Each of these experiments potentially gives rise to a new genome-scale dataset and a further challenge for bioinformaticians.

Expression data

Technologies and systems: SAGE, chips and arrays in yeast and beyond

Genome-wide expression information is principally generated by three technologies: cDNA microarrays [1], GeneChips (also called high-density oligonucleotide arrays) [2] and SAGE (serial analysis of gene expression) [3]. These technologies, which are all new and rapidly evolving, have been recently reviewed [4–6]. The large number of ESTs (expressed sequence tags) **[Au: OK?]** in 3
different cells and tissues provides a further source of large-scale expression information [7].

Expression monitoring on a genome-wide scale was first successfully demonstrated in yeast [8–10]. Later experiments have been performed on other organisms, including mycobacteria [11], Escherichia coli [12], worm (see http://bioinfo.mbb.yale.edu/genome/expression 4
[Au: OK?]), fly [13], mouse [14] and human [15,16]. There are a number of technical difficulties associated with certain systems (e.g. the lack of poly-A tails in bacteria) but, in principle, these experiments can be applied repeatedly in a wide variety of organisms.

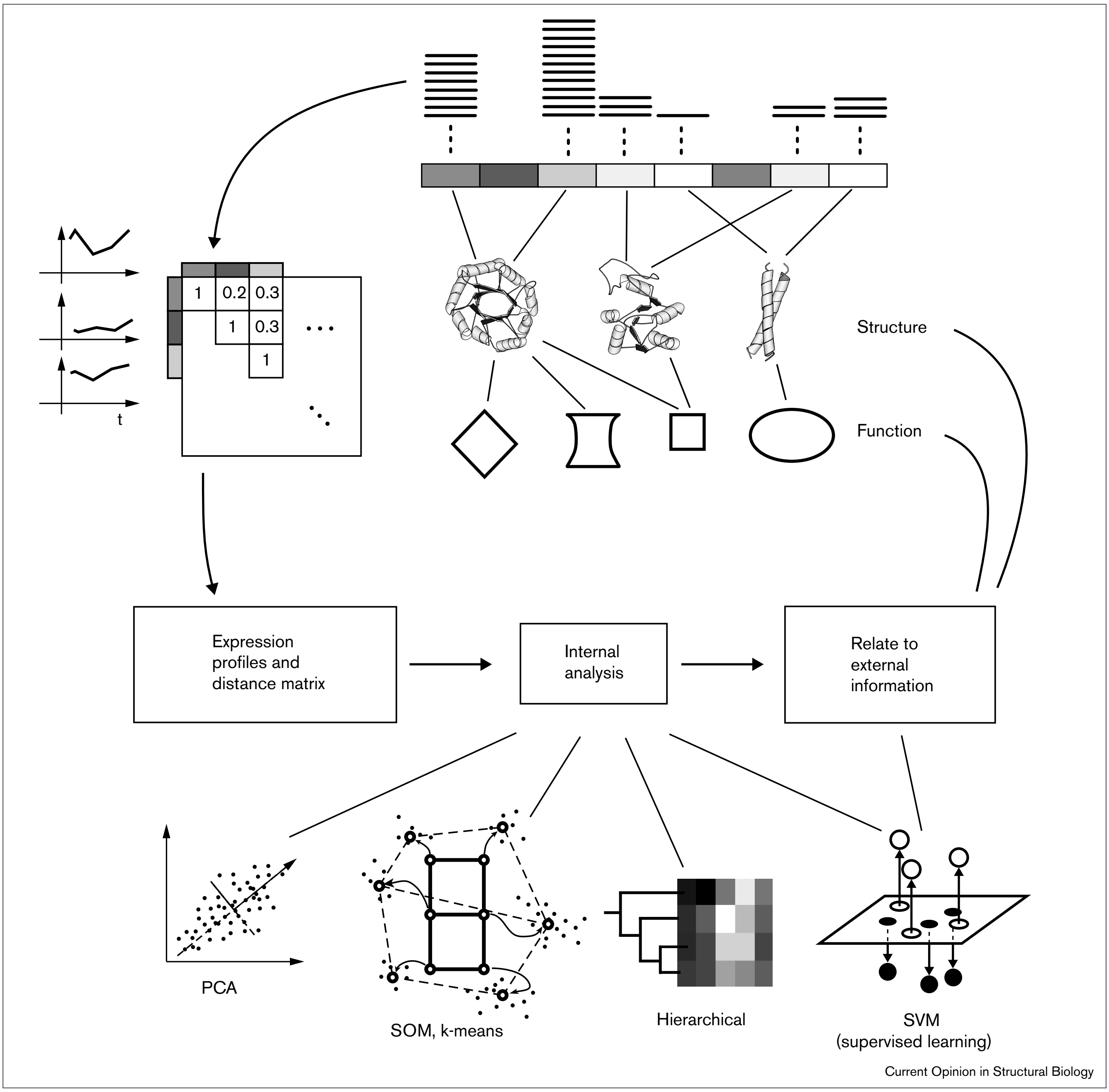
Relevant for computations: absolute versus relative, population averages and [Au: OK?] databases 5

From a computational perspective, the three expression technologies all produce a profile (or vector) of expression levels for many genes. In principle, GeneChips and SAGE allow the measurement of absolute expression levels (in units of mRNA transcripts per cell), whereas cDNA microarrays primarily measure changes relative to a reference state (yielding an 'expression ratio'). Although valuable, absolute transcript abundance measurements do not completely measure mRNA concentration, which also depends on cellular compartment volume.

Expression experiments measure cell population averages, not individual cells, so another important issue is the degree to which all cells in the investigated population are in the same 'state'. For single-cell organisms, temporal synchronization can often be achieved artificially, for
example, **[Au: OK?]** in yeast cell-cycle experiments, 6
cyclins were used for synchronization [10,17]. Work in multicellular organisms has the added complexity that expression measurements may combine many different tissues. Recent papers have discussed statistical aspects of expression data in detail [18,19•].

The first major bioinformatics task related to expression data is organization and storage. This is currently the sub-

Figure 1



ject of much discussion and there are a number of pilot databases: GEO (the NCBI Gene Expression Omnibus); ExpressDB [20] (Harvard); GeneX (NCGR); the Stanford Microarray Database; and ArrayExpress [21] (see
7 http://bioinfo.mbb.yale.edu/genome/expression **[Au: Does this link refer to all the databases or just ArrayExpress?]**).
8 Some issues being considered include **[Au: OK?]** whether to normalize and standardize the data, whether it should be stored in a central archive or federation of web sites and to what degree details about experimental design should be kept. Storing the raw array intensities lends itself nicely to standard relational tables. However, information related to the experimental conditions (tissues, drug treatments, etc.) is more complicated. To some degree, how to best archive the data will be determined by the most popular analyses that bioinformaticians end up performing.

Computational issues: internal versus external, supervised versus unsupervised

Analysis of expression datasets **[Au: OK?]** encourages more 9
exploratory, data-driven styles of research than traditional

Figure 1 legend

An overview of some principles of expression data analysis. The top part of the figure shows a representation of the input data. Expression data consist of expression level measurements for various genes arranged in 'profiles' across either different conditions or different times. One can determine the distance between each pair of profiles and put this into a large 'distance matrix', which then forms the basis for many of the clustering algorithms. (This is also known as a 'correlation matrix' or 'kernel matrix' in various calculations.) The top part also gives a schematization of the types of external information expression profiles can be related to. It shows part of a genome with the relationship between transcribed gene sequences and protein structure and function. Note how a number of genes can share the same protein fold, how certain protein folds can have many functions and how two different folds can have the same function. The bottom part of the figure illustrates a number of ways of analyzing expression data. Broadly, these can be divided into calculations dealing purely with the internal structure of the profiles and calculations relating the profiles to external, nonexpression information. Specific examples of various methods for analyzing the correlation matrix of expression profiles are PCA [34•], k-means clustering [30••] and SOMs [31••], hierarchical clustering [22••,25•] and SVMs [24•]. PCA tries to find the directions of greatest variance implied by the correlation matrix and to then 'visualize' the data in terms of their projection on these directions. Hierarchical clustering successively groups together the profiles that are the most similar, generating a tree-like description of the data. There are a variety of ways of making this determination of similarity; for example, in UPGMA, it is based on the distance to an existing averaged group center, whereas in single-linkage, it is based only on the distance to the nearest representative of a given cluster. K-means clustering algorithms make few assumptions about the data. They start with a number (k) of randomly positioned cluster centers and then update their positions to fit the data. SOMs are similar, but they impose a bit more structure on the clustering, requiring that the updated position of a cluster center be affected by the position of the other cluster centers. (In relation to the subschematic illustrating SOMs, adapted from [31••], note that SOMs would have constraints related to the dotted lines, whereas in k-means these would be absent). SVMs assume that the profiles are 'tagged' with already known classification information, such as a functional class. They then implicitly transform the data into a higher dimensional representation in which a simple plane can be found to separate the differently tagged groups. (In practice, this is accomplished by considering nonlinear measures for distance, beyond simple correlation.) SVMs are considered a type of supervised learning, in that they explicitly train and test against the external data. In contrast, SOMs, k-means and hierarchical clustering are considered unsupervised clustering, in that they do not relate the learned clusters to the external data until after they have been derived. However, one could imagine various unsupervised algorithms that simultaneously consider expression data and additional features derived from external information (such as localization) in learning clusters.

hypothesis-driven approaches. Expression data analysis can be loosely divided into two parts. In a first internal 10 **[Au: House-style dictates that italics can only be used for species names and Latin terms, would you like to use inverted commas instead?]** part, one analyzes the numerical structure of the data (e.g. by clustering) without explicitly relating expression levels to other biological information concerning protein function, structure, regulation and so on. In contrast, a second external 11 **[Au: See above]** part is primarily concerned with relating expression measurements to these 'external' information sources. The internal-external division is related to, but not the same as, 12 **[Au: OK?]** the supervised-unsupervised distinction, often used in machine learning [22••]. In supervised learning, an algorithm tries to find patterns in the data, given explicit sets of training and test examples preclassified on the basis of external data. Such 'tagged' data is not present in the unsupervised case. However, it is possible to subsequently relate patterns found in unsupervised learning to external data or to do unsupervised learning on a dataset consisting of expression profiles plus additional features.

Clustering: bottom-up hierarchies versus top-down partitions

The main type of internal analyses involves clustering and partitioning the data. As schematized in Figure 1, the starting point for clustering methods is defining a similarity measure among expression profiles and then constructing a matrix giving a distance between each pair of profiles. In general, there are many possible metrics [23••,24•]. A common one is the Pearson correlation coefficient [22••,25•]; **[Au: OK?]** an interesting modification of this is the 'jack- 13 knife correlation', which is robust with respect to data outliers [26].

Hierarchical methods group profiles in a 'bottom-up' fashion, joining the most similar profiles into clusters first and then including more diverse ones [27]. There are a variety of specific approaches (e.g. UPGMA **[Au; Would it be** 14 **helpful to briefly explain what this is?]**, single-linkage, multiple-linkage, etc.), which were mostly derived from phylogenetic tree construction [28]. These were the first methods applied to expression data [22••,25•,29•] and they have the advantage that the number of clusters needs not be specified beforehand. However, their drawback is that there is no reason to believe that expression data — in contrast to evolutionary information — is naturally organized in bifurcating trees. The trees produced by hierarchical clustering can only be broken into clusters in some ad hoc fashion. Furthermore, decisions made early in bottom-up clustering cannot be undone and sometimes adversely affect the final result.

In contrast to bottom-up clustering, partitioning approaches are 'top down'. Important examples applied to expression analysis are k-means [30••] and self-organizing maps (SOMs) [31••,32•]. A tree structure is not assumed in these methods; however, they often require an *a priori* decision on the number and structure of distinct clusters. It remains a problem to objectively determine the optimum number of clusters for these algorithms [19•]. Recently, partitioning algorithms have been developed in

Figure 2

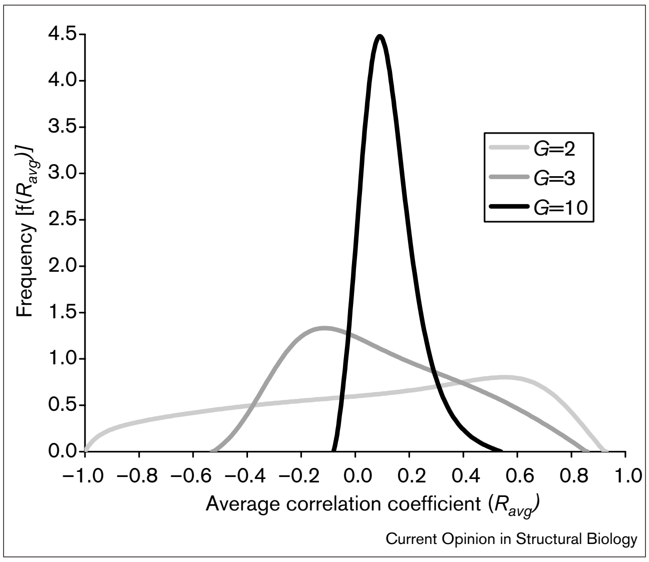


Figure 2 legend

**[Au: Please check the equations carefully. They are notoriously 19
difficult to typeset and mistakes can inadvertently be introduced.]** Sample distributions of the average correlation coefficient for groups of genes for expression data from the diauxic shift experiment. Most clustering algorithms are based on computing distances between the expression profiles of genes; in many cases, the Pearson correlation coefficient is used as a distance metric (see, for instance, [22••]). For two normalized expression ratio profiles $\mathbf{X}_\mathbf{i}$ and $\mathbf{X}_\mathbf{j}$ (each with average 0 and standard deviation 1), the Pearson correlation coefficient R_{ij} is given by the dot product:

$$R_{ij} = \frac{1}{N-1} \mathbf{X}_\mathbf{i} \bullet \mathbf{X}_\mathbf{j}$$

where N is the number of elements in the profiles $\mathbf{X}_\mathbf{i}$ and $\mathbf{X}_\mathbf{j}$. The normalized profile $\mathbf{X}$ can be computed as a 'Z-score' from the measured expression ratio profile $\mathbf{x}$ through the relation

$$X(k) = \frac{x(k) - x_{avg}}{\acute{o}_x}$$

where x_{avg} denotes the average and σ_x the standard deviation of values in $\mathbf{x}$, and $X(k)$ and $x(k)$ are the kth components of their respective profiles. Given a group of G genes, we can compute the correlation coefficient matrix $\mathbf{R}$, where each element (R_{ij}) of the matrix denotes the Pearson correlation coefficient between genes i and j. We can then compute an average correlation coefficient (R_{avg}) by averaging the matrix elements (excluding the main diagonal). This statistic gives an idea of the overall similarity of the expression profiles in a group of genes. Although there are $O(G^2)$ elements in $\mathbf{R}$, the computation time for R_{avg} can be kept proportional to $O(G)$ by calculating R_{avg} as follows:

$$R_{avg} = \frac{1}{G^2 - G}\left(\sum_{i,j}^{G} R_{ij} - G\right) = \frac{1}{G^2 - G}\left(\frac{1}{N-1}\mathbf{X}_\mathbf{Tot} \bullet \mathbf{X}_\mathbf{Tot} - G\right)$$

where

$$\mathbf{X}_\mathbf{Tot} = \sum_{g=1}^{G} \mathbf{X}_\mathbf{g}$$

is the sum of all expression profiles in the group of G genes. The figure shows the distribution of this statistic for the expression data measured during the diauxic shift in yeast [9]. Groups of genes of size G were randomly chosen from the genome. For $G = 2$, the statistic is simply the Pearson correlation coefficient itself. For increasing G, the distributions become narrower. The distributions were generated by sampling R_{avg} 10,000 times from the full distance matrix relating the expression profiles of all approximately 6000 genes in yeast. Functions of the form

$$F(y) = \sum_{i=0}^{6} a_i y^i \Big/ \sum_{i=0}^{6} b_i y^i$$

can be fit to the cumulative distributions, where

$$y = \ln\left(\frac{2}{1 - R_{avg}}\right)$$

is a transformation of the average correlation coefficient R_{avg}, with $a_6 = b_6 = 1$, $a_0 = a_1 = 0$. For the graph shown in the figure ($G = 2$), we used parameters $a_2 = 19.92$, $a_3 = 5.66$, $a_4 = -3.22$, $a_5 = -1.02$, $b_0 = 2.07$, $b_1 = 28.97$, $b_2 = 3.47$, $b_3 = 4.56$, $b_4 = -1.56$, $b_5 = -1.21$.

which the number of clusters is determined by the algorithm itself [33•].

An additional method of internal analysis is principal component analysis (PCA) [34•]. This method can be
15 understood **[Au: Would 'used' perhaps be better here?]** as a way of compressing the data and filtering out noise by projection onto a low-dimensional subspace. It can be used for data visualization and initial exploration of clusters.

Phenotype characterization: cancer diagnosis

16 Another type of internal analysis uses **[Au: OK?]** expression patterns to distinguish between cell types and disease states. In this context, entire expression profiles can be used to compare different experiments **[Au: Use inverted
17 commas?]** (in contrast to clustering genes). There have already been many applications in cancer diagnosis [35••,36,37•]; however, a full discussion is beyond the scope of this review.

Relating expression profiles to protein function

Thus far, we have only discussed computations aimed at
18 revealing the internal structure of **[Au: OK?]** expression data. Expert biological knowledge is applied afterwards to interpret the results. The next type of analysis tries to explicitly integrate information about protein function, structure and so forth directly into the expression data computations. First, we will look at work relating expression profiles to protein function. As a prelude, it is worthwhile to briefly discuss how protein functions are classified.

Functional classification and its problems

There are a number of schemes for classifying protein function, which have been recently reviewed [38]. Briefly, most of the schemes concentrate on a single organism, for example, MIPS for yeast, GenProtEC for E. coli, FlyBase for Drosophila and EGAD for human ESTs [39,40] (see
http://bioinfo.mbb.yale.edu/genome/expression). Other 20
schemes classify a subset of functions across a variety of organisms, for example, ENZYME for enzyme function and EcoCyc, WIT and KEGG for pathways [41–44]. There

Figure 3

(a)

MIPS category	Cell cycle (CDC28)	Cell cycle (CDC15)	Diauxic shift	Spo-rulation
	Experiment			
Cell growth, division and DNA synthesis	>4	>4	>4	>4
Protein synthesis	>4	>4	>4	>4
Transcription	>4	>4	>4	1.6
Cellular organization	>4	>4	0.3	0.3
Energy	>4	>4	0.1	0.9
Cell rescue, defense, death	>4	>4	0	0
Intracellular transport	>4	>4	0	0
Ionic homeostasis	>4	>4	0	0.8
Metabolism	>4	>4	0	0
Transport facilitation	>4	>4	0	0
Signal transduction	2.5	1.6	0.1	0.6
Unclassifield	2.3	>4	0	0
Cellular biogenesis	2.0	>4	0.4	0.2
Protein destination	0.3	>4	0.2	0.6
Retrotransposon and plasmid	0	2.8	1.9	1.0

(b)

MIPS category	Cell cycle (CDC28)	Cell cycle (CDC15)	Diauxic shift	Spo-rulation
	Experiment			
Respiration	>4	>4	>4	3.4
TCA pathway	>4	>4	>4	0.6
Glycogen, trehalose metabolism	>4	>4	1.2	0.7
Glycolysis	>4	>4	0.9	2.1
Gluconeogenesis	3.7	>4	0.1	1.7
Glyoxylate cycle	1.6	0.7	3.0	2.3
Pentose-phosphate pathway	1.5	0.8	0	0.6
Fermentation	1.3	>4	0	2.2
Other energy generation activities	0.7	0.1	0.1	0.2
Beta-oxidation of fatty acids	0.5	0.4	0.4	0.2

(c)

	Fraction of significant groups				Total # groups
	CDC28	CDC15	Diauxic shift	Sporu-lation	
MIPS 1	63%	81%	19%	13%	16
MIPS 2	50%	63%	17%	13%	102
MIPS 3	23%	33%	5%	4%	73
"Energy" (2nd level)	40%	60%	20%	0%	10
SOM	93%	-	-	-	30
Hierarchical clustering	80%				25

Current Opinion in Structural Biology

21 have been **[Au: OK?]** some attempts to merge functional classifications for different organisms into one common
22 source (the Gene Ontology Project [45], see http://bioinfo.mbb.yale.edu/genome/expression), although the creation of a complete universal functional system will be a difficult task [38,46]. However, there have been some attempts in terms of creating unique keyword combinations or sequence variability signatures for functions [47,48].

Beyond the lack of scope of the current classification schemes, it is important to realize that there are many profound difficulties in functional classification. First, the concept of 'function' is itself rather vague. Sometimes it is defined in terms of biochemical mechanism (e.g. 'adenylate kinase'); at other times, in terms of either **[Au: OK?]** 23
involvement in pathways or overall cellular role(e.g. part of 'glycolysis' or 'cellular metabolism'); and, finally, sometimes in terms of the phenotype of the organism when the associated gene is disabled (e.g. 'causes cancer'). **[Au: I**

Figure 3 legend

The degree of expression profile similarity is different for genes from different functional groups and also varies between different expression experiments. We illustrate this concept in the context of the
24 MIPS functional classification scheme. Each part shows **[Au: OK?]** the negative logarithm of the one-sided P-values [−log(P)] based on distributions of the average correlation coefficient for different experiments, as explained in the legend to Figure 2. The P-values give the probability that an average correlation greater than that observed for each functional group could have arisen from a randomly selected group of genes of the same size. Accordingly, lower P-values or higher values of −log(P) indicate a greater significance of the similarity between expression profiles. The P-values range from 0 to 1; correspondingly, −log(P) ranges from infinity to 0. For values of −log(P) greater than four, we cannot determine the value with certainty because of the limited scale of our computation; we indicate this in the table by '>4' (for highly significant groupings). **[Au: What does the
25 shading in the table represent? The degree of significance of the grouping. Please explain.]** Each row in parts **(a)** and **(b)** of the figure corresponds to an MIPS functional category and each column corresponds to a different expression experiment on yeast. The first experiment is a GeneChip experiment [17] to monitor the cell cycle synchronized by the cyclin CDC28. The other experiments are microarray experiments: the cell cycle synchronized by CDC15 [10], the diauxic shift [9] and the process of sporulation [8]. Part (a) shows the most general MIPS categories, whereas (b) shows the subcategories of the top-level MIPS category 'energy'. Part **(c)** summarizes the fraction of functional categories that represent 'significant' groupings with respect to expression. We define a grouping as significant if we find values of $-\log(P) > 3$, a less than 1 in 1000 chance that the observed average correlation arises randomly. The first column indicates the level in the MIPS hierarchy. (MIPS 1 is the first level, MIPS 2 is the second level, etc.) The next columns show the fraction of significant groups for each experiment and the last column shows the total number of groups in each MIPS level. The fraction of significant groups decreases as the detail of classification increases from the first to the third MIPS level. This is because (for the quantitative assessment presented here) a high significance for a more specialized MIPS category tends to also show up in a high significance for the more general MIPS category one level above. In part (c), we show the significance of the clustering determined by various methods described in the text – in particular, hierarchical clustering [21], k-means [30••] and SOMs [2]. The hierarchical clustering was applied to all four experiments and to additional data on the mitotic cell cycle, and temperature and reducing shocks (see
http://bioinfo.mbb.yale.edu/genome/expression). To apply the 29
methodology, the hierarchical tree was cut off such that 25 'subtrees' or gene clusters remained. Clearly, both these methods produce much more statistically significant clustering with respect to expression than the MIPS functional groups. The only functional categories for which we find high significance in all four experiments are at the top of the table: 'cell growth, division and DNA synthesis' and 'protein synthesis' (including ribosomal proteins). In contrast, some categories are not significant in any of the experiments (such as 'beta-oxidation of fatty acids', another subcategory of 'energy'). In general, there seems to be a higher degree of correlation for the two cell-cycle experiments than for the other two experiments (e.g. for the 'metabolism' category), perhaps because the mechanics of the cell cycle forces a high degree of transcriptional coexpression on many functional systems. However, a few functional groups show a higher significance in the diauxic shift and sporulation experiments (such as the group 'glyoxylate cycle', which is a subcategory of 'energy'). It can be clearly seen that many functional groups show different degrees of coexpression under different experimental conditions, highlighting the importance of experimental design.

**have bracketed the examples in order to make a long sen-
26 tence more digestible. Please let me know if you think that I have failed!]** Second, many proteins are multifunctional, having more than one function, sometimes in unrelated areas [49]. For instance, the protease thrombin is primarily associated with blood clotting, but also interacts with receptors for cell activation and neural development [50]. Third, conversely, multiple gene products often collectively carry out a single function (e.g. the ribosome). Fourth, the naming of functions is currently unsystematic and inappropriate for quantitative comparisons. Humorous
27 examples of this come from the fly, for which **[Au: OK?]** genes have most bizarre names, for example, 'suppressor-of-white-apricot' and 'darkener-of-apricot', which are, respectively, an RNA-binding protein and a kinase involved in eye-color determination (SUWA_DROME and
28 DOA_DROME **[Au: Please explain what these terms relate to.]**). There have been some attempts in terms of creating unique keyword combinations or objective sequence variability signatures for functions [47,51,52].

Supervised learning (support vector machines)

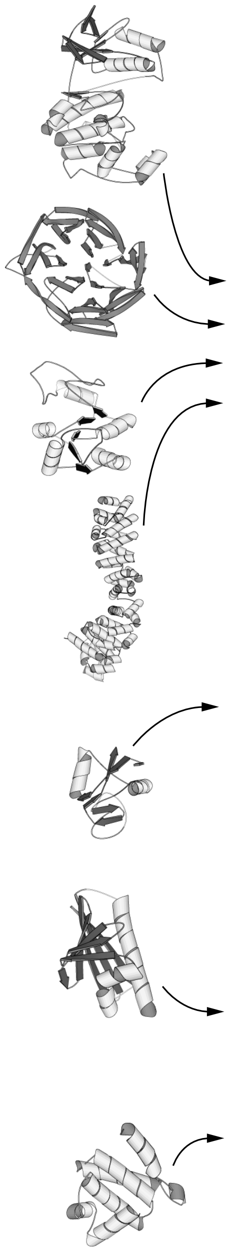
Given a function classification, one would like to know how well clusters of expression profiles relate to functional categories and, if there is a relation, the degree to which it can be used to predict the functions of genes. Some initial reports on expression analysis suggested that certain prominent expression clusters did relate to functional categories and that function prediction was possible [22••,53,54]. More recent work has tried to systematically test this proposition using explicit training and testing sets. As diagramed in Figure 1, one technique that has been applied is support vector machines (SVMs) [24•]. This supervised learning technique positions a hyperplane to partition the data and minimize the number of misclassified proteins on the basis of a known functional classification or empirical measurements not included in the dataset.

Other supervised learning approaches include decision trees, Parzen windows and Fisher's linear discriminant [24•]. More general approaches that make use of prior information include Bayesian networks [55].

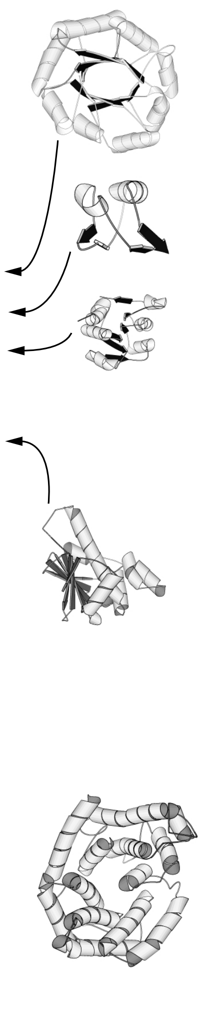
Global characterization of the expression/function relationship

The calculations relating expression and function have largely focused on specific cases or functional categories. Figures 2 and 3 attempt to give an overview of how they
relate in a 'global' sense. On the basis of **[Au: OK?]** the 26
results of a whole-genome expression experiment, one can determine the distribution of similarity values for each pair
of genes, that is, the distribution of the **[Au: OK?]** correla- 27
tions of their expression profiles. For groupings larger than

Figure 4



Fold of	Genome	Transcriptome	Experiment					Rep. PDB
			CDC28	CDC15	Diauxic shift	Sporulation	*E. coli* heat shock	
Protein kinases (catalytic core)	1	18	94	98	139	60	100	1p38
β-Propeller	2	5	160	108	109	82	-	1mda
P-loop NTP hydrolases	3	2	100	88	91	57	39	1gky
α–α superhelix	4	6	136	90	110	44	55	2bct
TIM-barrel	5	1	58	57	39	24	91	1byb
Ferredoxin-like	6	3	135	61	63	70	144	1fxd
Rossmann fold	8	4	55	99	43	56	92	1xel
Ribonucleotide reductase (R1)	100	143	1	-	-	-	35	1rlr
ATPase domain of HSP90	100	91	2	4	72	73	2	1ah6
Homing endonuclease-like	130	164	3	136	85	175	41	1af5
Aminoacid dehydrogenases; dim. dom.	-	-	4	169	121	3	51	1hup
DNA topo I (N term)	-	-	175	1	148	126	-	1ois
DNA clamp	130	115	8	2	87	11	60	2pol
Metallothionein	100	14	89	3	33	12	-	1mhu
Phosphoenolpyruvate carboxykinase	130	190	51	26	1	96	169	1ayl
Citrate synthase	81	120	14	8	2	28	51	1csh
N-carbamoylsarcosine amidohydrolase	130	112	9	-	3	138	118	1nba
TBP-like	81	91	46	38	4	75	100	1bv1
5'–3' exonuclease	67	150	32	125	162	1	157	1tfr
α/α toroid	62	132	169	145	114	2	100	1gai
Cyclin-like	20	61	20	15	129	4	-	1vin
ATPase domain of GroEL	36	34	183	143	61	151	1	1aon
Head domain of GrpE	130	135	196	31	165	165	3	1dkg
HSP70 (C term)	31	10	16	11	56	117	4	1dkz



Current Opinion in Structural Biology

pairs (e.g. triplets), this can be generalized to the distribution of the average value of the correlation. Sample
28 distributions based on the yeast **[Au: OK?]** diauxic shift timecourse [9] are shown in Figure 2. And, as shown in Figure 3, with respect to any particular expression experiment, distributions can be used to evaluate the statistical significance of a given clustering of genes. Most of the clusters automatically generated using the algorithms discussed earlier (e.g. hierarchical clusters or SOMs) appear to be significant. For instance, on the basis of a P < .001 threshold, 28 of the 30 SOM clusters for the cell-cycle data are significant (93%). However, fewer gene groupings based on the functional categories in MIPS are significant, for example, only 10 of the 16 top-level MIPS clusters have P < .001 (63%) for the same experiment. Some functional groups are always highly correlated with expression profiles (e.g. 'cell growth' and 'protein synthesis'). However, other MIPS groups are only correlated in certain experi-

Figure 4 legend

This figure shows how expression data can be related to protein structure. It shows a number of protein folds in the yeast, *E. coli* and *C. elegans* genomes ranked by various measures related to expression 33 **[Au: Again, what does the different shading of the rankings show?]**. At least the four most common folds for each type of ranking are shown. The first column shows the rank of the fold in terms of how many times it is found in the yeast genome (i.e. by duplication), based on recent PSI-BLAST structural assignments [63]. The next column shows its ranking in the transcriptome [64•], that is, the occurrence of each fold weighted by the number of copies of mRNA associated with it, based on GeneChip data [75]. Folds can be also ranked in terms of their fluctuation in mRNA levels over an experiment, rather than their total number of mRNA copies, using the average standard deviation of the expression ratios as an indication of the degree of fluctuation. Such rankings are shown in the next four columns. Columns 3 and 4 show a ranking based on the fluctuation in expression in the yeast cell cycle (CDC28 [17] and CDC15 [10]). Columns 5 and 6 show rankings based on other yeast experiments, the diauxic shift [9] and sporulation [8]. For comparison, columns 7 and 8 show the ranking for other organisms, *E. coli* (based on fluctuation in the heat shock experiment 34 [12]) and *C. elegans* (based on the fluctuations during successive larval stages of the worm; V Reinke, personal communication). **[Au:** 35 **Does column 8 not give the PDB codes of the proteins, rather than the *C. elegans* ranking? Have I got the correct figure? Please can you check and adjust the legend accordingly.]** Note how different all the rankings are. The most common folds in the transcriptome have a mixed α/βstructural architecture and are mostly cytosolic enzymes. The most abundant fold is the TIM barrel, which is also known to be the most versatile fold, associated with 16 different enzymatic functions [63]. In terms of the fluctuation rankings, one fold that changes considerably in expression is that of 'ATPase domain of HSP90/DNA topoisomerase II', which is highly ranked in both cell-cycle experiments (CDC28 and CDC15) and the *E. coli* experiment. The folds are selected from the current 520 folds and 771 superfamilies as of 1 November 1999 in SCOP 1.48 [56]. For the yeast fluctuation rankings, we excluded genes with an absolute expression level lower than 100 units of intensity, as given by the CDC28 GeneChip, because the signal fluctuations of lowly expressed genes are most likely due to measurement uncertainties. (The absolute expression level is defined as the difference between the intensity of the oligonucleotide-perfect match [PM] and the background intensity measured by a single mismatch probe [MM].) For the *E. coli* experiment, we simply ranked the expression ratio because no time series measurements were taken. For the *C. elegans* fluctuation ranking, we excluded signals with less than 250,000 units **[Au: Again,** 38 **is there a *C. elegans* ranking in the figure?]. [Au: What is 'dim dom' with respect to amino acid dehydrogenase? What is TBP? What is shown by the Rep. PDB column?]**

ments, for example, the 'metabolism' category and the 'glycolysis' subcategory are only correlated with expression in the cell-cycle experiments.

The lack of correlation between expression profiles and functional categories can be explained, to some degree, in terms of the different conditions of each experiment. However, it also reflects the problematic aspects of functional classification described earlier. Many of the MIPS classes comprise genes that one would not expect to be correlated, for example, 'regulation of phosphate utilization' (P = .23), and it will be difficult to standardize the functional categories enough that these inconsistencies disappear.

Relating expression data to protein structure

Although function is, in a sense, the most obvious aspect of proteins to relate to expression, many other attributes of proteins can be cross-referenced against expression data (e.g. their structure, localization, regulation, interactions and so forth). It is particularly worthwhile to relate protein structure to expression profiles for two reasons.

First, many of the classification ambiguities with respect to 36 **[Au: OK?]** function are not present with respect to 37 **[Au: OK?]** structure, so the foundation of the analysis is more precise. In particular, there are a number of 'universal' (across-organism) schemes classifying all known structures into approximately 500 folds (e.g. SCOP, CATH, FSSP and VAST [56–58]). These schemes, which have been reviewed elsewhere [59], principally differ in the degree to which they are based on automatic or manual curation, and are considerably more systematic and objective than any of the functional classification schemes. Furthermore, their annotation can be 'transferred' to genomes as a function of sequence similarity, which is **[Au: OK?]** based on well-established quantitative relation- 39 ships [46,60,61]. Finally, recent surveys of the relationship between fold and function indicate that most folds have only a single biochemical function, whereas a few generic scaffolds, such as the TIM barrel or α/β hydrolase, can accommodate many functions (>10) [62,63]. Thus, much of the lumping together of disparate genes into a single erroneous 'category' can be avoided if one first **[Au: OK?]** 40 classifies sequences based on single-function folds, rather than jumping directly to function.

Building on the classification of structures, it is possible to determine whether there are shared structural characteristics of highly expressed proteins. Recent surveys [64•,65•] have shown that highly expressed **[Au: OK?]** proteins in 41 yeast are of mixed helix-sheet architecture, enriched in alanine, relatively short and involved in metabolic and synthetic functions. In contrast, folds of membrane proteins or of proteins **[Au: OK?]** with all-helical or all-sheet architec- 42 ture are expressed at considerably lower levels. Figure 4 highlights these results, showing particular folds that are highly expressed and also folds that change in expression considerably over a timecourse. Note that these two groups are essentially disjoint; there being no folds that are both highly expressed and highly variable in expression over a timecourse. In particular, the most highly expressed fold in yeast, the TIM barrel, is not the same as the most commonly duplicated fold in the genome nor is it the same as the folds that vary most in expression in the various experiments.

Relating expression data to other external information

Another attribute of a gene that can be related to its expression profile is its regulation. This subject has been reviewed in detail [66], so we will only touch upon it briefly here. Almost by definition, genes that have similar expression profiles probably share upstream regulatory elements. This fact has been exploited to search for new regulatory sequences [30••,67–69]. For genes that have similar expression profiles but do not share an obvious regulatory element, one can use an unsupervised motif learner, such as a Gibbs sampler [70], to discover new regulatory motifs in upstream sequences.

Other attributes of proteins that have been related to expression include subcellular localization and protein–protein interactions. As was the case with protein structure, these attributes of proteins can be more precisely systematized than function. For yeast, systematic information on localization and interactions is tabulated in the MIPS, YPD and SwissProt databases [40,71,72]. With regard to localization, it has been found that cytosolic proteins tend to be expressed at high levels, whereas proteins destined for membranes and mitochondria are expressed at
43 lower levels [73•]. Proteins in the secretory pathway have high fluctuations in expression level over timecourses. Collectively, this information can, in fact, be combined to help predict the localization of proteins for which there is expression information available, but no known localization [74].

Conclusions

The advent of whole-genome expression experiments has led to a new class of bioinformatics analyses. These fall into two main groups: internal clustering and comparison of expression data, and cross-referencing of expression data to other information on protein structure and function. With respect to the experiments on yeast, clusters of genes that have similar expression profiles often fall into the same functional category. However, this is not always true in a 'global' sense. The discrepancies reflect particular functional categories highlighted by certain experiments. More importantly, they also result from the difficulty in consistently defining function across a wide variety of proteins. We believe this latter difficulty is quite significant and probably the major current impediment to interpreting expression data in terms of protein function. We can side-step this to some degree by focusing on attributes of proteins other than function, such as structure, regulation and localization. Many of these can be defined in a much more consistent fashion than function and, perhaps because of this, show a clearer relation to gene expression.

44 **Acknowledgements**

On the web, we will make available supplementary data related to the review (extended versions of Figures 2, 3 and 4, with a list of fold expression levels and function significance values for the whole yeast genome) and a 'links page' to web sites referred to in the text. Go to http://bioinfo.mbb.yale.edu/genome/expression **[Au: This is probably the best place to put this (there is no precedent for this sort of supplementary material in the Current Opinion journals). It is a very good idea.]** 45

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

• of special interest
•• of outstanding interest

1. Shalon D, Smith SJ, Brown PO: **A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization.** *Genome Res* 1996, **6**:639-645.

2. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H *et al.*: **Expression monitoring by hybridization to high-density oligonucleotide arrays.** *Nat Biotechnol* 1996, **14**:1675-1680.

3. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B, Kinzler KW: **Characterization of the yeast transcriptome.** *Cell* 1997, **88**:243-251.

4. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: **Expression profiling using cDNA microarrays.** *Nat Genet* 1999, **21**:10-14.

5. Brown PO, Botstein D: **Exploring the new world of the genome with DNA microarrays.** *Nat Genet* 1999, **21**:33-37.

6. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ: **High density synthetic oligonucleotide arrays.** *Nat Genet* 1999, **21**:20-24.

7. Ewing RM, Kahla AB, Poirot O, Lopez F, Audic S, Claverie JM: **Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression.** *Genome Res* 1999, **9**:950-959.

8. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I: **The transcriptional program of sporulation in budding yeast.** *Science* 1998, **282**:699-705.

9. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278**:680-686.

10. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9**:3273-3297.

11. Gingeras TR, Ghandour G, Wang E, Berno A, Small PM, Drobniewski F, Alland D, Desmond E, Holodniy M, Drenkow J: **Simultaneous genotyping and species identification using hybridization pattern recognition analysis of generic Mycobacterium DNA arrays.** *Genome Res* 1998, **8**:435-448.

12. Richmond CS, Glasner JD, Mau R, Jin H, Blattner FR: **Genome-wide expression profiling in *Escherichia coli* K-12.** *Nucleic Acids Res* 1999, **27**:3821-3835.

13. White KP, Rifkin SA, Hurban P, Hogness DS: **Microarray analysis of Drosophila development during metamorphosis.** *Science* 1999, **286**:2179-2184.

14. Lee CK, Klopp RG, Weindruch R, Prolla TA: **Gene expression profile of aging and its retardation by caloric restriction.** *Science* 1999, **285**:1390-1393.

15. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson J Jr, Boguski MS *et al.*: **The transcriptional program in the response of human fibroblasts to serum.** *Science* 1999, **283**:83-87.

16. Kaminski N, Allard JD, Pittet JF, Zuo F, Griffiths MJ, Morris D, Huang X, Sheppard D, Heller RA: **Global analysis of gene expression in pulmonary fibrosis reveals distinct programs regulating lung inflammation and fibrosis.** *Proc Natl Acad Sci* USA 2000, **97**:1778-1783.

17. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ *et al.*: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mol Cell* 1998, **2**:65-73.

18. Chen Y, Dougherty E, Bittner M: **Ratio-based decisions and the quantitative analysis of cDNA microarray images.** *J Biomed Optics* 1997, **2**:364-374.

19. Claverie JM: **Computational methods for the identification of**
46 • **differential and coordinated gene expression.** *Hum Mol Genet* 1999, **8**:1821-1832.
[Au: Each bulleted reference must be accompanied by an annotation. Please either delete the bullet or provide a few sentences describing the results of the paper and their implications.]

20. Aach J, Rindone W, Church GM: **Systematic management and analysis of yeast gene expression data.** *Genome Res* 2000, **10**:431-445.

21. Brazma A, Robinson A, Cameron G, Ashburner M: **One-stop shop for microarray data.** *Nature* 2000, **403**:699-700.

22. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and**
•• **display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95**:14863-14868.
The authors show the clustering of cDNA microarray data and graphical representation of the results using the Pearson correlation coefficient to define distances between expression profiles. They then used **[Au: OK?]** an algorithm similar to the average-linkage method to generate a tree. First, the pair of genes with the highest similarity was identified from the correlation coefficient matrix and the first node in the dendrogram was formed. Then the two genes were summarized by their average expression profile and the matrix was recomputed. This procedure was iterated until the dendrogram was finished. It was shown that this procedure groups a number of functionally related genes, such as those encoding ribosomal proteins and proteins involved in translation, the proteasome, the mini-chromosome maintenance
47 DNA replication complex, numerous glycolytic enzymes and enzymes of **[Au: OK?]** the TCA cycle **[Au: List OK?]**.

48

23. Michaels G, Carr D, Askenazi M, Fuhrman S, Wen X, Somogyi R:
•• **Cluster analysis and data visualization of large-scale gene expression data.** *Pac Symp Biocomput* 1998:42-53.
The authors analyzed gene expression patterns for the rat cervical spinal cord generated by RT-PCR. The paper demonstrates clustering of these patterns using the FITCH software, which was initially designed for the generation of evolutionary trees. In addition to the Euclidian distance, the authors applied mutual information (based on information theoretic entropy
49 **[Au: OK?]**) as a similarity measure among expression patterns. This has the advantage that not only positive, linear relations, but also negatively, non-linearly correlated expression patterns are recognized as proximal.

24. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS,
•• Ares M Jr, Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc Natl Acad Sci USA* 2000, **97**:262-267.
[Au: Each bulleted reference must be accompanied by an annotation.
50 **Please either delete the bullet or provide a few sentences describing the results of the paper and their implications.]**

25. Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL,
• Somogyi R: **Large-scale temporal gene expression mapping of central nervous system development.** *Proc Natl Acad Sci USA* 1998, **95**:334-339.
This paper describes a clustering algorithm that uses a 'jack-knife correlation coefficient'. This is the minimum of a set of correlations computed for two vectors, with one value pair removed each time. This procedure is robust to single outliers. The clustering algorithm assesses the quality of clusters by measuring the minimum similarity between two genes in the cluster.

26. Heyer LJ, Kruglyak S, Yooseph S: **Exploring expression data: identification and analysis of coexpressed genes.** *Genome Res* 1999, **9**:1106-1115.

27. Kaufman L, Rousseeuw P: *Finding Groups in Data*. New York; 1990. **[Au: Please provide the name of the publishers of this book.]**

28. Felsenstein J: **Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods.** *Methods Enzymol* 1996, **266**:418-427.

[Au: I am afraid that only papers published during or after 1999
51 **(although I let a few 1998 annotations slip through) or papers that are 'in press' may be annotated and bulleted. Therefore, I have had to delete this annotation.]**

29. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ:
• **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.** *Proc Natl Acad Sci USA* 1999, **96**:6745-6750.
[Au: Each bulleted reference must be accompanied by an annotation.
52 **Please either delete the bullet or provide a few sentences describing the results of the paper and their implications.]**

30. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM:
•• **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22**:281-285.
The authors demonstrate the application of self-organizing maps (SOMs) to cluster gene expression data of the yeast cell cycle and hematopoietic differentiation. An advantage of the method is the short computation time. SOMs are one method of extracting the most prominent patterns from a dataset. They are similar to k-means, but assume a more structured (correlated) distribution of cluster centers. Although this approach does appear to work well on expression profiles, it is not clear what in expression data justifies the added structure in the SOM approach.

31. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E,
•• Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci USA* 1999, **96**:2907-2912.
The authors demonstrate the application of self-organizing maps (SOMs) to cluster gene expression data of the yeast cell cycle and hematopoietic differentiation. An advantage of the method is the short computation time. SOMs are one method of extracting the most prominent patterns from a data set. They are similar to k-means, but assume a more structured (correlated) distribution of cluster centers. Although this approach does appear to work well on expression profiles, it is not clear what in expression data justifies the added structure in the SOM approach.

32. Toronen P, Kolehmainen M, Wong G, Castren E: **Analysis of gene**
•• **expression data using self-organizing maps.** *FEBS Lett* 1999, **451**:142-146.
Independently of the work described in [31••], the authors demonstrate the application of self-organizing maps to the expression data for the diauxic shift [9]. The self-organizing maps algorithm is an unsupervised neural network learning algorithm.

33. Ben-Dor A, Shamir R, Yakhini Z: **Clustering gene expression**
• **patterns.** *J Comput Biol* 1999, **6**:281-297.
The authors describe the theory and practical application of a partitioning clustering algorithm that recovers an original cluster structure with high probability when the expression data exhibits stochastic errors.

34. Raychaudhuri S, Stuart J, Altman R: **Principal component analysis to**
• **summarize microarray experiments: application to sporulation time series.** *Pac Sym Biocomput* 2000. **[Au: Do you have any further details for this reference or is it 'in press'?]** 53
The authors describe the application of principal component analysis (PCA) to gene expression data measured in a sporulation timecourse experiment 54
[8] **[Au: OK?]** as a tool to extract the experimental conditions that contain most of the information from a multicondition expression experiment. The principal components are a subset of the N eigenvectors of the N x N covariance matrix of experimental conditions (N = 7 time points), chosen such that most of the variance can be accounted for. (Here 90% of the variance can be accounted for by the first two of seven components.) The variances accounted for by each eigenvector are derived from the associated eigenvalues. The authors suggest that PCA can be used to decide whether the data are suitable for clustering. They imply that the expression data do not have an obvious clustering.

35. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M,
• Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA *et al.*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286**:531-537.
[Au: Each bulleted reference must be accompanied by an annotation.
Please either delete the bullet or provide a few sentences describing 55
the results of the paper and their implications.]

36. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van De Rijn M, Waltham M *et al.*: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet* 2000, **24**:227-235.

37. Califano A, Stolovitzky G, Tu Y: **Analysis of gene expression**
• **microarrays for phenotype classification.** *Ismb* 2000. **[Au: Do you have any further details for this reference or is it 'in press'?]**
[Au: Each bulleted reference must be accompanied by an annotation.
Please either delete the bullet or provide a few sentences describing 56
the results of the paper and their implications.]

38. Riley M: **Systems for categorizing functions of gene products.** *Curr Opin Struct Biol* 1998, **8**:388-392.

39. Riley M: **Genes and proteins of *Escherichia coli* K-12.** *Nucleic Acids Res* 1998, **26**:54.

40. Mewes HW, Frishman D, Gruber C, Geier B, Haase D, Kaps A, Lemcke K, Mannhaupt G, Pfeiffer F, Schuller C *et al.*: **MIPS: a database for genomes and protein sequences.** *Nucleic Acids Res* 2000, **28**:37-40.

41. Bairoch A: **The ENZYME database in 2000.** *Nucleic Acids Res* 2000, **28**:304-305.

42. Karp PD, Riley M, Saier M, Paulsen IT, Paley SM, Pellegrini-Toole A: **The EcoCyc and MetaCyc databases.** *Nucleic Acids Res* 2000, **28**:56-59.

43. Selkov E Jr, Grechkin Y, Mikhailova N, Selkov E: **MPW: the Metabolic Pathways Database.** *Nucleic Acids Res* 1998, **26**:43-45.

44. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Res* 1999, **27**:29-34.

45. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al.*: **Gene ontology: tool for the unification of biology.** *Nat Genet* 2000, **25**:25-29.

46. Wilson CA, Kreychman J, Gerstein M: **Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores.** *J Mol Biol* 2000, **297**:233-249.

47. Naylor G, Gerstein M: **Measuring shifts in function and evolutionary opportunity using variability profiles: a case study of the globins.** *J Mol Evol* 2000, in press. **[Au: Do you have any**
57 **further details (volume and page range) for this reference?]**

48. Tamames J, Ouzounis C, Casari G, Sander C, Valencia A: **EUCLID: automatic classification of proteins in functional classes by their database annotations.** *Bioinformatics* 1998, **14**:542-543.

49. Jeffery CJ: **Moonlighting proteins.** *Trends Biochem Sci* 1999, **24**:8-11.

50. Coughlin SR, Vu TK, Hung DT, Wheaton VI: **Characterization of a functional thrombin receptor. Issues and opportunities.** *J Clin Invest* 1992, **89**:351-355.

51. des Jardins M, Karp PD, Krummenacker M, Lee TJ, Ouzounis CA: **Prediction of enzyme classification from protein sequence without the use of sequence similarity.** *Ismb* 1997, **5**:92-99.

52. Tamames J, Casari G, Ouzounis C, Valencia A: **Conserved clusters of functionally related genes in two bacterial genomes.** *J Mol Evol* 1997, **44**:66-73.

53. Niehrs C, Pollet N: **Synexpression groups in eukaryotes.** *Nature* 1999, **402**:483-487.

54. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function.** *Nature* 1999, **402**:83-86.

55. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** 1999. **[Au: Do you have any more**
58 **details for this reference? I have looked in PubMed and Medline, but couldn't find anything. Please supply the journal name, volume and page range.]**

56. Lo Conte L, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C: **SCOP: a structural classification of proteins database.** *Nucleic Acids Res* 2000, **28**:257-259.

57. Pearl FM, Lee D, Bray JE, Sillitoe I, Todd AE, Harrison AP, Thornton JM, Orengo CA: **Assigning genomic sequences to CATH.** *Nucleic Acids Res* 2000, **28**:277-282.

58. Holm L, Sander C: **Touring protein fold space with Dali/FSSP.** *Nucleic Acids Res* 1998, **26**:316-319.

59. Brenner SE, Chothia C, Hubbard TJ: **Population statistics of protein structures: lessons from structural classifications.** *Curr Opin Struct Biol* 1997, **7**:369-376.

60. Chothia C, Lesk AM: **The relation between the divergence of sequence and structure in proteins.** *EMBO J* 1986, **5**:823-826.

61. Wood TC, Pearson WR: **Evolution of protein sequences and structures.** *J Mol Biol* 1999, **291**:977-995.

62. Martin AC, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JB, Taroni C, Thornton JM: **Protein folds and functions.** *Structure* 1998, **6**:875-884.

63. Hegyi H, Gerstein M: **The relationship between protein structure and function: a comprehensive survey with application to the yeast genome.** *J Mol Biol* 1999, **288**:147-164.

64. • Jansen R, Gerstein M: **Analysis of the yeast transcriptome with structural and functional categories.** *Nucleic Acids Res* 2000, **28**:1481-1488.

Databases of structural and functional categories were cross-referenced with 10 genome expression datasets in order to show which features are more prevalent in the transcriptome than in the genome. The transcriptome is shown to be enriched with the amino acids alanine and glycine, soluble protein folds with mixed α/β architecture and proteins involved in protein synthesis, cell structure and energy production. Conversely, proteins enriched in asparagine, membrane proteins and proteins involved in transport, transcription and signaling have lower than average expression levels. The TIM barrel is shown to be the most common fold in the transcriptome.

65. • Gerstein M: **Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census.** *Proteins* 1998, **33**:518-534.

An early analysis of the gene expression data in terms of protein folds, focusing on the diauxic shift. The most common folds based on duplication and expression are ranked. It is shown that this ranking changes at different parts in the diauxic shift experiment.

66. Bucher P: **Regulatory elements and expression profiles.** *Curr Opin Struct Biol* 1999, **9**:400-407.

67. Brazma A, Jonassen I, Vilo J, Ukkonen E: **Predicting gene regulatory elements *in silico* on a genomic scale.** *Genome Res* 1998, **8**:1202-1215.

68. Roth FP, Hughes JD, Estep PW, Church GM: **Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.** *Nat Biotechnol* 1998, **16**:939-945.

69. van Helden J, Andre B, Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies.** *J Mol Biol* 1998, **281**:827-842.

70. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262**:208-214.

71. Costanzo MC, Hogan JD, Cusick ME, Davis BP, Fancher AM, Hodges PE, Kondu P, Lengieza C, Lew-Smith JE, Lingner C *et al.*: **The yeast proteome database (YPD) and *Caenorhabditis elegans* proteome database (WormPD): comprehensive resources for the organization and comparison of model organism protein information.** *Nucleic Acids Res* 2000, **28**:73-76.

72. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28**:45-48.

73. • Drawid A, Gerstein M: A **Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome.** *J Mol Biol* 2000, in press. **[Au: Do you have any more details (volume and page** 59
range) for this paper?]

It was observed that highly expressed proteins tend to be cytosolic, whereas lowly expressed ones tend to be localized in the nucleus or the membranes. The relation between a gene's expression level and the subcellular localization of its associated protein is then used to help predict the localization of the more than 4000 yeast proteins with unknown localization. A probabilistic Bayesian formalism is used, whereby the localization of a protein is iteratively updated depending on a variety of features it has. These features include expression level and also the presence of **[Au: OK?]** traditional sequence motifs (e.g. HDEL).

74. Drawid A, Jansen R, Gerstein M: **Gene expression levels are related to protein subcellular localization.** *Trends Genet* 2000, in press.
[Au: Do you have any more details (volume and page range) for 60
this paper?]

75. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA: **Dissecting the regulatory circuitry of a eukaryotic genome.** *Cell* 1998, **95**:717-728.