BIOINFORMATICS

Vol. 18 no. 4 2002
Pages 585–596

Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts

Dov Greenbaum[3,†], Ronald Jansen[1,†] and Mark Gerstein[1,2,*]

[1]Departments of Molecular Biophysics & Biochemistry, [2]Computer Science and [3]Genetics, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA

Received on July 2, 2001; revised on October 5, 2001; accepted on October 22, 2001

ABSTRACT

Motivation: Protein abundance is related to mRNA expression through many different cellular processes. Up to now, there have been conflicting results on how correlated the levels of these two quantities are. Given that expression and abundance data are significantly more complex and noisy than the underlying genomic sequence information, it is reasonable to simplify and average them in terms of broad proteomic categories and features (e.g. functions or secondary structures), for understanding their relationship. Furthermore, it will be essential to integrate, within a common framework, the results of many varied experiments by different investigators. This will allow one to survey the characteristics of highly expressed genes and proteins.

Results: To this end, we outline a formalism for merging and scaling many different gene expression and protein abundance data sets into a comprehensive reference set, and we develop an approach for analyzing this in terms of broad categories, such as composition, function, structure and localization. As the various experiments are not always done using the same set of genes, sampling bias becomes a central issue, and our formalism is designed to explicitly show this and correct for it. We apply our formalism to the currently available gene expression and protein abundance data for yeast. Overall, we found substantial agreement between gene expression and protein abundance, in terms of the enrichment of structural and functional categories. This agreement, which was considerably greater than the simple correlation between these quantities for individual genes, reflects the way broad categories collect many individual measurements into simple, robust averages. In particular, we found that in comparison to the population of genes in the yeast genome, the cellular populations of transcripts and proteins (weighted by their respective abundances, the transcriptome and what we dub the translatome) were both enriched in: (i) the small amino acids Val, Gly, and Ala; (ii) low molecular weight proteins; (iii) helices and sheets relative to coils; (iv) cytoplasmic proteins relative to nuclear ones; and (v) proteins involved in 'protein synthesis,' 'cell structure,' and 'energy production.'

Supplementary information: http://genecensus.org/expression/translatome

Contact: mark.gerstein@yale.edu

INTRODUCTION

High throughput experimentation, measuring mRNA (Schena *et al.*, 1995; Eisen and Brown, 1999; Ferea and Brown, 1999; Lipshutz *et al.*, 1999) and protein expression (Anderson and Seilhamer, 1997; Futcher *et al.*, 1999; Gygi *et al.*, 1999a; Ross-Macdonald *et al.*, 1999; Lopez, 2000; MacBeath and Schreiber, 2000; Nelson *et al.*, 2000; Zhu *et al.*, 2000) are currently the single richest source of genomic information. However, how to best interpret this data is still an open question (Bassett *et al.*, 1996; Wittes and Friedman, 1999; Zhang, 1999; Gerstein and Jansen, 2000; Searls, 2000; Sherlock, 2000; Claverie, 1999; Einarson and Golemis, 2000; Epstein and Butow, 2000; Shapiro and Harris, 2000). Understanding how protein abundance is related to mRNA transcript levels is essential for interpreting gene expression, protein interactions, structures and functions in a cellular system (Hatzimanikatis *et al.*, 1999). Moreover, as protein concentration is the more relevant variable with respect to enzyme activity, it connects genomics to the physical chemistry of the cell (Kidd *et al.*, 2001). Protein abundance may also be invaluable for diagnostics and for determining drug targets (Corthals *et al.*, 2000).

[*]To whom correspondence should be addressed.
[†] These authors contributed equally to this work.

Previously, we surveyed the population of protein features—such as folds, amino acid composition, and functions—in yeast, and other recently sequenced genomes (Gerstein, 1997, 1998a,b; Gerstein and Hegyi, 1998; Hegyi and Gerstein, 1999; Das and Gerstein, 2000; Lin and Gerstein, 2000), and we extended this concept to compare the population of features in the yeast transcriptome to that in the genome (Drawid *et al.*, 2000; Jansen and Gerstein, 2000). Others have also done related work (Frishman and Mewes, 1997; Tatusov *et al.*, 1997; Jones, 1998; Wallin and von Heijne, 1998; Frishman and Mewes, 1999; Wolf *et al.*, 1999). Here, we present a new methodology to compare the features of the mRNA expression population with the protein abundance population.

Precise terminology is essential for this comparison. Unfortunately, 'proteome' is used inconsistently. Proteome can logically be used to describe all the distinct proteins in the genome (Qi *et al.*, 1996; Cavalcoli *et al.*, 1997; Fey *et al.*, 1997; Garrels *et al.*, 1997; Gaasterland, 1999; Jones, 1999; Sali, 1999; Tekaia *et al.*, 1999; Bairoch, 2000; Cambillau and Claverie, 2000; Doolittle, 2000; Pandey and Mann, 2000; Rubin *et al.*, 2000) and, in this context, it is equivalent to what others may refer to as the coding part of the genome. However, in papers on two-dimensional (2D) electrophoresis, it is often used to describe the sum total of proteins in a cell, taking into account the different levels of protein abundance (Shevchenko *et al.*, 1996; Gygi *et al.*, 2000a; Lopez, 2000; Washburn and Yates, 2000). In an effort to be clear, we propose the term 'translatome' for this second usage of proteome.

With this definition, we are able to refer compactly to three different cellular populations. These are illustrated in Figure 1.

(i) We use the term *genome* when we refer to the population of open reading frames, where each ORF counts once.

(ii) We use the term *transcriptome* when we refer to the population of mRNA transcripts. This term was originally coined by Velculescu *et al.* (1997). Note that each ORF may give rise to different numbers of transcripts. Consequently, the transcriptome is essentially the same as the genome but with each ORF weighted by its expression level.

(iii) The next level is the cellular population of proteins. As each protein represents a translated transcript, we make an analogy with the term transcriptome and use the term *translatome* as described above to describe this third population. Thus, the translatome is a subset of the genome where each ORF is weighted by its associated level of protein abundance.

Note that one could also less compactly call the translatome a 'weighted proteome.' However, doing so assumes one of the two aforementioned definitions of proteome. To avoid ambiguity, we studiously avoid the use of proteome altogether in the paper.

Differences between the translatome and the transcriptome exist given that transcripts from different genes can give rise to different numbers of proteins, due to different rates of translation and protein degradation. Post-transcriptional modifications further affect the translatome.

In our analysis of the transcriptome and translatome, we focus on global protein features rather than the comparison of individual genes. Previous analyses have shown that differences between mRNA expression and protein abundance levels can be quite dramatic for individual genes. This may either be due to the noise in the data or to fundamental biological processes. However, our analyses show that the variation between transcriptome and translatome is much smaller for global properties that are computed by averaging over the properties of many individual genes.

METHODS

Data sources used

For our analysis we culled many divergent data sets, representing protein abundance and mRNA expression experiments and also other sources of genome annotation. These are all summarized in Table 1.

Biases in the data

The databases that annotate the specific genes may not always be accurate (Ishii *et al.*, 2000). Gene Chip experiments suffer with regard to cross hybridization and the saturation of probes. SAGE data degrades for lowly expressed mRNAs. 2D gels are unable to resolve membrane proteins (approximately 30% of the genome) and basic proteins (Gerstein, 1998c; Krogh *et al.*, 2001). In addition, the procedures for identification and quantification of the protein spots are subject to uncertainties (Haynes and Yates, 2000). Human biases include the lack of low abundance proteins (Fey and Larsen, 2001; Gygi *et al.*, 2000b; Harry *et al.*, 2000) and the differences between laboratories in sample preparation. Our reference expression data set attempts to resolve these problems.

Data set scaling

A reference set for mRNA expression. With many different mRNA expression data sets available, it is worthwhile to integrate them into a single unified reference set, with the intention of reducing the noise and errors contained in the individual data sets and to obtain a unified estimate of the normal expression state in a cell.

We adopt an iterative scaling and merging formalism,

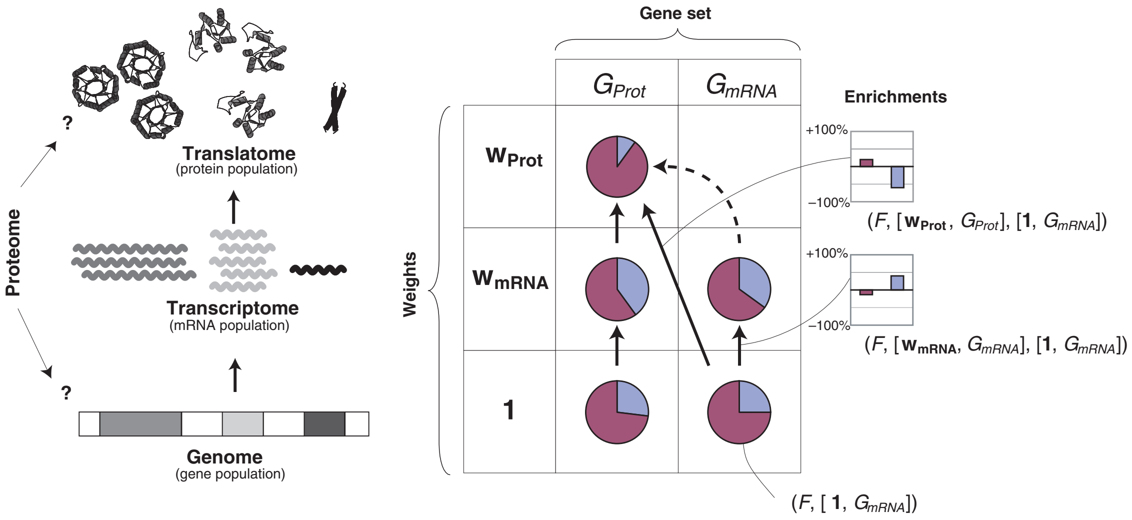
Proteome
?
Translatome
(protein population)
Transcriptome
(mRNA population)
?
Genome
(gene population)
Gene set
G_{Prot}
G_{mRNA}
Enrichments
Weights
$\mathbf{W}_{\mathbf{Prot}}$
$\mathbf{W}_{\mathbf{mRNA}}$
1
+100%
−100%
(F, [$\mathbf{W}_{\mathbf{Prot}}$, G_{Prot}], [1, G_{mRNA}])
+100%
−100%
(F, [$\mathbf{W}_{\mathbf{mRNA}}$, G_{mRNA}], [1, G_{mRNA}])
(F, [1, G_{mRNA}])

Fig. 1. Schematic overview of the analysis. On the left-side we outline the terms we use to describe the process of gene expression. The coding section of the genome is transcribed into a population of mRNA transcripts called the 'transcriptome.' The transcripts in turn are translated to a population of proteins; we use the term 'translatome' for this protein population rather than the alternative 'proteome' because the latter term may be confounded with the protein complement of the genome (which is not necessarily associated with a quantitative abundance level).
The matrix in the middle schematically shows an analysis of the three stages of expression. In general, we define a protein 'population' as a set of genes associated with a corresponding number of expression or abundance levels ('weights'). In the matrix each row represents a weight and each column a gene set. In particular, we differentiate between the mRNA reference expression set ($G_{\mathrm{mRNA}} = G_{\mathrm{Gen}}$), which essentially covers the complete genome, and the reference protein abundance set (G_{Prot}) which contains the proteins in data sets 2-DE #1 and 2-DE #2 (see Table 1) because the protein abundance set is a significantly smaller subset of the genome. By definition, this subset contains only proteins that can be identified by 2-D gel electrophoresis and is therefore biased in this sense. The enrichment figures throughout this paper, through a comparison of the right- and left-sides of this figure, show the results of the experimental biases of 2D gels on the data set.
Each pie chart represents a composition of a particular protein feature F (for instance, an amino acid composition) in a population (represented by the symbol μ). We can further look at the 'enrichment' of this feature in one population relative to another (represented by the symbol Δ, see Section '**Methods**' for an explanation of the formalism).

which we summarize below. We present a more detailed review of the methods on our web site.

We start with the values of one gene chip data set U_i where i is used throughout as a subscript to denote gene number. We then transform the values of the next Gene Chip data set X_i to Y_i with the following non-linear regression: $\min \sum_i (Y_i - U_i)^2$ with $Y_i = AX_i^B$ where A and B are the parameters of the regression. Note that two Gene Chip sets may not be defined for the same set of genes, so we have to perform the fit only over the genes common to both sets. The motivation for scaling is that the dynamic range of observed expression levels varies somewhat between different data sets, although cell types and growth conditions are very similar. Reasons for disparity may include different calibration procedures for relating fluorescence intensity to a cellular concentration (measured in copies of transcripts per cell) or different protocols for harvesting and reverse-transcribing the cellular mRNA.

We then merge and average the data to create a new reference set V as follows:

$$\text{If } U_i \text{ and } Y_i \text{ are both defined for gene } i \text{ and } \frac{|Y_i - U_i|}{Y_i + U_i} < \alpha$$

Then $V_i = \frac{1}{2}(Y_i + U_i)$
Else if only Y_i exists, $V_i = Y_i$
Else $V_i = U_i$.

As presented above, where only one data set has a value for the corresponding ORF, we incorporated that value and did not exclude it. When both data sets have values for an ORF, we averaged the values if they were within 15% of each other; otherwise, we just stayed with the original chip data set U_i. We used $\alpha = 15\%$ in order to prevent outliers from skewing the result. This 15% value is a reasonable threshold for excluding outliers though other values (e.g. 10 or 20%) would give similar results (data not shown). Other data sets are subsequently included in the same procedure, continuing the iteration from the new expression values V_i. The initial iteration starts with the Young Expression Set, as U_i, since we have the highest confidence in its accuracy.

The SAGE data (Velculescu *et al.*, 1997) was not included in the above procedure since it is of a fundamentally different nature. An advantage of the SAGE technology over Gene Chips is that there is no possible signal saturation for high expression levels, as is possible for chips (Futcher *et al.*, 1999). Conversely, SAGE values are less reliable for lowly expressed genes since there is a chance that one might not sequence a SAGE tag corresponding to such a gene altogether. Therefore, if after the last iteration, the average Gene Chip expression level V_i was both above a certain threshold β and below the SAGE expression level S_i for the same gene, it was replaced with the SAGE value; otherwise the average Gene Chip value was kept. This gave us our final expression set $\mathbf{w_{mRNA}}$. Our treatment of the SAGE data is modeled after that in Futcher *et al.* (1999), and like them, we used $\beta = 16$.

This incorporation of the SAGE data into the reference data set ensures that the highly expressed outliers are as accurate as possible.

Rather than plain arithmetic averaging, this overall scaling procedure with the α cutoff avoids 'artificial averages' that combine very different values for a particular gene. Some expression values might be statistical outliers. In addition, it may be possible that the expression levels of a variety of genes can only be within mutually exclusive ranges or modes, such as when two alternative pathways are switched on or off. Simply averaging these would give values that are less representative of the particular mode values. This situation is analogous to that in averaging together an ensemble of protein structures (i.e from NMR structure determination). Each structure could be stereochemically correct, with all side-chain atoms in predefined rotamer configurations. However, an average of all structures could yield one that is stereochemically incorrect if this involved averaging over particular side-chains in different rotameric states.

With regard to our regression analysis, we have investigated both non-linear and linear fits but found a non-linear procedure to be more advantageous. The non-linear relationship between different expression data sets perhaps reflects saturation in one or more of the Gene Chips—not an uncommon phenomenon. This non-linearity is immediately evident on scatter plots of two data sets against one another (see website). Accordingly, the non-linear fit produces a smaller residual than the linear fit: 98 297 (non-linear) versus 122 182 (linear) for the scaling of the Church data set and 59 828 (non-linear) versus 67 462 (linear) for the Samson data set.

A reference set for protein abundance. We followed a similar procedure to calculate a reference protein abundance set from the two gel electrophoresis data sets. We first scaled the two data sets against the mRNA expression reference data set, getting regression parameters C_j and D_j:

$$\min \sum_i (P_{i,j} - C_j w_{\mathrm{mRNA},i}^{D_j})^2$$

where the subscript j indicates the data set 2-DE #1 or 2-DE #2 respectively; $P_{i,j}$ is the protein abundance value in data set j, and $w_{\mathrm{mRNA},i}$ the corresponding reference expression value, and C_j and D_j are the parameters of the non-linear regression.

Using these parameters, we transformed the values of set 2-DE #2 onto 2-DE #1. Then we combined both sets into the reference protein set $\mathbf{w_{Prot}}$ by averaging them, if both values existed. Otherwise, by using the existing value, viz:

$$Q_{i,2} \equiv C_1 \left(\frac{P_{i,2}}{C_2} \right)^{D_1/D_2}$$

$w_{\mathrm{Prot},i} = (P_{i,1} + Q_{i,2})/2$ if both $P_{i,1}$ and $Q_{i,2}$ exist.

Else if only $P_{i,1}$ exists, $w_{\mathrm{Prot},i} = P_{i,1}$

Else if $Q_{i,2}$ exists, $w_{\mathrm{Prot},i} = Q_{i,2}$.

Enrichment of features

Formalism. In the next part of our analysis, we want to group a number of proteins together into various categories based on common features and characterize those features that are enriched in one population relative to another, i.e. the translatome population of proteins as measured by 2D gels relative to the transcriptome population of transcripts or the genome population of genes. To this end, we set up a formalism that could be applied universally to all the attributes that we were interested in. Due to the limitations of the experiments, the translatome, transcriptome, and genome populations are defined on different sets of genes, and sometimes we want to remove this 'selection bias' by forcing them to be compared on exactly the same set of genes. This is a key aspect of our formalism as presented in Figure 1.

We call an entity like [$\mathbf{w}$, G] a 'population,' where G is a set describing a particular selection of genes from the genome and $\mathbf{w}$ is vector of weights associated with each element of this population. In particular, we focus on three main populations here:

(i) [$\mathbf{1}$, G_{Gen}] is the population of genes in the genome, all 6280 genes weighted once ($\mathbf{w} = \mathbf{1}$);

(ii) [$\mathbf{w_{mRNA}}$, G_{mRNA}] is the observed population of the transcripts in the transcriptome, i.e. the 6249 genes in the reference expression set weighted by their reference expression value;

(iii) $[\mathbf{w}_{\mathbf{Prot}}, G_{\text{Prot}}]$ is the observed cellular population of the proteins in the translatome, i.e. the 181 genes in the reference abundance set weighted by their reference abundance value.

(The set of genes in the genome G_{Gen} is approximately equal to the genes in set G_{mRNA}, such that we can use both symbols interchangeably.) We can also use this notation to describe specific experiments—e.g. $[\mathbf{w}_{\mathbf{lacZ}}, G_{\text{lacZ}}]$ describes the gene set and weights relating to the transposon abundance set.

Furthermore, we define F_j as the value of a feature F in ORF j. For example, F could be the composition of leucine (a real number) or a binary value (0 or 1) indicating whether an ORF contains a trans-membrane segment. Given these definitions, the weighted average of feature F in population $[\mathbf{w}, G]$ is:

$$\mu(F, [\mathbf{w}, G]) \equiv \frac{\sum_{j\in G} w_j F_j}{\sum_{j\in G} w_j}.$$

The weighted averages of two populations $[\mathbf{w}, G]$ and $[\mathbf{v}, S]$ can be compared by simply looking at their relative difference Δ:

$$\Delta(F, [\mathbf{v}, S], [\mathbf{w}, G]) = \frac{\mu(F, [\mathbf{v}, S]) - \mu(F, [\mathbf{w}, G])}{\mu(F, [\mathbf{w}, G])}$$

where $\mathbf{v}$ and $\mathbf{w}$ are weights for the sets of ORFs S and G respectively. We call Δ the 'enrichment' of feature F because it indicates whether F is enriched (if Δ is positive) or depleted (if Δ is negative) in population $[\mathbf{v}, S]$ relative to $[\mathbf{w}, G]$.

Usually, the gene set G is defined by the particular experiment, for which the weight $\mathbf{w}$ was measured. However, it is also possible to combine the gene set associated with one experiment with expression levels from another set. One may want to do this to compute the enrichment only on the genes common to both populations, for which there are defined values for both $\mathbf{w}$ and $\mathbf{v}$, viz: $\Delta(F, [v, S \cap G], [w, S \cap G])$. In practice, this is most relevant for comparing G_{Prot} and G_{mRNA}. Since G_{Prot} is completely a subset of G_{mRNA}, we need not explicitly deal with intersections if we calculate all statistics directly over G_{Prot}.

One can adjust the weight vectors to take into account different types of averaging. For instance, when computing the amino acid composition ($F = aa$) from the amino acid compositions of individual ORFs $F_j = aa_j$ ($\forall j \in G$), we weight by ORF length. In the case of expression weights, we have:

$$w_j = N_j w_{\text{mRNA},j} \quad \forall j \in G$$

where N_j is a measure of the length of ORF j (such as the number of amino acids).

On the other hand, when computing the average molecular weight per amino acid, we need to normalize by the number of amino acids per ORF, which is equivalent to choosing the following weights:

$$w_j = \frac{w_{\text{mRNA},j}}{N_j} \quad \forall j \in G.$$

Application of methodology to quantitative abundance sets

Having defined our formalism, we applied it to a diverse set of protein features in yeast.

Amino acid enrichment. As shown in Figure 2a, we used our methodology to measure the enrichment of individual amino acids in both the translatome and the transcriptome relative to the genome. We found that three amino acids—valine, glycine and alanine—were consistently enriched in both transcriptome and translatome populations.

In Figure 2a we compare different gene sets. In Figure 2b we focus mainly on the variation in enrichments when all the comparisons are restricted to the set of 181 genes ($G_{\text{Prot}} \cap G_{\text{mRNA}} = G_{\text{Prot}}$) common to all data sets. Thus, the differences between the populations now only reflect the effects of differential transcription of certain genes and differential translation of certain transcripts. We find here an enrichment specifically of cysteine in the translatome in relation to the transcriptome.

To measure the statistical significance of the results on amino acid enrichment, we have performed a control analysis on a randomized data set (Figure 2d). We randomly permutated the expression values of the ORFs 1000 times and then recomputed the enrichments. This allowed us to compute distributions for the amino acid enrichments and, from integrating these, one-sided p-values indicating the significance of the observed enrichments.

Amino acid enrichment in Transposon data set. We also tried to extend our methodology, ineffectively, to cope with the semi-quantitative Transposon set. We used only those 450 ORFs that consistently yielded either no expression or high expression, as binary data, on or off. We show the enrichments of amino acids computed from this filtered Transposon abundance set in Figure 2a. Overall, the enrichments from this set seemed to be attenuated in comparison to other data.

Biomass enrichment. A corollary to amino acid enrichments is the determination of the average biomass of the transcriptome and translatome populations (shown in Figure 2c). We found that the average molecular weight of a protein in both populations was, on average, lower than in the genome population. These preliminary observations suggest a cell preference to use less energetically expensive proteins for those that are highly transcribed or trans-

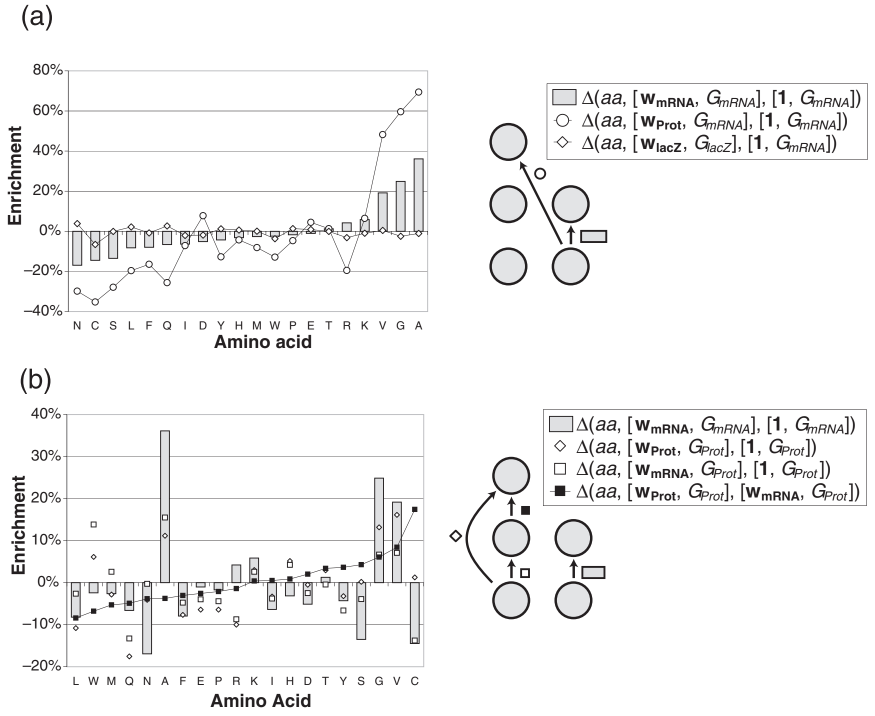

Fig. 2. Amino acid and biomass enrichment. (a) Shows the amino acid enrichments between different populations as indicated by the legend to the right of the plot (the legend is ordered in the same way as the schematic illustration in Figure 1). The bars indicate the enrichment of the transcriptome relative to the genome, whereas the circles indicate the enrichment of the translatome relative to the genome. In addition, we also show the enrichment for protein abundance from the Transposon abundance set, represented by the circles with the line through them. (b) Shows a different view of amino acid enrichment from that contained in (a), now focusing on changes, and thus restricting the comparison to the genes common to all the data sets. The graph is ordered according to the enrichment from transcriptome to translatome (black squares). We focus here only on the changes for the abundance gene set (G_{Prot}) to exclude the effects that arise from looking at different subsets. In this view the enrichments from genome to transcriptome (white squares) and from genome to translatome (white diamonds) look more similar than do the analogous sets in (a). To make comparison with (a) easier we again show the enrichment from genome to the transcriptome for the complete gene set (G_{Gen}, shown in bars). (c) Shows biomass enrichment. The left panel depicts the average molecular weight per ORF (in units of kDa) and the right panel, the average molecular weight per amino acid (in units of Daltons) in each of the three stages of gene expression. The numbers inside the circles indicate the average molecular weights. The values next to the arrows indicate the enrichments in biomass between different populations. Both the circle diameters and the arrow widths are functions of the corresponding values (the hollow arrow indicates a positive value). It is very clear that the average molecular weight per ORF is much lower in the translatome (by 20 or 15%) and transcriptome (by 29%) than in the genome. This relative depletion of biomass mainly takes place as a result of transcription; the effect of translation is less clear, depending on the populations compared. On the other hand, the depletion in the average molecular weight per amino acid (−3.3% from genome to translatome) is an order of magnitude smaller than in the average weight per ORF. This shows that the yeast cell favors the expression of shorter ORFs over longer ones, and agrees with our earlier observation that there is a negative correlation between maximum ORF length and mRNA expression (Jansen and Gerstein, 2000); it seems that this effect mainly takes place during transcription rather than translation. (d) This plot shows that the amino acid enrichments are statistically significant. We have assessed significance by randomly permuting the expression levels among the genes and then recomputing the amino acid enrichments. This procedure can be repeated and used to generate distributions of random enrichments that can then be compared against the observed enrichments. In the plot the gray bars represent the observed enrichments already shown in Figure 3a. On top of the gray bars we show standard boxplots of enrichment distributions based on 1000 random permutations. (The middle line represents the distribution median. The upper and lower sides of the box coincide with the upper and lower quartiles. Outliers are shown as dots and defined as data points that are outside the range of the whiskers, the length of which is 1.5 the interquartile distance.) Based on the random distributions, we can compute one-sided p-values for the observed enrichments. Amino acids for which the p-values are less than 10^{-3} are shown in bold font.

lated. However, we also found that the average molecular weight *per amino acid* differed much less between the transcriptome and the translatome on the one hand, and the genome on the other hand (though it was still slightly less). This finding indicates that lower molecular weights in the translatome and transcriptome relative to the genome are

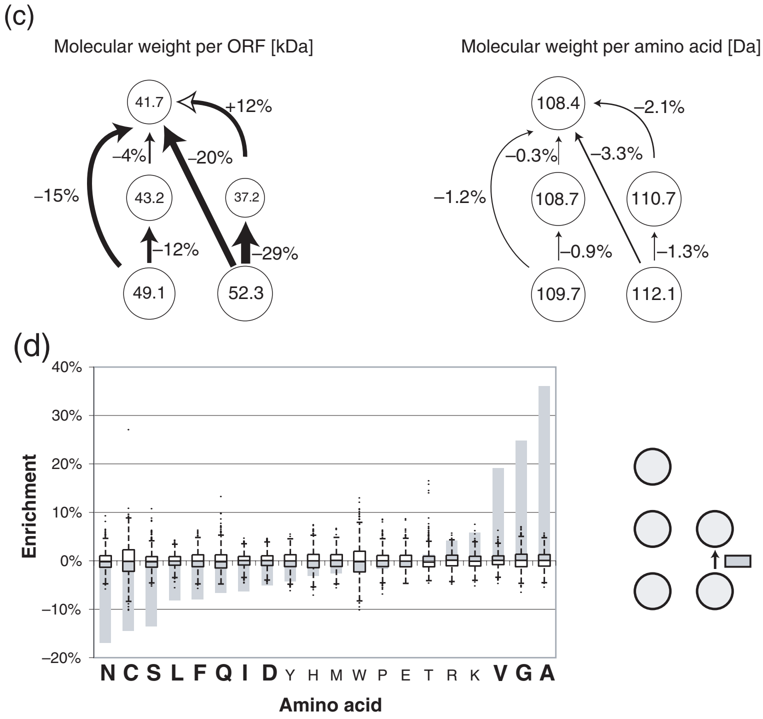

Fig. 2. Cont.

predominantly due to greater expression of shorter proteins rather than the incorporation of smaller amino acids.

Secondary structure composition. We also used our methodology to study the enrichment of secondary-structural features. Secondary structural annotation was derived from structure prediction applied uniformly to all the ORFs in the yeast genome as described in Table 1. As shown in Figure 3a, all three populations—genome, transcriptome, and translatome—had a fairly similar composition of secondary structures—sheets, helices, and coils. The differences between populations were marginal and based only on the small subset of genes.

We also found that Transmembrane (TM) proteins were significantly depleted in the transcriptome (see website and caption). These results are consistent with our previous analyses (Jansen and Gerstein, 2000). The protein abundance data does not have any membrane proteins.

Subcellular localization. Figure 3c shows the enrichment of proteins associated with the various subcellular compartments. For clarity, we divided the cell into five distinct subcellular compartments, (see Table 1). We found that, in comparison to the genome, both the transcriptome and translatome are enriched in cytoplasmic proteins. This is true whether we make our comparisons in relation to the relatively large reference mRNA expression set or the smaller reference protein abundance set. As Figure 3c shows, the 2D gel experiments are clearly biased towards proteins from the cytoplasm. However, in the biased subset G_{Prot} transcription and translation lead to an even higher fraction of cytoplasmic proteins in the translatome.

Functional categories. Finally, we compared the enrichment of various functional categories in both the translatome and the transcriptome (see Figure 3b). This gives us a broad yet informative view of the cell as a whole. As described in Table 1, we used the top-level of the MIPS scheme for the functional category definitions. We found broad differences between the various populations, with some of the functional categories showing strikingly high enrichments.

DISCUSSION AND CONCLUSION

We developed: (i) a methodology for integrating many different types of gene expression and protein abundance into a common framework and applied this to a preliminary analysis; (ii) a procedure for scaling and merging different mRNA and protein sets together; and (iii) an approach for computing the enrichment of various proteomic features in the population of transcripts and proteins. We showed that by analyzing broad categories instead of individual noisy

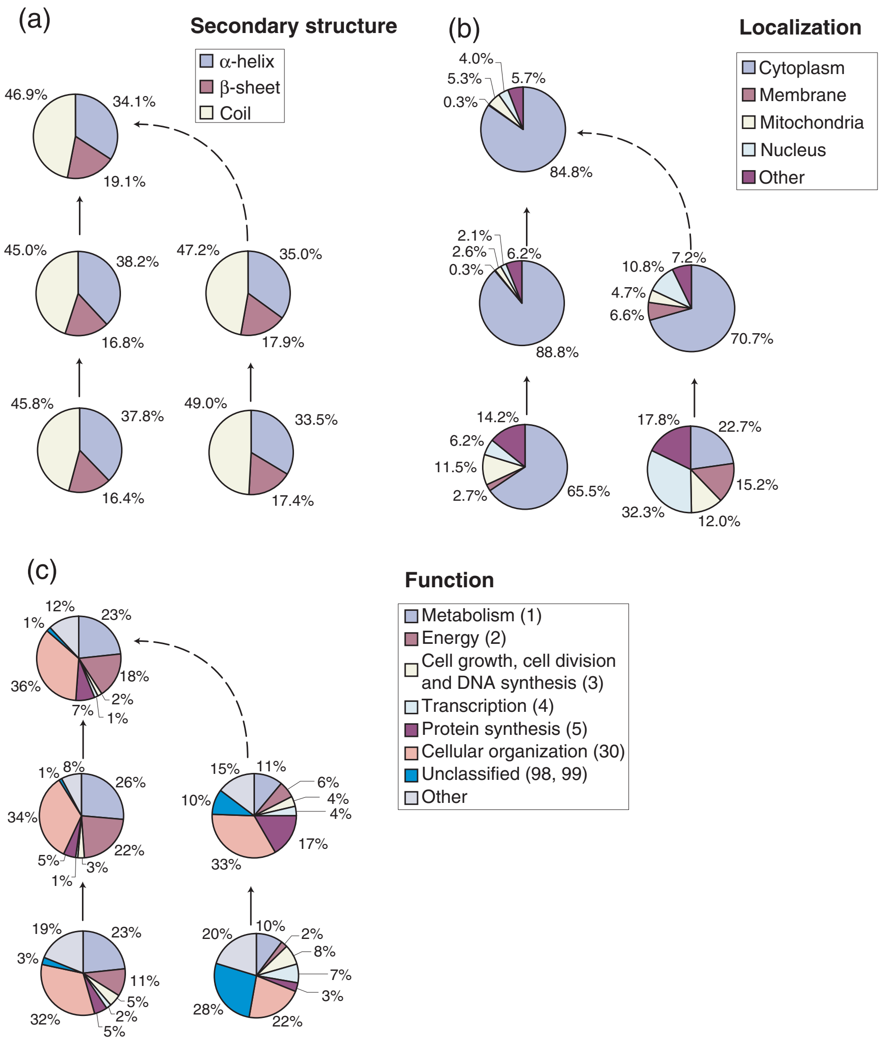

Fig. 3. Breakdown of the transcriptome and translatome in terms of broad categories relating to structure, localization, and function. All of the subfigures are analogous to the schematic illustration in Figure 1. (a) Represents the composition of secondary structure in the different populations. (b) Represents the distribution of subcellular localizations associated with proteins in the various populations. We used standardized localizations developed earlier (Drawid and Gerstein, 2000), which, in turn, were derived from the MIPS, YPD, and SwissProt databases (Bairoch and Apweiler, 2000; Costanzo *et al.*, 2000; Mewes *et al.*, 2000). The subcellular localization has been experimentally determined for less than half of the yeast proteins, so our analysis applies only to this subset. (c) Shows the division of ORFs into different functional categories (according to the MIPS classification) in the various populations. Only the largest functional categories of the top level of the MIPS classification are shown. The group 'other' contains the smaller top-level categories lumped together. This 'other' group is different from the group 'unclassified,' which contains genes without any functional description.

data points, we could find logical trends in the underlying data. For example, individual transcription factors might have higher or lower protein abundance than one expects from their mRNA expression, but the category 'transcription factors' as a whole has a similar representation in the transcriptome and translatome.

We found, as previously described (Futcher *et al.*, 1999; Gygi *et al.*, 1999b; Greenbaum *et al.*, 2001), a weak correlation between individual measurements of mRNA and protein abundance. The outliers of this correlation tend to be associated with cellular organization. One might conceive of using these outliers (i.e. those with significantly different transcriptional and translational behavior) to find consensus regulatory sequences. One possible method would involve using predicted mRNA structures (Jaeger *et al.*, 1990; Zuker, 2000) to find and investigate consensus structural elements in these outliers to which the yeast translational machinery is known to be

Table 1. Data sets

Data set	Description	Size [ORFs]	Reference
mRNA expression			
Young	Gene chip profiles yeast cells with mutations that affect transcription	5455	Holstege *et al.* (1998)
Church	Gene chip profiles of yeast cells under four different conditions	6263	Roth *et al.* (1998)
Samson	Comparing gene chip profiles for yeast cells subjected to alkylating agent	6090	Jelinsky and Samson (1999)
SAGE	Yeast cells during vegetative growth	3778	Velculescu *et al.* (1997)
Reference expression	Scaling and integrating the mRNA expression set into one data source	6249	–
Protein abundance			
2-DE #1	Measurement of yeast protein abundance by 2D gel electrophoresis and mass spectrometry	156	Gygi *et al.* (1999a,b)
2-DE #2	Similar to 2-DE set #1	71	Futcher *et al.* (1999)
Transposon	Large-scale fusions of yeast genes with *lacZ* by transposon insertion	1410	Ross-Macdonald *et al.* (1999)
Reference abundance	Scaling and integrating the 2-DE data sets into one data source	181	–
Annotation			
Annotated localization	Subcellular localizations of yeast proteins	2133 (6280)	Drawid and Gerstein (2000)
TM segments	Predicted TM and soluble proteins in yeast	2710 (6280)	Gerstein (1998a,b,c)
MIPS functions	Functional categories for yeast ORFs	3519 (6194)	Mewes *et al.* (2000)
GOR secondary structure	Predicted secondary structure yeast ORFs	6280	Gerstein (1998a,b,c)

This table provides an overview of the data sets used in our analysis. The table is divided into three sections. The top section lists different mRNA expression sets. The middle section shows the protein abundance data sets used. The bottom section contains different annotations of protein features. The column 'Data set' lists a shorthand reference to each data set used throughout this paper. The next columns contain a brief description of the data sets, the number of ORFs contained in each of them, and the literature reference. In contrast to the other data we investigated, the reference expression and abundance data sets have been calculated for the purpose of our analysis (see text). An expanded version of the table is available on our web site.

Some further information on the genome annotations:

Localization. Protein localization information from YPD, MIPS and SwissProt were merged, filtered and standardized (Bairoch and Apweiler, 2000; Costanzo *et al.*, 2000; Mewes *et al.*, 2000) into five simplified compartments—cytoplasm, nucleus, membrane, extracellular (including proteins in ER and golgi), and mitochondrial—according to the protocol in Drawid *et al.* (2000). This yielded a standardized annotation of protein subcellular localization for 2133 out of 6280 ORFs.

TM segments. In 2710 out of 6280 yeast ORFs TM segments are predicted to occur, ranging from low to high confidence (732 ORFs). The TM prediction was performed as follows: the values from the scale for amino acids in a window of size 20 (the typical size of a TM helix) were averaged and then compared against a cutoff of -1 kcal mol^{-1}. A value under this cutoff was taken to indicate the existence of a TM helix. Initial hydrophobic stretches corresponding to signal sequences for membrane insertion were excluded. (These have the pattern of a charged residue within the first seven, followed by a stretch of 14 with an average hydrophobicity under the cutoff.) These parameters have been used, tested, and refined on surveys of membrane protein in genomes. 'Sure' membrane proteins had at least two TM-segments with an average hydrophobicity less than -2 kcal mol^{-1} (Rost *et al.*, 1995; Gerstein *et al.*, 2000; Santoni *et al.*, 2000; Senes *et al.*, 2000).

Functions. MIPS functional categories have been assigned to 3519 out of 6194 ORFs. (The remainder are assigned to category '98' or '99,' which corresponds to unclassified function.)

sensitive (McCarthy, 1998).

In relation to functional categories, we found three trends that were particularly notable: (i) the 'cellular organization,' 'protein synthesis,' and 'energy production' categories were increasingly enriched as we moved from genome to transcriptome to translatome. In the transcriptome and translatome population relative to the genome; (ii) proteins with 'unclassified function' are significantly depleted, perhaps reflecting a bias against studying them; (iii) proteins in the 'transcription' and 'cell growth, cell division, and DNA synthesis' categories were consistently depleted. This reflects the fact that many of these proteins, such as transcription factors, act as 'switches' such that only small quantities of the protein are necessary to activate or deactivate a process. These results concur with previous calculations (Jansen and Gerstein, 2000) wherein we found the transcriptome is enriched specifically with proteins involved in protein synthesis and energy.

Limitations given the small size of the protein abundance data

Even with the extended coverage made possible by merging many data sets together into reference sets, the analysis is still limited by the minimal data. This was most applicable to the protein abundance measurements, potentially biasing our statistical results towards certain protein families. Moreover, the 181 proteins in G_{Prot} do not represent a random sample. They are skewed towards highly expressed, well-studied proteins. Our methodology attempts to control for this gene-selection bias through our enrichment formalism, which allows one to rather precisely gauge various aspects of the bias. Conversely, many protein features in both the translatome and the transcriptome are dominated by highly expressed proteins. Under these circumstances, it is often sufficient to look at this smaller number of dominating proteins to characterize the whole population. This is similar to the development of the codon adaptation index for yeast (Sharp and Li, 1987). While based on only 24 highly expressed proteins, it has proven to be robust in predicting expression levels for the entire genome.

We believe that the essential formalism and approach that we develop will remain quite relevant for future data sets (Smith, 2000).

ACKNOWLEDGEMENT

M.G. thanks the Keck foundation for support.

REFERENCES

An,H., Scopes,R.K. *et al.* (1991) Gel electrophoretic analysis of *Zymomonas mobilis* glycolytic and fermentative enzymes: identification of alcohol dehydrogenase II as a stress protein. *J. Bacteriol.*, **173**, 5975–5982.

Anderson,L. and Seilhamer,J. (1997) A comparison of selected mRNA and protein abundances in human liver. *Electrophoresis*, **18**, 533–537.

Bairoch,A. (2000) Serendipity in bioinformatics, the tribulations of a Swiss bioinformatician through exciting times! *Bioinformatics*, **16**, 48–64.

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

Bassett,D.E. Jr., Basrai,M.A. *et al.* (1996) Exploiting the complete yeast genome sequence. *Curr. Opin. Genet. Dev.*, **6**, 763–766.

Batke,J., Benito,V.A. *et al.* (1992) A possible *in vivo* mechanism of intermediate transfer by glycolytic enzyme complexes: steady state fluorescence anisotropy analysis of an enzyme complex formation. *Arch. Biochem. Biophys.*, **296**, 654–659.

Cambillau,C. and Claverie,J.M. (2000) Structural and genomic correlates of hyperthermostability. *J. Biol. Chem.*, **275**, 32 383–32 386.

Cavalcoli,J.D., VanBogelen,R.A. *et al.* (1997) Unique identification of proteins from small genome organisms: theoretical feasibility of high throughput proteome analysis. *Electrophoresis*, **18**, 2703–2708.

Claverie,J.M. (1999) Computational methods for the identification of differential and coordinated gene expression [in process citation]. *Hum. Mol. Genet.*, **8**, 1821–1832.

Corthals,G., Wasinger,V.C., Hochstrasser,D.F. and Sanchez,J.C. (2000) The dynamic range of protein expression: a challenge for proteomic research. *Electrophoreisis*, **21**, 1104–1115.

Costanzo,M.C., Hogan,J.D. *et al.* (2000) 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.*, **28**, 73–76.

Das,R. and Gerstein,M. (2000) The stability of thermophilic proteins: a study based on comprehensive genome comparison. *Funct. Int. Genom.*, **1**, 33–45.

Doolittle,W.F. (2000) The nature of the universal ancestor and the evolution of the proteome. *Curr. Opin. Struct. Biol.*, **10**, 355–358.

Drawid,A. and Gerstein,M. (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J. Mol. Biol.*, **301**, 1059–1075.

Drawid,A., Jansen,R. *et al.* (2000) Gene expression levels are correlated with protein subcellular localization. *Trends Genet.*, **10**, 426–430.

Einarson,M. and Golemis,E. (2000) Encroaching genomics: adapting large-scale science to small academic laboratories. *Physiol. Genom.*, **2**, 85–92.

Eisen,M.B. and Brown,P.O. (1999) DNA arrays for analysis of gene expression. *Meth. Enzymol.*, **303**, 179–205.

Epstein,C. and Butow,R. (2000) Microarray technology—enhanced versatility, persistent challenge. *Curr. Opin. Biotechnol.*, **11**, 36–41.

Ferea,T. and Brown,P. (1999) Observing the living genome. *Curr. Opin. Genet. Dev.*, **9**, 715–722.

Fey,S.J., Nawrocki,A. *et al.* (1997) Proteome analysis of *Saccharomyces cerevisiae*: a methodological outline. *Electrophoresis*, **18**, 1361–72.

Fey,S.J. and Larsen,P.M. (2001) 2D or not 2D. Two-dimensional gel electrophoresis. *Curr. Opin. Chem. Biol.*, **5**, 26–33.

Frishman,D. and Mewes,H.W. (1997) Protein structural classes in five complete genomes [letter]. *Nat. Struct. Biol.*, **4**, 626–628.

Frishman,D. and Mewes,H.W. (1999) Genome-based structural biology. *Prog. Biophys. Mol. Biol.*, **72**, 1–17.

Futcher,B., Latter,G. *et al.* (1999) A sampling of the yeast proteome. *Mol. Cell Biol.*, **19**, 7357–7368.

Gaasterland,T. (1999) Archaeal genomics. *Curr. Opin. Microbiol.*, **2**, 542–547.

Garrels,J.I., McLaughlin,C.S. *et al.* (1997) Proteome studies of *Saccharomyces cerevisiae*: identification and characterization of abundant proteins. *Electrophoresis*, **18**, 1347–1360.

Gerstein,M. (1997) A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure. *J. Mol. Biol.*, **274**, 562–576.

Gerstein,M. (1998a) How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold. Des.*, **3**, 497–512.

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

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

Gerstein,M. and Hegyi,H. (1998) Comparing genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol. Rev.*, **22**, 277–304.

Gerstein,M. and Jansen,R. (2000) The current excitement in bioinformatics, analysis of whole-genome expression data: how does it relate to protein structure and function. *Curr. Opin. Struct. Biol.*, **10**, 574–584.

Gerstein,M., Lin,J. *et al.* (2000) Protein folds in the worm genome. *Pac. Symp. Biocomput.*, 30–41.

Greenbaum,D., Luscombe,N. *et al.* (2001) Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. *Genome Res.*, **11**, 1463–1468.

Gygi,S.P., Rist,B. *et al.* (1999a) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnol.*, **17**, 994–999.

Gygi,S.P., Rochon,Y. *et al.* (1999b) Correlation between protein and mRNA abundance in yeast. *Mol. Cell Biol.*, **19**, 1720–1730.

Gygi,S.P., Corthals,G.L. *et al.* (2000a) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl Acad. Sci. USA*, **97**, 9390–9395.

Gygi,S.P., Rist,B. *et al.* (2000b) Measuring gene expression by quantitative proteome analysis. *Curr. Opin. Biotechnol.*, **11**, 396–401.

Harry,J.L., Wilkins,M.R. *et al.* (2000) Proteomics: capacity versus utility. *Electrophoreisis*, **21**, 1071–1081.

Hatzimanikatis,V., Choe,L.H. *et al.* (1999) Proteomics: theoretical and experimental considerations. *Biotechnol. Prog.*, **15**, 312–318.

Haynes,P.A. and Yates,J.R. (2000) Proteome profiling-pitfalls and progress. *Yeast*, **17**, 81–87.

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

Holstege,F.C., Jennings,E.G. *et al.* (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell*, **95**, 717–728.

Ishii,M., Hashimoto,S. *et al.* (2000) Direct comparison of genechip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics*, **68**, 136–143.

Ito,T., Tashiro,K. *et al.* (2000) Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. *Proc. Natl Acad. Sci. USA*, **97**, 1143–1147.

Jaeger,J.A., Turner,D.H. *et al.* (1990) Predicting optimal and suboptimal secondary structure for RNA. *Meth. Enzymol.*, **183**, 281–306.

Jansen,R. and Gerstein,M. (2000) Analysis of the yeast transcriptome with structural and functional categories: characterizing highly expressed proteins. *Nucleic Acids Res.*, **28**, 1481–1488.

Jelinsky,S.A. and Samson,L.D. (1999) Global response of *Saccharomyces cerevisiae* to an alkylating agent. *Proc. Natl Acad. Sci. USA*, **96**, 1486–1491.

Jones,D.T. (1998) Do transmembrane protein superfolds exist? *FEBS Lett.*, **423**, 281–285.

Jones,D.T. (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.*, **287**, 797–815.

Kidd,D. *et al.* (2001) Profiling serine hydrolase activities in complex proteomes. *Biochemistry*, **40**, 4005–4015.

Klose,J. (1975) Protein mapping by combined isoelectric focusing and electrophoresis of mouse tissues. A novel approach to testing for induced point mutations in mammals. *Humangenetik*, **26**, 231–243.

Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Lin,J. and Gerstein,M. (2000) Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res.*, **10**, 808–818.

Lipshutz,R.F. S., Gingeras,T.R. and Lockhart,D.J. (1999) High density synthetic oligonucleotide arrays. *Nature Genet.*, **21**, 20–24.

Lopez,M.F. (2000) Better approaches to finding the needle in a haystack: optimizing proteome analysis through automation. *Electrophoreisis*, **21**, 1082–1093.

MacBeath,G. and Schreiber,S.L. (2000) Printing proteins as microarrays for high-throughput function determination. *Science*, **289**, 1760–1763.

Matton,D.P., Constabel,P. *et al.* (1990) Alcohol dehydrogenase gene expression in potato following elicitor and stress treatment. *Plant Mol. Biol.*, **14**, 775–783.

McCarthy,J.E. (1998) Posttranscriptional control of gene expression in yeast. *Microbiol. Mol. Biol. Rev.*, **62**, 1492–1553.

Mewes,H.W., Frishman,D. *et al.* (2000) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **28**, 27–40.

Millar,A.A., Olive,M.R. *et al.* (1994) The expression and anaerobic induction of alcohol dehydrogenase in cotton. *Biochem. Genet.*, **32**, 279–300.

Molloy,M.P. (2000) Two-dimensional electrophoresis of membrane proteins using immobilized pH gradients. *Anal. Biochem.*, **280**, 1–10.

Nauchitel,V.V. and Somorjai,R.L. (1994) Spatial and free energy distribution patterns of amino acid residues in water soluble proteins. *Biophys. Chem.*, **51**, 327–336.

Nelson,R.W., Nedelkov,D. *et al.* (2000) Biosensor chip mass spectrometry: a chip-based proteomics approach. *Electrophoresis*, **21**, 1155–1163.

O'Farrell,P.H. (1975) High resolution two-dimensional electrophoresis of proteins. *J. Biol. Chem.*, **250**, 4007–4021.

Pandey,A. and Mann,M. (2000) Proteomics to study genes and genomes. *Nature*, **405**, 837–846.

Qi,S.Y., Moir,A. *et al.* (1996) Proteome of *Salmonella typhimurium* SL1344: identification of novel abundant cell envelope proteins and assignment to a two-dimensional reference map. *J. Bacteriol.*, **178**, 5032–5038.

Ross-Macdonald,P., Coelho,P.S. *et al.* (1999) Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature*, **402**, 413–418.

Rost,B., Casadio,R. *et al.* (1995) Transmembrane helices predicted at 95% accuracy. *Protein Sci.*, **4**, 521–533.

Roth,F.P., Hughes,J.D. *et al.* (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.*, **16**, 939–945.

Rubin,G.M., Yandell,M.D. *et al.* (2000) Comparative genomics of the eukaryotes. *Science*, **287**, 2204–2215.

Sali,A. (1999) Functional links between proteins. *Nature*, **402**, 25–26.

Santoni,V., Molloy,M. *et al.* (2000) Membrane proteins and proteomics: un amour impossible? *Electrophoreisis*, **21**, 1054–1070.

Schena,M., Shalon,D. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

Searls,D.B. (2000) Using bioinformatics in gene and drug discovery. *Drug Discov. Today*, **5**, 135–143.

Senes,A., Gerstein,M. *et al.* (2000) Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J. Mol. Biol.*, **296**, 921–936.

Shapiro,L. and Harris,T. (2000) Finding function through structural genomics. *Curr. Opin. Biotechnol.*, **11**, 31–35.

Sharp,P.M. and Li,W.H. (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.

Sherlock,G. (2000) Analysis of large-scale gene expression data. *Curr. Opin. Immunol.*, **12**, 201–205.

Shevchenko,A., Jensen,O.N. *et al.* (1996) Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl Acad. Sci. USA*, **93**, 14 440–14 445.

Smith,R.D. (2000) Probing proteomes-seeing the whole picture? *Nature Biotechnol.*, **18**, 1041–1042.

Tatusov,R.L., Koonin,E.V. *et al.* (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Tekaia,F., Lazcano,A. *et al.* (1999) The genomic tree as revealed from whole proteome comparisons. *Genome Res.*, **9**, 550–557.

Velculescu,V.E., Zhang,L. *et al.* (1997) Characterization of the yeast transcriptome. *Cell*, **88**, 243–251.

Wallin,E. and von Heijne,G. (1998) Genome-wide analysis of integral membrane proteins from eubacterial, archaean, and eukaryotic organisms. *Protein Sci.*, **7**, 1029–1038.

Washburn,M.P., Wolters,D. *et al.* (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnol.*, **19**, 242–247.

Washburn,M.P. and Yates,J.R. 3rd (2000) Analysis of the microbial proteome. *Curr. Opin. Microbiol.*, **3**, 292–297.

Wittes,J. and Friedman,H.P. (1999) Searching for evidence of altered gene expression: a comment on statistical analysis of microarray data [editorial; comment]. *J. Natl Cancer. Inst.*, **91**, 400–401.

Wolf,Y.I., Brenner,S.E. *et al.* (1999) Distribution of protein folds in the three superkingdoms of life. *Genome Res.*, **9**, 17–26.

Young,K.H. (1998) Yeast two-hybrid: so many interactions, (in) so little time . . . . *Biol. Reprod.*, **58**, 302–311.

Zhang,M.Q. (1999) Large-scale gene expression data analysis: a new challenge to computational biologists (published erratum appears in *Genome Res.*, 1999, **9**, 1156). *Genome Res.*, **9**, 681–688.

Zhu,H., Klemic,J.F. *et al.* (2000) Analysis of yeast protein kinases using protein chips. *Nature Genet.*, **26**, 283–289.

Zuker,M. (2000) Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.*, **10**, 303–310.