Interrelating Different Types of Genomic Data, from Proteome to Secretome: ’Oming in on Function

Dov Greenbaum,[2] Nicholas M. Luscombe,[1] Ronald Jansen,[1] Jiang Qian,[1] and Mark Gerstein[1,3]

[1] *Department of Molecular Biophysics and Biochemistry,* [2] *Department of Genetics, Yale University, New Haven, Connecticut 06520-8114, USA*

With the completion of genome sequences, the current challenge for biology is to determine the functions of all gene products and to understand how they contribute in making an organism viable. For the first time, biological systems can be viewed as being finite, with a limited set of molecular parts. However, the full range of biological processes controlled by these parts is extremely complex. Thus, a key approach in genomic research is to divide the cellular contents into distinct sub-populations, which are often given an “-omic” term. For example, the proteome is the full complement of proteins encoded by the genome, and the secretome is the part of it secreted from the cell. Carrying this further, we suggest the term “translatome” to describe the members of the proteome weighted by their abundance, and the “functome” to describe all the functions carried out by these. Once the individual sub-populations are defined and analyzed, we can then try to reconstruct the full organism by interrelating them, eventually allowing for a full and dynamic view of the cell. All this is, of course, made possible because of the increasing amount of large-scale data resulting from functional genomics experiments. However, there are still many difficulties resulting from the noisiness and complexity of the information. To some degree, these can be overcome through averaging with broad proteomic categories such as those implicit in functional and structural classifications. For illustration, we discuss one example in detail, interrelating transcript and cellular protein populations (transcriptome and translatome). Further information is available at http://bioinfo.mbb.yale.edu/what-is-it.

> “[It] does not consist of individuals, but expresses the sum of interrelations, the relations within which these individuals stand.”
> —adapted from Karl Marx, *Grundrisse* (1857)

The raw data produced by genome sequencing projects currently provides little insight into the precise workings of an organism at the molecular level (Luscombe et al., in press). Therefore, the goal of functional genomics is to complement the genomic sequence by assigning useful biological information to every gene. Through this, we aim to improve our understanding of how the different biological molecules contained within the cell (i.e., DNA, RNA, proteins, and metabolites) combine to make the organism viable. Clearly, the main challenge is the elucidation of all molecular, cellular, and physiological functions of each gene product. However, there are many subsidiary goals as part of this challenge, such as defining the three-dimensional structures of these macromolecules, their subcellular localizations, intermolecular interactions, and expression levels. Although gathering and classifying the necessary information is central to this process, it is impractical to rely on individual experiments for the potentially thousands of genes in each organism. Furthermore, with large-scale proteomic experiments still yet to be used widely, computational techniques — while sometimes based on less than ideal information — provide a crucial resource for assigning biological data.

The paper by Antelmann et al. in this issue of *Genome Research* (Antelmann et al. 2001) evaluates their earlier attempts to assign protein functions through computational means. Previously, the group used computational methods to predict all exported proteins(or members of the secretome) in *Bacillus subtilis* by searching for signal peptides and cell retention signals in the protein sequences. A better understanding of how and why a protein is secreted is valuable as the bacterium’s ability to export numerous enzymes enables it to degrade extracellular substrates and survive in a continuously changing environment. Moreover, it will eventually allow these bacteria to be employed as “cellular factories” for secreting commercially valuable proteins in large quantities (Tjalsma et al. 2000).

Antelmann et al.’s present paper aims to verify their previous predictions by experimentally characterising the entire population of secreted proteins using 2D gel electrophoresis and mass spectrometry. They showed that the original predictions correctly identified about 50% of all secreted proteins. Most of the disagreements were due to the inability to predict the secretion of proteins lacking the appropriate signal, or those containing seemingly inappropriate signals (cell retention signals). In summary, Antelmann et al.’s work highlights both the encouraging aspects of computational assignments of biological data, and reveals some of the shortcomings in the current methods.

The Path to Function is Filled with ’omes

To describe their studies, Antelmann et al. coined the term “secretome”. This ’omic term is an example of the new lexicon that has appeared recently to define the varied popu-

[3] **Corresponding author.**
E-MAIL mark.gerstein@yale.edu
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.207401.

lations and sub-populations in the cell (Fig. 1). These terms are generally suffixed with "-ome", with an associated research topic of "-omics".

Broadly, the existing 'omes can be divided into those that represent a population of molecules, and those that define their actions (Fig. 1). For the first category, populations provide an inventory or "parts list" of molecules contained within an organism (Gerstein and Hegyi 1998; Skolnick and Fetrow 2000; Vukmirovic and Tilghman 2000; Qian et al. 2001). The genome, the entire DNA sequence of an organism, presents a basis for defining the proteome, a list of coding DNA regions that result in protein products. Transcription of these coding sequences produces the transcriptome (Velculescu et al. 1997), which is the cellular complement of all mRNA under a variety of cellular conditions. Note, this population is weighted by the expression level of each molecule and, ideally, should incorporate the results of alternative splicing. Following translation of the transcriptome, we suggest the term "translatome" to describe the cellular population of proteins expressed in the organism at a given time, explicitly weighted by their abundance. It is important to note that, whereas the membership of the genome and proteome are virtually static, the transcriptome and translatome are dynamic and continually change in response to internal and external events. Additional 'omes describe the presence of molecules that are not encoded by the genome, but are nonetheless essential, for instance, the metabolome (Tweeddale et al. 1998). Because of the newness of most 'omic terms, a few still have competing definitions. This is most evident for the proteome (see Table 1).

The second group of 'omes are fewer in number and describe the actions of the protein products. For example, the secretome is a subset of the proteome that is defined by its action, that is, it is actively exported from the cell. The interactome (Sanchez et al. 1999) lists all of the specific interactions that are made between macromolecules in the cell. More abstractly, the regulome (Web references only; see Table 1) defines the genome-wide regulatory network of the cell and most notably includes transcription regulation pathways.

The elucidation of each of these 'omes contributes to the ultimate goal of functional genomics, defining the functome,which describes all of the functions that are assigned to each gene in the genome (the Rison et al. 2000, http://www.biochem.ucl.ac.uk/~rison). The functions of a gene can be described at many levels, including their biochemical, cellular and physiological roles (Ashburner et al. 2000), and also depend on additional factors that are not immediately associated with their basic functions, such as subcellular localization and intermolecular interactions. Therefore, aspects of the functome may be expressed in terms of other 'omes, for example those that group similar biochemical functions, for example the immunome (Pederson 1999); similar localizations, for example the secretome; and similar interactions, for example, the interactome. For the record, we coin our own term here; at present, a large proportion of genes can only be described as members of the "unknome": those with currently no functional information!

Computational Methods for Defining 'omes

There are a variety of computational approaches for defining 'omes (Gerstein and Honig 2001):

(1) Algorithmic methods for predicting genes, protein structure, interactions, or localization based on patterns in indi-

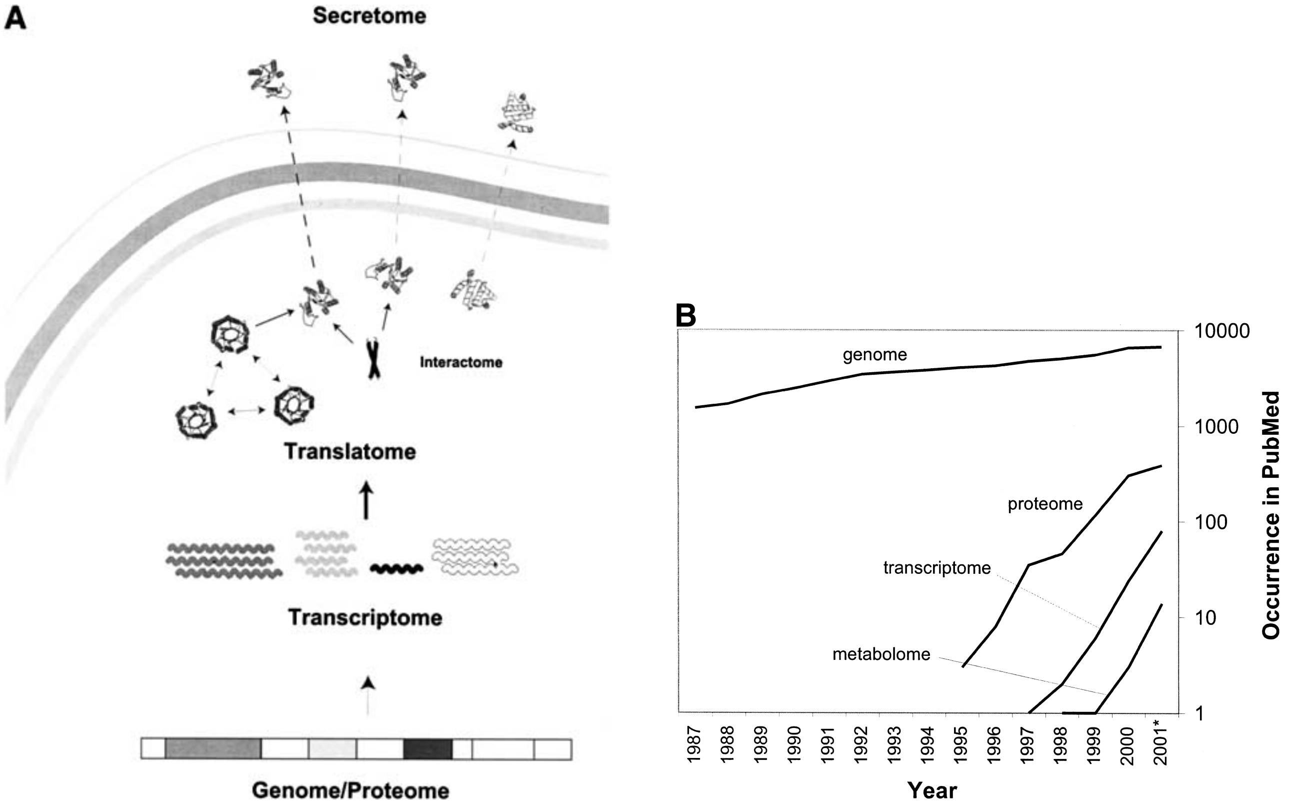

Figure 1 An overview of the current 'omic terminology. (*A*) A schematic of the main 'omes in the process of gene expression. (*B*) The literature citations of four of the most widely used 'omes over time.

Table 1. A Table of 'omes, Together with their Occurrence in the Literature and on the World Wide Web

Term	Description	Google	PubMed	Year of first PubMed citation
Genome	The full complement of genetic information both coding and non coding in the organism	~1880000	66171	1932**
Proteome	The protein-coding regions of the genome	~63,000	703	1995
Transcriptome	The population of mRNA transcripts in the cell, weighted by their expression levels	3520	72	1997
Physiome	Quantitative description of the physiological dynamics or functions of the whole organism	2980	15	1997
Metabolome	The quantitative complement of all the small molecules present in a cell in a specific physiological state	349	12	1998
Phenome	Qualitative identification of the form and function derived from genes, but lacking a quantitative, integrative definition	4980	6	1995
Morphome	The quantitative description of anatomical structure, biochemical and chemical composition of an intact organism, including its genome, proteome, cell, tissue and organ structures	238	2	1996
Interactome	List of interactions between all macromolecules in a cell	56	2	1999
Glycome	The population of carbohydrate molecules in the cell	46	1	2000
Secretome	The population of gene products that are secreted from the cell	21	1	2000
Ribonome	The population of RNA-coding regions of the genome	1	1	2000
Orfeome	The sum total of open reading frames in the genome, without regard to whether or not they code; a subset of this is the proteome	42	–	–
Regulome	Genome-wide regulatory network of the cell	18	–	–
Cellome	The entire complement of molecules and their interactions within a cell	17	–	–
Operome	The characterization of proteins with unknown biological function	8	–	–
Transportome	The population of the gene products that are transported; this includes the secretome	1	–	–
Pseudome	The complement of pseudogenes in the proteome	–	–	–
Functome	The population of gene products classified by their functions	1	–	–
Translatome	The population of proteins in the cell, weighted by their expression levels	–	–	–
Foldome	The population of gene products classified through their tertiary structure	–	–	–
Unknome*	Genes of unknown function	–	–	–

Updated versions of this table will be available through our Web site at http://bioinfo.mbb.yale.edu/what-is-it. Note that we define five new 'omes: the translatome, the foldome, the pseudome, the functome, and the unknome. Our definition of the translatome is motivated partially by the ambiguities in term proteome, which has two competing definitions. First, broadly favored by computational biologists, it is a list of all the proteins encoded in the genome (Gaasterland 1999; Doolittle 2000). In this context, it is equivalent to what some refer to as the orfeome, (i.e., the set of genes excluding noncoding regions). Experimentalists, especially those involved in large-scale experiments such as expression analysis and 2D electrophoresis, favor a second definition. Here, it is used to describe the actual cellular contents of proteins, taking into account the different levels of protein concentrations (Yates 2000). We prefer the former definition for proteome, and use the term translatome for the latter. See http://www.genomic glossaries.com/content/omes.asp for a listing of other 'omes and their definitions.

*This term is also used in other fields with different meanings. **First citation according to the Oxford English Dictionary.

vidual sequences or structures; for example, defining the proteome or orfeome using a gene-finding algorithm on the genome (Claverie 1997; Guigo et al. 2000; Harrison et al. 2001; Yeh et al. 2001), determining the foldome from structure prediction of the proteome (Simons et al. 2001), determining the interactome from the foldome, using known binding sites (Teichmann et al. 2001), and determining the secretome through identifying signal sequences in the proteome (Tjalsma et al. 2000).

(2) Annotation transfer through homology, that is, inferring structure or function based on sequence and structural information of homologous proteins (Genstein 1997, 1998; Brenner 1999; Hegyi and Gerstein 1999; Wilson et al. 2000; Thornton 2001; Hegyi and Gerstein, in press).

(3) Using a "guilt-by-association" method based on clustering where functions or interactions are inferred from clusters of functional genomic data, such as expression information. For example, similar functions can sometimes be inferred through interactions with other proteins or similar expression profiles (Eisen et al. 1998; Marcotte et al. 1999; Gerstein and Jansen 2000; Ito et al. 2001).

Experimental Methods for Defining 'omes

Although still in their infancy, several large-scale experimental techniques are designed to assess the nature of different 'omes. Gene expression studies are now well established and microarray or GeneChip technologies can be used to measure mRNA abundance in the cell and hence define the transcriptome

(Epstein and Butow 2000). Detection of protein concentration and definition of the translatome is more difficult, however, as evidenced by the dearth of such data. At present, the most prominent method employs two-dimensional electrophoresis to isolate proteins followed by mass spectrometry for their identification (Futcher et al. 1999; Gygi et al. 1999; Naaby-Hansen et al. 2001) followed by quantification (Appel et al. 1997; Aebersold et al. 2000; Gygi et al. 2000). The two-hybrid system enables detection of specific protein–protein associations to build the interactome (Uetz et al. 2000; Ito et al. 2001; Walhout and Vidal 2001). Antelmann et al. (2001) used two-dimensional electrophoresis to determine the membership of the secretome.

Given the goal of determining the functome, perhaps the most exciting technology is the protein chip system, which is capable of high-throughput screening of protein biochemical activity. (Zhu et al. 2000; Zhu 2001, in press). Other methods for obtaining large-scale protein functional characterization include a transposon insertion methodology (Ross-Macdonald et al. 1999).

Although we discuss the computational and experimental methods separately, there is, in fact, an inseparable relationship between the two. On the one hand, data resulting from high-throughput experimentation require intensive computational interpretation and evaluation (Carson et al. 2001). On the other hand, computational methods use empirical data to build a knowledge base for predictions. Furthermore, they sometimes produce questionable predictions that should be reviewed and confirmed through experiments, as Antelmann et al. point out. In addition to these high-throughput techniques, another interesting tactic is to aggregate the results of individual experiments through comprehensive literature searches. Although there clearly are difficulties with differing experimental conditions and varying interpretations, preliminary results have shown this to be an effective method (Jenssen et al. 2001; Marcotte et al. 2001; Ono et al. 2001).

Interrelating Different 'omes

Having categorized the organism into different sub-populations, a fundamental approach in genomics is to establish relationships between the different 'omes. In other words, by piecing the individual 'omes together, we hope to build a full and dynamic view of the complex processes that support the organism. For example, how do the proteome and regulome combine to produce the translatome?

As with defining the 'omes, these relationships can be explored in different ways:

(1) Defining or assigning one 'ome based on another, as described above.

(2) Comparing one 'ome with another to better understand the processes that shift one population into its successor. For instance, this could be done by correlating expression measurements for the transcriptome and translatome (see below).

(3) Calculating "missing" (experimentally unattainable) information in one 'ome based on information in another one – for example, using the known relationships between gene expression level and subcellular location to help predict the destination of proteins of unknown localization (Drawid and Gerstein 2000; Drawid et al. 2000).

(4) Describing the intersection between multiple populations. For example, combining data from the transcriptome and the functome could describe the array of biochemical, and potentially, physiological functions that are available to the cell at any given time (Hegyi and Gerstein 1999).

The Use of Broad Categories to Interpret Noisy Data

Functional genomics experiments generally give rise to very complicated data that are inherently hard to interpret. Furthermore, these data are often plagued with noise (Kerret al. 2000). Both factors can lead to inaccuracies and conflicting interpretations.

A good example is gene expression measurements, which are known to fluctuate between experiments even if the conditions are apparently identical (Baldi and Long 2001). These fluctuations are often due to measurement errors, but there are also inherent biological variations of expression levels, relating to the stochastic nature of gene expression (Szallasi 1999). One cause is the very low cellular concentrations of many transcription factors, meaning, that they bind promoters very rarely. Such events approximate to a Poisson process, and in fact, macroscopic chemical kinetics would fail to describe the resulting expression level of the gene (McAdams and Arkin 1999; Thattai and van Oudenaarden 2001). In another example, the interactome, when determined using the yeast two-hybrid technique, is notorious for false positives and negatives (Ito et al. 2000; Serebriiskii et al. 2000; Ito et al. 2001; Legrain et al. 2001).

A useful way to tackle noise and complexity of functional genomics information is to average the data from many different genes into broad 'omic categories (Jansen and Gerstein 2000). For instance, instead of looking at how the level of expression of an individual gene changes over a timecourse, we can average all the genes in a functional category (e.g., glycolysis) together. This gives a more robust answer about the degree to which a functional system changes over the timecourse. Likewise, if one wants to investigate the relationship between a gene's essentiality — whether or not it is essential (Winzeler et al. 1999) — and its subcellular localization, it might be useful to combine the results for all proteins in the same compartment. This would give the average degree of essentiality of all nuclear proteins, cytoplasmic proteins, and so forth. In an actual study for predicting protein subcellular localization, we obtained more accurate predictions for the overall populations (96% accuracy) of a given subcellular compartment than for individual genes (75% accuracy) (Drawid et al. 2000).

Thus, the strength of genomic studies lies in the global comparisons between biological systems rather than detailed examination of single genes or proteins. Genomic information is often misused when applied exclusively to individual genes. If one is interested only in one particular gene, there are many more conclusive experiments that should be consulted before using the results from genomics datasets. Therefore, genomic data should not be used in lieu of traditional biochemistry, but as an initial guideline to identify areas for deeper investigation and to see how those results fit in with the rest of the genome.

Moreover, most genomics datasets give relative rather than absolute information, which means that information about a single gene has little meaning in isolation. For example, they are best used to identify "outlier" genes that are particularly highly-expressed, or have especially many interactions, rather than to focus on the individual measurements for a particular gene. A gene that makes a particularly large number of interactions may indicate that it is a key component of the cell. One numerical technique that is particularly useful with regard to dealing with this information is expressing results through ranks (i.e., not giving the number of interactions of a particular gene product, but how it ranks when compared with others). Furthermore, it provides a powerful way to combine many different heterogeneous sources of information into a common and statistically robust numerical framework (Gerstein and Levitt 1997; Gerstein and Hegyi 1998; Qian et al. 2001).

These observations should be kept in mind when interacting with genomics tools and databases. Many websites focus on providing a lot of information for a single gene sequence or protein, in a "non-genomic" fashion. Rather, such sites should be designed to simultaneously display and manipulate large populations of genes. In the absence of such an 'omic interface, it is important that information resources at least accommodate bulk downloading of standardized data.

A Case Study: Interrelating the Transcriptome and the Translatome

A specific example of comparing the transcriptome and translatome will illustrate the points we made about interrelating 'omes and using categories to interpret noisy data. Here the question is to what degree do highly expressed genes (transcriptome) correspond to highly expressed proteins (translatome)? We can get very different answers depending on the perspective we take:

Theoretical View

Turning to the entire mRNA and protein populations, the change in protein concentration over time is equal to the rate of translation minus the rate of degradation. Borrowing from chemical kinetics, this is approximately expressed by the equation $dP(i,t)/dt = SE(i,t) - DP(i,t)$, where P is the abundance of protein i at time t, E is the corresponding expression level of this protein, S is a general rate of protein synthesis per mRNA, and D is a general rate of protein degradation per protein. Obviously, this is highly simplified and in a more general context one would expect that the rates of synthesis and degradation to be different for each gene and dependent on the regulatory effects of other genes over time. In addition, the equation does not take into account the stochastic nature of gene expression (see above) (Chen et al. 1999).

Direct Comparison of Individual mRNA and Protein Data

At the moment, we do not have good enough data to apply models such as the equation above. However, there is an intuitive sense that highly expressed genes correspond to highly abundant proteins. (One can see this by imagining the situation at steady-state, when the lefthand side of the equation is zero and a positive correlation between E and P results.) Figure 2A shows the direct comparison between raw measurements of mRNA expression and protein abundance data for 181 genes in yeast drawn from two recent studies (Futcher et al. 1999; Gygi et al. 1999). The two variables show a high degree of variation for individual data pairs and investigators have come to different conclusions about the general correlation between them. This is, to some degree, dependent on the subjective way of analyzing the data.

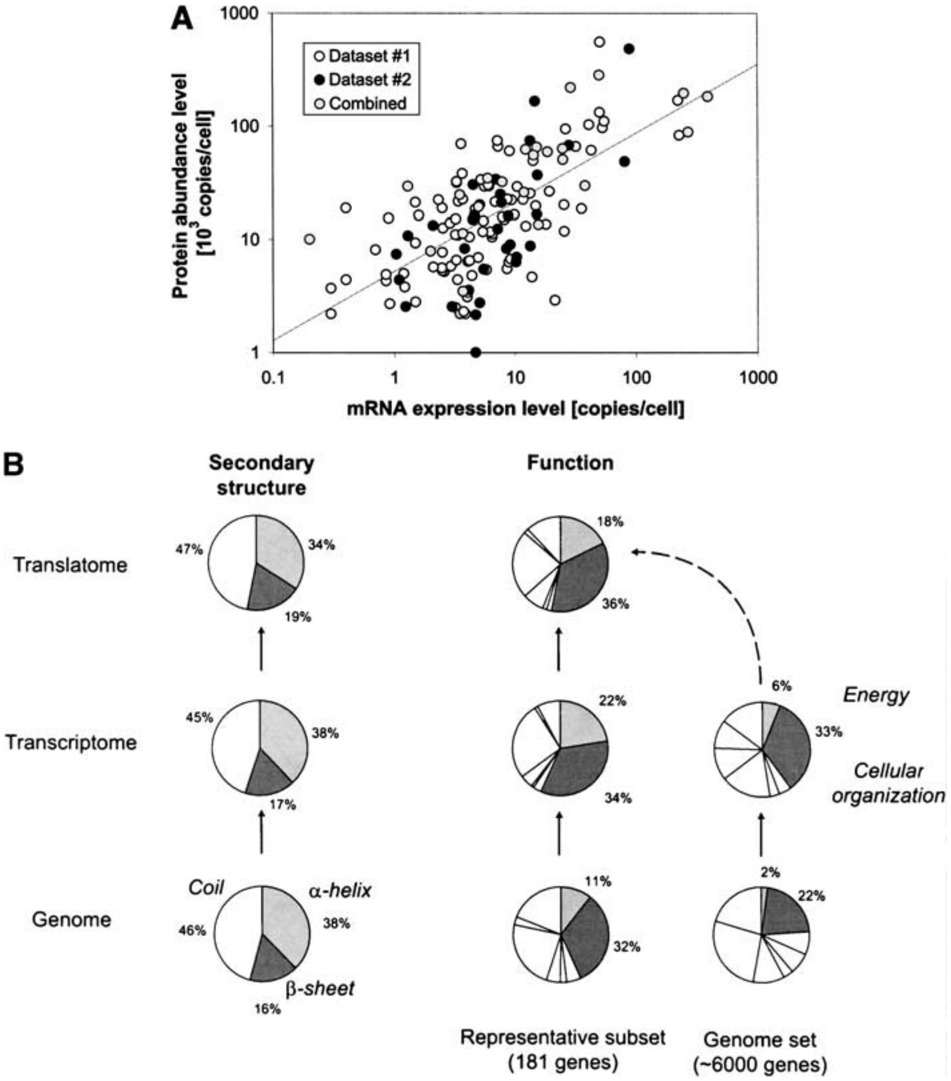

Figure 2 Interrelating the transcriptome and the translatome.(*A*) A direct comparison of protein abundance and mRNA expression. The abundance data is from two recent studies (datasets 1 and 2) of a global comparison of protein and mRNA expression levels in yeast (Futcher et al. 1999; Gygi et al. 1999). The combined protein abundance dataset is an average of the data points from the two studies if the given gene product appears in both studies. The mRNA expression data is mainly derived from Holstege (1998). Although there is a general trend for protein concentration to rise with mRNA levels, the actual correlation is weak and protein concentrations can sometimes vary by more than two orders of magnitude for a given mRNA level. Similar observations were reported by a study in human liver cells (Anderson and Seilhamer 1997). The mRNA expression data was scaled and the process is described on our Web site (http://bioinfo.mbb.yale.edu/expression). (*B*) The composition of the genome (proteome), transcriptome and translatome in terms of broad categories: protein secondary structures and functions. This is based on the analysis in Jansen and Gerstein (2000) with updates to include protein abundance data. The *bottom* piecharts give the composition in the genome, the *middle* charts in the transcriptome and the *top* charts in the translatome. The compositions for the transcriptome and the translatome are calculated by weighting each mRNA/protein with its respective expression level. The secondary structure composition does not vary significantly between the different 'omes, mainly because transcription and translation are independent of secondary structure. The *right* five piecharts analyse the functional composition. We highlight the Energy and Cellular Organization categories determined from MIPS (Mewes et al. 2000). A problem in comparing the different 'omes is that each represents a different set of genes. For instance, protein levels have been measured only for a fraction of genes whereas mRNA levels are known for almost all genes. The piecharts show the compositions for the whole genome in the right column and a representative subset of genes with known protein levels in the left column. Comparing the left to the right immediately shows the experimental bias of two-dimensional electrophoresis (the method for measuring protein abundance) with respect to certain functional categories. There is good agreement between the composition in the translatome and the transcriptome, despite the low correlation of protein and mRNA levels for individual genes. In comparison, the compositions in the genome are much lower.

Analysis of the Data in Terms of Categories

Although the relationship between mRNA and protein levels is vague for individual genes, some of the statistics for broad categories of protein properties are much more robust. Figure 2B shows the protein secondary structure and functional composition in the genome, the transcriptome (i.e., weighted by mRNA abundance), and in the translatome (i.e., weighted by protein abundance). In contrast to the differences between mRNA and protein data for individual genes, the broad categories show that the transcriptome and translatome populations are remarkably similar; both contain roughly the same proportions of secondary structure and functional categories. Moreover, this contrasts with the genome, which appears to have a distinctly different composition of functional categories. This illustrates that we get a more consistent picture when we average across the population; that is, there is broad similarity between the characteristics of highly expressed mRNA and highly abundant proteins.

Conclusion

The ultimate goal of genomics is the elucidation of the functome, but there are many intermediate steps. By viewing the cell in terms of a list of distinct parts, we can define, part by part, each 'ome in an effort to determine and categorize functional information for each gene. High-throughput experimentation and computational techniques are valuable and complementary; that is, conclusive results often cannot be made based on a single methodology. It must be noted that this data is only valuable with regard to large populations, and as such, should only be used as a secondary source for single gene queries. Moreover, genomic approaches result in inaccurate and noisy data. This noise, while deafening on the single gene level, can be tolerated through the use of broad categories to analyze the data.

ACKNOWLEDGMENTS

R.J. acknowledges IBM Graduate Research Fellowship.

REFERENCES

Aebersold, R., Rist, B., and Gygi, S.P. 2000. *Ann. NY Acad. Sci.* **919:** 33–47.

Anderson, L. and Seilhamer, J. 1997. *Electrophoresis* **18:** 533–537.

Antelmann, H., Tjalsma, H., Voight, B., Ohlmeier, S., Bron, S., van Dijl, J.M., and Hecker, M. 2001. A proteomic view on genome-based signal peptide predictions. *Genome Res.* **11:** 1484–1502.

Appel, R.D., Vargas, J.R., Palagi, P.M., Walther, D., and Hochstrasser, D.F. 1997. *Electrophoresis* **18:** 2735–2748.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. *Nat. Genet.* **25:** 25–29.

Baldi, P. and Long, A.D. 2001. *Bioinformatics* **17:** 509–519.

Brenner, S. E. 1999. *Trends Genet.* **15:** 132–3.

Carson, J.H., Cowan, A., and Loew, L.M. 2001. *Trends Cell. Biol.* **11:** 236–238.

Chen, T., He, H.L., and Church, G.M. 1999. *Pac. Symp. Biocomput.* 29–40.

Claverie, J.M. 1997. *Hum. Mol. Genet.* **6:** 1735–44.

Doolittle, W.F. 2000. *Curr. Opin. Struct. Biol.* **10:** 355–358.

Drawid, A. and Gerstein, M. 2000. *J. Mol. Biol.* **301:** 1059–1075.

Drawid, A., Jansen, R., and Gerstein, M. 2000. *Trends Genet.* **16:** 426–430.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. 1998. *Proc. Natl. Acad. Sci.* **95:** 14863–14868.

Epstein, C. and Butow, R. 2000. *Curr. Opin. Biotechn.* **11:** 36–41.

Futcher, B., Latter, G.I., Monardo, P., McLaughlin, C.S., and Garrels, J.I. 1999. *Mol. Cell. Biol.* **19:** 7357–7368.

Gaasterland, T. 1999. *Curr. Opin. Microbiol.* **2:** 542–547.

Gerstein, M. 1997. *J. Mol. Biol.* **274:** 562–576.

———. 1998. *Proteins* **33:** 518–534.

Gerstein, M. and Levitt, M. 1997. *Proc. Natl. Acad. Sci.* **94:** 11911–11916.

Gerstein, M. and Hegyi, H. 1998. *FEMS Microbiol. Rev.* **22:** 277–304.

Gerstein, M. and Jansen, R. 2000. *Curr. Opin. Struct. Biol.* **10:** 574–584.

Gerstein, M. and Honig, B. 2001. *Curr. Opin. Struct. Biol.* **11:** 327–329.

Guigo, R., Agarwal, P., Abril, J.F., Burset, M., and Fickett, J.W. 2000. *Genome Res.* **10:** 1631–1642.

Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. 1999. *Mol. Cell. Biol.* **19:** 1720–1730.

Gygi, S.P., Rist, B., and Aebersold, R. 2000. *Curr. Opin. Biotechnol.* **11:** 396–401.

Harrison, P.M., Echols, N., and Gerstein, M.B. 2001. *Nucl. Acids Res.* **29:** 818–830.

Hegyi, H. and Gerstein, M. 1999. *J. Mol. Biol.* **288:** 147–164.

———. 2001. *Genome Res.*. (in press)

Holstege, F.C., Jennings, E.G. Wyrick, J.J. Lee, T., I. Hengartner, C. J. Green, M. R. Golub, T. R. Lander, E. S. Young, R.A. 1998. *Cell* **95:** 717–728.

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S., and Sakaki, Y. 2000. *Proc. Natl. Acad. Sci.* **97:** 1143–1147.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. 2001. *Proc. Natl. Acad. Sci.* **98:** 4569–4574.

Jansen, R. and Gerstein, M. 2000. *Nucl. Acids Res.* **28:** 1481–1488.

Jenssen, T.K., Laegreid, A., Komorowski, J., and Hovig, E. 2001. *Nat. Genet.* **28:** 21–8.

Kerr, M.K., Martin, M., and Churchill, G.A. 2000. *J. Comput. Biol.* **7:** 819–837.

Legrain, P., Wojcik, J., and Gauthier, J. 2001. *Trends Genet.* **17:** 346–352.

Luscombe, N., Greebaum, D., and Gerstein, M. *Methods of Information in Medicine*. (in press)

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., and Eisenberg, D. 1999. *Science* **285:** 751–753.

Marcotte, E.M., Xenarios, I., and Eisenberg, D. 2001. *Bioinformatics* **17:** 359–363.

McAdams, H.H. and Arkin, A. 1999. *Trends Genet.* **15:** 65–69.

Mewes, H.W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schuller, C., et al. 2000. *Nucl. Acids Resc* **28:** 37–40.

Naaby-Hansen, S., Waterfield, M.D., and Cramer, R. 2001. *Trends Pharmacol. Sci.* **22:** 376–384.

Ono, T., Hishigaki, H., Tanigami, A., and Takagi, T. 2001. *Bioinformatics* **17:** 155–161.

Pederson, T. 1999. *Mol. Immunol.* **36:** 1127–8.

Qian, J., Stenger, B., Wilson, C.A., Lin, J., Jansen, R., Teichmann, S.A., Park, J., Krebs, W.G., Yu, H., Alexandrov, V.,et al. 2001. *Nucl. Acids Res.* **29:** 1750–1764.

Rison, S.C.G, Hodgman, T.C., and Thornton, J.M. 2000. *Funct. Integr. Genomics* **1:** 56–59.

Ross-Macdonald, P., Coelho, P.S., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K.H., Sheehan, A., Symoniatis, D., Umansky, L., et al. 1999. *Nature* **402:** 413–418.

Sanchez, C., Lachaize, C., Janody, F., Bellon, B., Roder, L., Euzenat, J., Rechenmann, F., and Jacq, B. 1999. *Nucl. Acids Res.* **27:** 89–94.

Serebriiskii, I., Estojak, J., Berman, M., and Golemis, E.A. 2000. *Biotechniques* **28:** 328–330, 332–336.

Simons, K.T., Strauss, C., and Baker, D. 2001. *J. Mol. Biol.* **306:** 1191–1199.

Skolnick, J. and Fetrow, J.S. 2000. *Trends Biotechnol.* **18:** 34–39.

Szallasi, Z. 1999. *Pac. Symp. Biocomput.*5–16.

Teichmann, S.A., Murzin, A.G., and Chothia, C. 2001. *Curr. Opin. Struct. Biol.* **11:** 354–363.

Thattai, M. and van Oudenaarden, A. 2001. *Proc. Natl. Acad. Sci.* **98:** 8614–8619.

Thornton, J.M. 2001. *Science* **292:** 2095–7.

Tjalsma, H., Bolhuis, A., Jongbloed, J.D., Bron, S., and van Dijl, J.M. 2000. *Microbiol. Mol. Biol. Rev.* **64:** 515–547.

Tweeddale, H., Notley-McRobb, L., and Ferenci, T. 1998. *J. Bacteriol.* **180:** 5109–5116.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. 2001. *Proc. Natl. Acad. Sci.* **98:** 4569–4574.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P.,et al. 2000. *Nature* **403:** 623–627.

Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Jr., Hieter, P., Vogelstein, B., and Kinzler, K.W. 1997. *Cell* **88:** 243–251.

Vukmirovic, O.G. and Tilghman, S.M. 2000. *Nature* **405:** 820–822.

Walhout, A.J. and Vidal, M. 2001. *Methods* **24:** 297–306.

Wilson, C.A., Kreychman, J., and Gerstein, M. 2000. *J. Mol. Biol.* **297:** 233–249.

Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., et al. 1999. *Science* **285:** 901–906.

Yates, 3rd, J.R. 2000. *Trends Genet.* **16:** 5–8.

Yeh, R.F., Lim, L.P., and Burge, C.B. 2001. *Genome Res.* **11:** 803–816.

Zhu, H., Klemic, J.F., Chang, S., Bertone, P., Casamayor, A., Klemic, K.G., Smith, D., Gerstein, M., Reed, M.A., and Snyder, M. 2000. *Nat. Genet.* **26:** 283–289.

Zhu, H.B., M Bangham, R Hall, D Casamayor, A Bertone, P Lan, N Jansen, R Bidlingmaier, S Houfek, T Mitchell, T Miller, P Dean, R Gerstein, M Snyder, M. *Science* (in press).