Nucleic Acids Research, 2001, Vol. 29, No. 17 **3513–3519**

A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*

Andrei Grigoriev*

GPC Biotech, Fraunhoferstraße 20, Martinsried 82152, Germany

Received May 29, 2001; Revised and Accepted July 10, 2001

ABSTRACT

The relationship between the similarity of expression patterns for a pair of genes and interaction of the proteins they encode is demonstrated both for the simple genome of the bacteriophage T7 and the considerably more complex genome of the yeast *Saccharomyces cerevisiae*. Statistical analysis of large-scale gene expression and protein interaction data shows that protein pairs encoded by co-expressed genes interact with each other more frequently than with random proteins. Furthermore, the mean similarity of expression profiles is significantly higher for respective interacting protein pairs than for random ones. Such coupled analysis of gene expression and protein interaction data may allow evaluation of the results of large-scale gene expression and protein interaction screens as demonstrated for several publicly available datasets. The role of this link between expression and interaction in the evolution from monomeric to oligomeric protein structures is also discussed.

INTRODUCTION

Many genome-wide expression profiling studies and protein–protein interaction (PPI) screens currently underway are expected to obtain substantial new information on the known and yet undiscovered cellular processes. Piecing together these results will be instrumental in creating a unified view of biomolecular pathways of living cells. The promise of effective integration of these heterogeneous datasets is based on the assumption that there is a link between interaction of two proteins *in vivo* and similar expression patterns of the genes encoding those proteins.

Since interacting proteins must be simultaneously present in a cell, their lifetimes, between synthesis (translation of respective mRNA and function-enabling modifications) and degradation, should overlap. In many cases, genes co-expressed 'by definition' (as in bacterial operons with one polycistronic mRNA for several adjacent cistrons) encode interacting proteins, often members of one pathway. A number of such gene pairs corresponding to interacting proteins are conserved across prokaryotes (1). Further, open reading frames (ORFs) of such genes sometimes merge, resulting in one amino acid chain. The latter possibly represents the ultimate transformation of a connection between co-expression and interaction of two entities into the covalently linked domains of one protein. It has recently been exploited to predict potentially interacting proteins in some prokaryotic species (2)

In these special cases the ORFs showing similar expression profiles are more likely to encode interacting proteins than random pairs of ORFs. How general is this observation, especially in more complex genomes without polycistronic messages?

Different levels of granularity in assessing genome-wide gene expression data are generally represented either by clusters of genes with sufficiently similar profiles or by pairwise correlations of such profiles. However, many clustering approaches are essentially based on a pairwise similarity and the smallest non-singleton cluster would consist of two genes. Hence a natural question to ask is: do similar (highly correlated) expression profile pairs indicate interacting protein pairs?

In this paper we show that interacting proteins, in general, correspond to gene pairs with higher mean similarity of expression profiles compared to random gene pairs. First, we demonstrate this for gene groups explicitly shown to be expressed and translated into proteins at the same time during the development of phage T7. Secondly, we statistically test this for calculated similarity of gene expression profiles in the yeast *Saccharomyces cerevisiae*. Further, we show here how to use this analysis of similarity distributions to assess the results of different large-scale efforts in protein interaction mapping. Then we consider homotypic interactions and related evolutionary implications. Finally, we look at these interconnections from a point of view that amalgamates various cellular processes into a unified picture, using phage T7 as an example.

RESULTS

Bacteriophage T7

Bacteriophage T7 was the target of the first genome-wide PPI study, which uncovered 25 interactions between phage

*Tel: +49 89 8565 2644; Fax: +49 89 8565 2610; Email: andrei.grigoriev@gpc-biotech.com

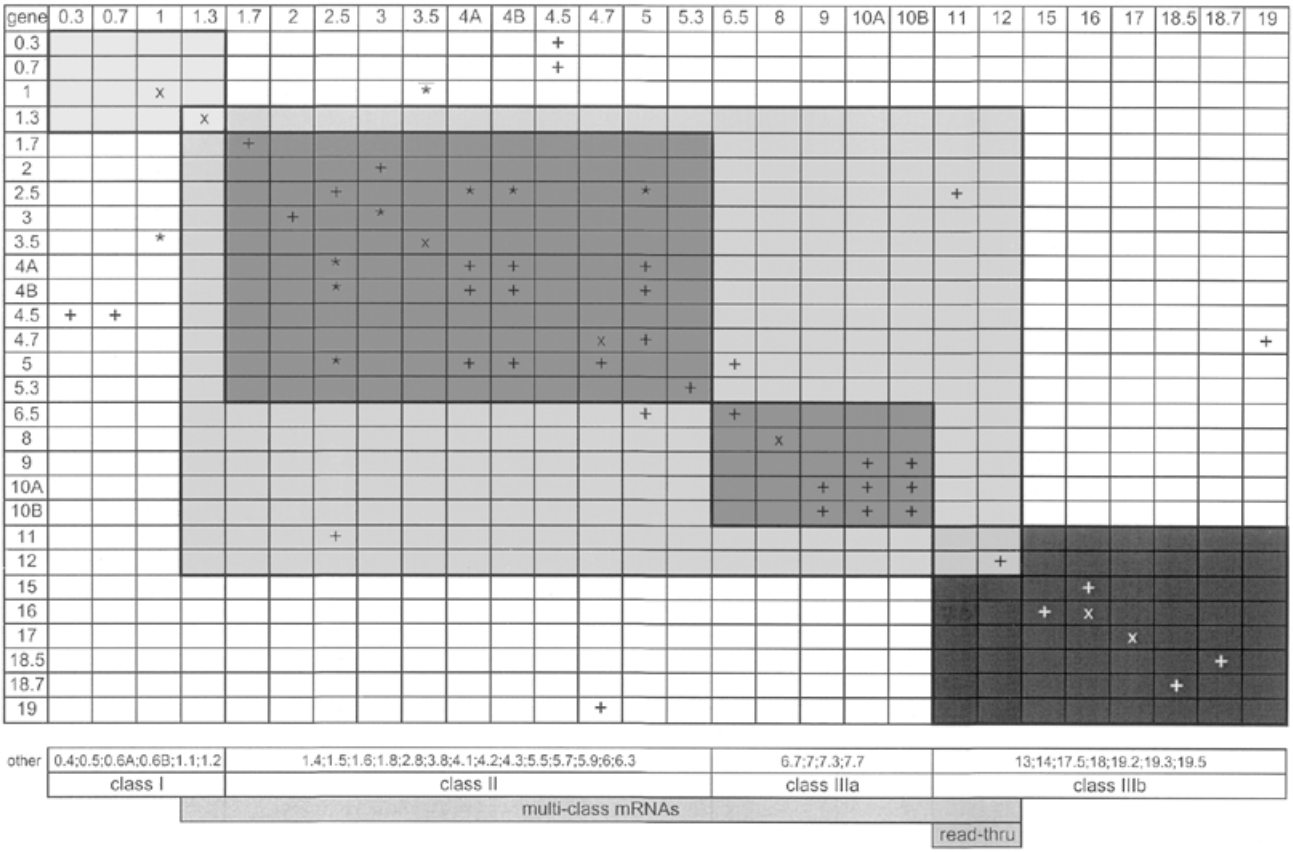

Figure 1. Protein–protein interaction matrix for the phage T7. Rows and columns correspond to individual genes, scored positive in yeast two-hybrid screens (3). Genes are arranged left to right and top to bottom according to their position in the genome; class boundaries are shown. Gene expression is coupled to the genome position in T7, since class members are co-expressed. Genes not detected in interaction screens are given in the row labelled 'other', also grouped by class. An interaction between two proteins *i* and *j* is presented in the (*i*,*j*) and the symmetrical (*j*,*i*) elements of the matrix as + (an empty cell indicates no detected interaction). A positive yeast two-hybrid result likely caused by intramolecular associations only (3) is shown as × and interactions known from other studies but missed in yeast two-hybrid screens are marked as *. Distinction between baits and targets is not taken into account so the matrix is symmetrical. Shaded rectangles correspond to the different gene classes and inter-class mRNAs are shown schematically as grey boxes at the bottom (see text).

proteins and protein fragments (3). Adding to those results several interactions known from structural and biochemical studies (4–10), we have produced an interaction matrix of the T7 proteins (Fig. 1).

This matrix also displays gene expression information, since the genetic organisation and pattern of transcription of phage T7 are understood at the nucleotide level (11). T7 genes can be divided into three classes according to the level and timing of transcription and translation. Each class contains genes of related functions. The genes of class I are transcribed first by the host RNA polymerase; their functions are directed towards overcoming host restriction systems and converting the host cell metabolism to the production of T7 proteins. Then the newly synthesised T7 RNA polymerase (T7 RNAP) transcribes the class II genes (with functions involved in DNA metabolism), followed by the class III genes (bacteriophage particle proteins and proteins involved in maturation and packaging of DNA). Co-occurrence and evanescence of the corresponding proteins in *Escherichia coli* cell extracts has been convincingly demonstrated (11).

Owing to the arrangement of strong T7 RNAP promoters and terminators the class III genes form two subclasses: IIIa with the higher and IIIb with the lower average level of transcription. These two subclasses have been clearly differentiated on the basis of mutations resulting from different transcription levels (12). On the other hand, genes 11 and 12 from class IIIb are transcribed entirely via read-through transcription past the terminator T_Φ upstream of them (11) and the corresponding mRNAs contain genes of class IIIa (and possibly II). In fact, some of the T7 mRNAs correspond to genes from different classes (11; see also Fig. 1).

The striking feature of this matrix displaying both gene expression and PPI data is that most of the interactions are contained in the rectangles corresponding to the co-regulated transcriptional classes of the T7 genes. Individual interactions have been discussed in detail (3) but such a clear relationship between the global gene expression patterns in T7 and interactions of the corresponding proteins is shown graphically for the first time in this paper.

Many symbols are located on the main diagonal of the matrix, signifying homotypic interactions (see also below). Nevertheless, even if the diagonal is not taken into account, rectangles corresponding to the distinct transcriptional classes are the most populated segments of the matrix.

Table 1. Protein–protein interaction datasets analysed

Dataset	Size, non-redundant interactions	Homotypic interactions (%)	Comments and references
MIPS	1227[a]	44 (3.6)	Collection of physical interactions on the MIPS site[b] (30)
UETZ-ALL	929[a]	45 (4.8)	Result of compilations in (17) and (18)[c], excluding 134 pairs added from (19)
ITO-CORE	806	52 (6.5)	'Core' subset of (19)[d]
ITO-NONCORE	3684	30 (0.8)	Full dataset of (19)[d] excluding ITO-CORE

[a]Differs from numbers in Schwikowski *et al*. (17) due to including homotypic interactions.
[b]http://mips.gsf.de/proj/yeast/tables/interaction/physical_interact.html.
[c]http://portal.curagen.com.
[d]http://genome.c.kanazawa-u.ac.jp/Y2H.

Class I genes do not follow this trend, since their products seem to avoid intra-class (as well as inter-class) interactions. Moreover, the two homotypic interactions detected by a yeast two-hybrid screen may be artefactual, e.g. T7 RNAP, encoded by gene 1, is a monomer according to extensive structural and biochemical evidence (6,13).

An explanation for this seeming peculiarity is that the proteins produced from the genes in this class only 'set the stage' for the later coordinated events of phage development and interact largely with host proteins, rather than with the other class members. That may be the reason why only four genes (out of a total of 10 in class I) have produced detectable interacting pairs and are shown in the matrix. Thus two interactions of the product of gene 4.5 (dispensable for growth, unknown function) with the proteins encoded by genes 0.3 (host endonuclease inhibition) and 0.7 (protein kinase, phosphorylating host proteins) seem puzzling, while the well studied interaction of T7 RNAP with the gene 3.5 lyzozyme (8,9) was not found in a yeast two-hybrid screen.

While suggesting a strong link between co-expression and interaction in the presented visual form, the T7 data are somewhat sparse, so another genome was chosen for a more reliable statistical analysis.

Saccharomyces cerevisiae

The yeast *S.cerevisiae* was the first eukaryotic organism to be completely sequenced at the genome level (14), the first target of large-scale gene expression studies (see 15 and references therein) and also the natural host for the yeast two-hybrid system (16). It is therefore uniquely characterised to allow for statistical testing of the hypothesis of a general relationship between PPI and gene expression patterns.

A widely used yeast gene expression reference dataset has been described by Eisen *et al*. (15). It contains 2467 ORFs with known function with 79 data points/ORF (experimental conditions include diauxic shift, mitotic cell division cycle, sporulation, temperature and reducing shocks). We analysed the expression profiles of these ORFs with regard to four reference PPI datasets (Table 1) by relating the level of similarity between pairs of gene expression profiles to the observed interactions between corresponding proteins. Two measures of pairwise similarity were used for expression profiles: the Pearson product–moment correlation coefficient and the metric of Eisen *et al*. (15). We calculated these measures of similarity for every gene pair in the dataset and compared their distributions (assumed normal due to the large number of pairs; see also Fig. 2) for all gene pairs and only for those corresponding to the interacting proteins.

Since all pairwise correlation coefficients represent a total population of ORF pairs, we can calculate the true mean of their distribution. Excluding homotypic interactions, with n ORFs there are $n(n-1)/2$ different pairs (2467 ORFs giving 3 041 811 different pairs in the analysed set). The vast majority of these pairs represent true non-interactions: if each protein from the total of n on average has k interaction partners, the proportion of interacting pairs among all pairs would be $k/(n-1)$. For example, for five interaction partners, on average this proportion would be just 5/2466 = 0.2% of this population, however, $k < 3$ for all four datasets.

The analysed PPI datasets represent samples obtained by different techniques that potentially extract members of the small interacting sub-population from the predominantly non-interacting general population. Thus comparing the distributions of correlation coefficients between the general population and a particular sample one can evaluate how successful is the extraction associated with the corresponding technique. The summary statistics for all four datasets are given in Table 2. Results for both the Pearson correlation and Eisen's metric are essentially the same (except for a small shift on the x-axis), so only the Pearson correlation is discussed below as a similarity measure.

All of the PPI datasets have mean pairwise similarities significantly higher than the true mean of all ORFs. The MIPS dataset has by far the highest mean correlation coefficient $R = 0.2$ (16 standard errors above the true mean, $P <$ e–30). On the opposite side is the ITO-NONCORE dataset with $R = 0.07$ (3.5 standard errors above the true mean, $P >$ 2e–4). Additionally, the MIPS dataset agrees with the expression data significantly better than the other three PPI datasets (the mean R is higher than those in the nearest scoring ITO-CORE and UETZ-ALL, $P <$ e–4).

Although the mean correlation coefficient is an essential parameter, more important is the proportion of strongly correlated pairs—these belong to clearly co-expressed genes. About one-fifth of all correlation coefficients in the MIPS dataset are 0.5 and above, signifying meaningful correlations. For three other samples such high coefficients are found for between 8 and 14% of the ORF pairs, while for the total population ~7%

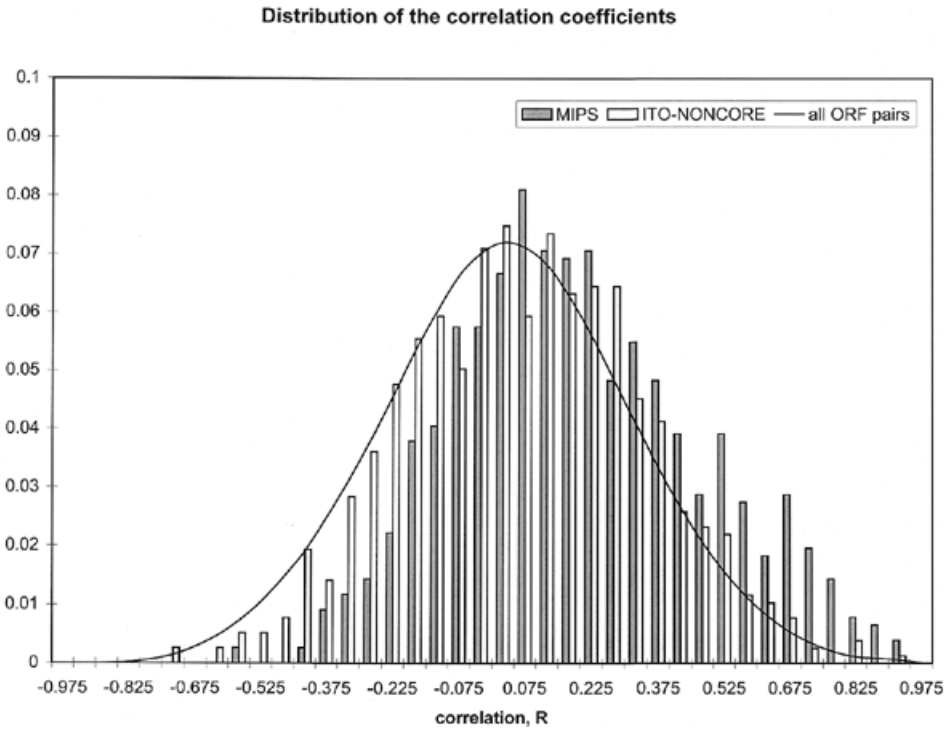

Figure 2. Distributions of the pairwise correlation coefficients of gene expression profiles for interacting proteins in the MIPS and ITO-NONCORE datasets and for all ORF pairs.

of the ORF pairs have $R \geq 0.5$. In other words, a strongly correlated gene pair encodes proteins that are three times more likely to be found interacting with each other in the MIPS dataset than a random pair.

While a significant positive correlation indicates possible co-expression and increases the chances of interaction, a significant negative correlation suggests that the genes are expressed in the opposite phase (while one is expressed, the other is not, and *vice versa*) so their products are less likely to interact. Notably, <1% of all pairs in the MIPS and ITO-CORE datasets have $R < -0.5$, while the total population and the other two datasets contain 1.5–3% of such negatively correlated pairs.

The data in the MIPS dataset have been compiled from a number of sources; their common feature (in contrast to the other three datasets) is in smaller scale data collection, which is often supported by additional evidence. The resulting high reliability is the likely main reason for the observed best agreement of this dataset with the gene expression data. The relative reliability argument apparently also works for the ITO-CORE dataset, where each interaction has been detected at least three times. In contrast, the ITO-NONCORE dataset shows the least correlation with expression profiles. The difference between the MIPS and ITO-NONCORE datasets is clearly seen on the plots of their distributions of the correlation coefficients (Fig. 2).

Thus, provided that gene expression data is a meaningful benchmark, comparing the expression profile similarity distribution for all gene pairs with that for detected interaction pairs one can estimate the relative efficiency of the interaction detection technique in question.

Self-interactions

One class of interactions cannot be tested by co-expression. For every gene in a genome, there is always one that is perfectly co-expressed with it. That is the gene itself. Interestingly, a relatively high percentage of proteins appear to be involved in homotypic interactions, especially in the T7 interaction matrix.

Notably, many interaction events correspond to the symbols on the main diagonal of the matrix (Fig. 1) and these represent self-interactions. Some of them may be solely due to intramolecular associations (3), while others are likely to reflect true oligomeric association of full-length proteins. On the other hand, in the compactly organised genome of T7 several genes encode two alternative products, which often form homo- and heterodimers as well as multimeric complexes (e.g. the diagonal and near-diagonal symbols for helicases/primases 4A and 4B, major and minor head proteins 10A and 10B and unknown genes 18.5 and 18.7).

Publications of large-scale PPI data tend to concentrate on creating interaction networks from heterotypic interactions: for example, Schwikowski *et al*. (17) have completely omitted self-interactions from the datasets analysed. Notably, in the yeast PPI datasets showing the best agreement with the expression profiling results (ITO-CORE, UETZ-ALL and MIPS), the proportion of self-interactions is the highest (Table 1), although not as impressive as the ~40% in phage T7 (counting all symbol types; Fig. 1). On the other hand, the number of false positives detected due to intramolecular associations should also be lower in those datasets, compared to T7.

Table 2. Summary statistics of the distributions of correlation coefficients and profile similarity between expression profiles of interacting proteins from the analysed datasets

		All ORFs	MIPS	UETZ-ALL	ITO-CORE	ITO-NONCORE
Total pairs[a]		3 041 811	765	187	187	775
Correlation coefficient, *R*						
	Mean	0.03	0.20	0.11	0.11	0.07
	SD	0.28	0.28	0.29	0.31	0.27
	SE	0	0.01	0.02	0.02	0.01
	Z[b]		15.92	3.43	3.52	3.51
	P[b]		1e–30	3e–04	2e–04	2e–04
	$R > 0.5$[c]	6.9%	19.5%	12.3%	14.4%	8.3%
	$R < -0.5$[c]	2.8%	0.3%	2.1%	0.5%	1.5%
Profile similarity, *S*						
	Mean	0.09	0.24	0.16	0.16	0.12
	SD	0.28	0.27	0.29	0.31	0.27
	SE	0	0.01	0.02	0.02	0.01
	Z[b]		15.27	3.38	3.06	3.05
	P[b]		1e–30	4e–04	1e–03	1e–03
	$S > 0.5$[c]	9.5%	22.2%	16.0%	17.6%	10.6%
	$S < -0.5$[c]	3.1%	0.3%	2.1%	1.1%	1.7%

[a]Pairs with both partners encoded by the genes included in the expression dataset (15).
[b]*Z* score and *P* value for the null hypothesis of no difference between the sample (dataset) mean and the true mean.
[c]Proportion of pairs with such *R* or *S* values.

DISCUSSION

Integrating interaction and expression data

In this paper I have compared existing genome-wide PPI datasets with patterns of expression of the corresponding genes in phage T7 and the yeast *S.cerevisiae*. We used the graphical display of an interaction matrix and the statistical analysis of distributions of the expression profile similarities to investigate links between gene expression and protein interaction. Both methods confirmed that even on the genome-wide scale of such different living systems co-expressed genes encode proteins that are more likely to interact with each other than with other random proteins.

The analyses described in this paper provide a basis for evaluating different PPI datasets in terms of their agreement with the expression data. The MIPS dataset, with interactions collected from the literature and often supported by additional evidence, shows by far the best agreement with the *S.cerevisiae* gene expression profiles. This indicates a high reliability of this dataset, if the gene expression profiles represent a meaningful benchmark. One should bear in mind that these PPI data have been produced based on knowledge and conjectures about members of protein complexes, rather than by painstakingly probing every pair of ORFs. This makes the MIPS dataset predestined to perform better than the other three and, since such a result was expected, the described approach to data comparison appears to be valid.

The three other datasets have been produced by a more exhaustive search for interacting pairs on a whole genome scale (Table 1). They also show significantly higher mean profile similarity, compared to all ORF pairs (Table 2). We put together PPI data from Schwikowski *et al*. and Uetz *et al.* (17,18) into the UETZ-ALL dataset to obtain a large enough intersection with the expression data. However, the data of Ito *et al*. (19) were separated into the ITO-CORE dataset, where each interaction has been detected at least three times, and ITO-NONCORE, with all remaining interactions. Following the relative reliability argument, ITO-NONCORE shows the least correlation with the gene expression profiles.

Large-scale biological experimentation is often accompanied by elevated noise levels. In genome-wide PPI screens a simple requirement of multiple detection of any interaction seems to decrease the noise and leads to a better agreement with expression data, as is the case for the ITO-CORE and UETZ-ALL datasets. In addition to higher mean profile similarity, these datasets also contain a higher proportion of strongly correlated pairs compared to ITO-NONCORE. Somewhat surprisingly in this respect, ITO-NONCORE contains a lower share of negatively correlated pairs than the UETZ-ALL dataset.

One feature of these three datasets is a low number of common interactions (19). Hazbun and Fields (20) have commented on this, suggesting that functional hypotheses arising from the large PPI networks would require validation by conventional single protein analyses. They have also speculated that the low degree of overlap between the datasets may be indicative of an underestimated size of the yeast interactome. While this remains a definite possibility, a higher error

level in large-scale PPI screening (not confirmed by additional experiments) is another plausible explanation.

On the proteome level, a general population of all possible protein pairs can be divided into two sub-populations: interacting and non-interacting. Whole-genome PPI screens attempt to find the border between the two populations by identifying all interacting pairs using a certain method. For various reasons, mainly related to inability within the yeast two-hybrid system to represent a wide range of environments required for different interactions (membrane association, post-translational modifications missing in yeast, potential toxicity and many others), false negatives may occur in a screen. Also, false positives may be obtained as interactions that do not actually take place in a cell (between two proteins not normally expressed in the same subcellular compartment or due to reporter activation dependent on only one of the fusion proteins, etc.). Thus, taking expression profiles as a reference and comparing the distributions of their correlation coefficients between the general population and a particular sample one can evaluate how successful the technique at hand is in identification of the interacting sub-population.

This increases the value of genome-wide gene expression profiling as a tool for studying functional properties of a proteome, such as PPI and protein complexes. However, expression data also contain errors, often related to incorrect signal measurement (background noise, poor reproducibility, hybridisation peculiarities, etc.). The described approach to integration of gene expression and PPI data offers an opportunity for cross-validation of the results of both experimental approaches.

Further, in addition to gene expression there are other factors affecting cellular protein levels (translational regulation, post-translational changes, stability, etc.). In many cases, protein interaction will depend on these factors. Data generated by additional experimental techniques (e.g. mass spectrometry) will be integrated on a large scale with gene expression and PPI results to produce more reliable biological models, followed by careful testing of individual pathways and complexes in the laboratory.

Self-interactions, quaternary structure and evolution

Both the T7 and yeast PPI datasets contain many self-interacting proteins. These are often neglected in the publications on large-scale PPI projects. However, biochemical and structural evidence together with PPI results suggest that proteins in a cell frequently form homodimer or oligomer structures and this is directly related to their function (7,21–25). There are several evolutionary considerations related to this observation.

Mutation pressure. Self-interaction represents an economical way to construct composite symmetrical shapes using one building block many times, compared to encoding the whole structure in a single polypeptide chain. Imagine that symmetry is required for an oligomeric complex of n units to function. A nonsense mutation anywhere in its sequence would result in a non-functional structure and that would happen n times less frequently if it is built from short blocks rather than one long chain (a similar argument applies to heteromeric protein complexes consisting of smaller proteins, compared to one polypeptide chain). Further, a mutation leading to an amino acid change may destroy the symmetry in a long chain while in the case of one small block it would still lead to the same change in every block, possibly still preserving functional symmetry.

Selection pressure. Imagine a mutation that leads to creation of a homodimer from previously non-interacting monomer units and this improves the organism's fitness. If several ribosomes participate in sequential translation of the same mRNA, multiple copies of the same protein are produced in the same area of the cell at the same time. Then the chance of them interacting is high, the fitness improvement is very likely to manifest itself and the mutation is likely to be fixed. The opposite case is when mutation in one gene could lead to a possible interaction of its product with another protein but they are not usually present in close proximity until either of them is degraded. Then the chances of the complex forming and the fitness improvement taking place are very slim. This line of reasoning clearly applies to any interaction between co-expressed proteins, but it is an intrinsic property of homotypic interactions.

A transition from monomeric ancestors to oligomeric protein structures is a non-trivial evolutionary step in terms of sequence and structural changes (24). Several models of homodimer formation have been proposed, including interchanging domains (21), covalent dimerisation (22) or a combination of these (24). However, for each of these models the selectionist argument above would apply as one of the factors in the evolution of the oligomeric quaternary structure.

Drawing the bigger picture

In elucidating the biomolecular circuitry of a living cell it will be important to determine the connection between genome-scale datasets of measured cellular effects, such as those provided by rapidly developing genomics and proteomics techniques. There have been attempts to combine these types of experimental data with computational analyses to predict gene function (26), sometimes with contradictory results (17). While relating the gene expression data to the large PPI datasets in yeast is one way to establish such connections, an example of the smaller genome of phage T7 gives a flavour of the things to expect for other genomes.

Although cellular protein levels depend on the corresponding gene expression levels, there is still the chance for a wide variation due to translational regulation, post-translational modification and protein (or complex) stability. Unlike in the yeast, genome-wide protein synthesis patterns have been directly observed during the life cycle of T7. Additionally, the synthesis, modification and stability of individual mRNAs have been studied for this phage and information about promoter strength, terminator efficiency and the exact locations of mRNA cleavage sites is available (11).

Although the connection between protein synthesis and interaction in T7 is impressive (Fig. 1), there are a few more remarkable phenomena known, which link its interaction matrix to the T7 life cycle. For example, it has recently been shown (12) that elevated transcription levels lead to increased mutation rates for the highly transcribed genes of T7 (classes IIIa and IIIb, Fig. 1). Interestingly, these genes encode the important structural, host range and DNA maturation proteins. Thus, higher transcription levels of these classes might

ultimately affect the infection and spread capabilities of T7. Even more striking is the succession of the genes expressed and the switch between transcription by *E.coli* RNAP and T7 RNAP. The order of transcription (left to right, in gene blocks, despite the fact that the strongest promoters are on the right) suggests that both RNAPs serve as molecular motors, gradually pulling into the host cell the phage DNA as it is being transcribed (27,28).

While simplified and schematic, this picture gives an idea of the type of interconnections between different cellular processes, both at the level of mechanism and effect, that will eventually be unveiled through large-scale genome and proteome studies of more complex organisms, including human.

This can only be achieved by integrating a number of various data sources with what is currently known about individual genes and proteins. The approach described in this paper could be used to combine and test the agreement of large-scale datasets with existing (and disperse) knowledge of the cellular processes, which may be extracted from the literature using automated systems (29).

ACKNOWLEDGEMENTS

I would like to thank S. Schlenker, C. Ahrens, D. Bankroft, I. Ivanov and C. Schaab for discussions or comments on the manuscript.

REFERENCES

1. Dandekar,T., Snel, B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.
2. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
3. Bartel,P.L., Roecklein,J.A., SenGupta,D. and Fields,S. (1996) A protein linkage map of *Escherichia coli* bacteriophage T7. *Nature Genet.*, **12**, 72–77.
4. Nakai,H. and Richardson,C.C. (1986) Interactions of the DNA polymerase and gene 4 protein of bacteriophage T7: protein-protein and protein-DNA interactions involved in RNA-primed DNA synthesis. *J. Biol. Chem.*, **261**, 15208–15216.
5. Kim,Y.T., Tabor,S., Churchich,J.E. and Richardson,C.C. (1992) Interactions of gene 2.5 protein and DNA polymerase of bacteriophage T7. *J. Biol. Chem.*, **267**, 15032–15040.
6. Sousa,R., Chung,Y.J., Rose,J.P. and Wang,B.C. (1993) Crystal structure of bacteriophage T7 RNA polymerase at 3.3 Å resolution. *Nature*, **364**, 593–599.
7. Egelman,H.H., Yu,X., Wild,R., Hingorani,M.M. and Patel,S.S. (1995) Bacteriophage T7 helicase/primase proteins form rings around single-stranded DNA that suggest a general structure for hexameric helicases. *Proc. Natl Acad. Sci. USA*, **92**, 3869–3873.
8. Zhang,X. and Studier,F.W. (1997) Mechanism of inhibition of bacteriophage T7 RNA polymerase by T7 lysozyme. *J. Mol. Biol.*, **269**, 10–27.
9. Jeruzalmi,D. and Steitz,T.A. (1998) Structure of T7 RNA polymerase complexed to the transcriptional inhibitor T7 lysozyme. *EMBO J.*, **17**, 4101–4113.
10. Hadden,J.M., Convery,M.A., Declais,A.C., Lilley,D.M. and Phillips,S.E. (2001) Crystal structure of the Holliday junction resolving enzyme T7 endonuclease I. *Nature Struct. Biol.*, **8**, 62–67.
11. Studier,F.W. and Dunn,J.J. (1983) Organization and expression of bacteriophage T7 DNA. *Cold Spring Harb. Symp. Quant. Biol.*, **47**, 999–1007.
12. Beletskii,A., Grigoriev,A., Joyce,S. and Bhagwat,A.S. (2000) Mutations induced by bacteriophage T7 RNA polymerase and their effects on the composition of the T7 genome. *J. Mol. Biol.*, **300**, 1057–1065.
13. Cheetham,G.M. and Steitz,T.A. (1999) Structure of a transcribing T7 RNA polymerase initiation complex. *Science*, **286**, 2305–2309.
14. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M. *et al.* (1996) Life with 6000 genes. *Science*, **274**, 563–567.
15. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
16. Fields,S. and Song,O. (1989) A novel genetic system to detect protein-protein interactions. *Nature*, **340**, 245–246.
17. Schwikowski,B., Uetz,P. and Fields,S. (2000) A network of protein-protein interactions in yeast. *Nat. Biotechnol.*, **18**, 1257–1261.
18. 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) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
19. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
20. Hazbun,T.R. and Fields,S. (2001) Networking proteins in yeast. *Proc. Natl Acad. Sci. USA*, **98**, 4277–4278.
21. Bennett,M.J., Choe,S. and Eisenberg,D. (1994) Domain swapping: entangling alliances between proteins. *Proc. Natl Acad. Sci. USA*, **91**, 3127–3131.
22. D'Alessio,G., (1995) Oligomer evolution in action? *Nature Struct. Biol.*, **2**, 11–13.
23. Barker,A., Fickert,R., Oehler,S. and Müller-Hill,B. (1998) Operator search by mutant Lac repressors. *J. Mol. Biol.*, **278**, 549–558.
24. Ciglic,M.I., Jackson,P.J., Raillard,S.A., Haugg,M., Jermann,T.M., Opitz,J.G., Trabesinger-Ruf,N. and Benner,S.A. (1998) Origin of dimeric structure in the ribonuclease superfamily. *Biochemistry*, **37**, 4008–4022.
25. Karlsson,C., Jornvall,H. and Hoog,J.O. (1991) Sorbitol dehydrogenase: cDNA coding for the rat enzyme. Variations within the alcohol dehydrogenase family independent of quaternary structure and metal content. *Eur. J. Biochem.*, **198**, 761–765.
26. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**, 751–753.
27. McAllister,W.T., Morris,C., Rosenberg,A.H. and Studier,F.W. (1981) Utilization of bacteriophage T7 late promoters in recombinant plasmids during infection. *J. Mol. Biol.*, **153**, 527–544.
28. Zavriev,S.K. and Shemyakin,M.F. (1982) RNA polymerase-dependent mechanism for the stepwise T7 phage DNA transport from the virion into *E. coli*. *Nucleic Acids Res.*, **10**, 1635–1652
29. Blaschke,C., Andrade,M.A., Ouzounis,C. and Valencia A. (1999) Automatic extraction of biological information from scientific text: protein-protein interactions. *Intell. Syst. Mol. Biol.*, **7**, 60–67.
30. Fellenberg,M., Albermann,K., Zollner,A., Mewes,H.M. and Hani,J. (2000) Integrative analysis of protein interaction data. *Intell. Syst. Mol. Biol.*, **8**, 152–161.