kluwer
the language of science

JOURNAL OF STRUCTURAL AND FUNCTIONAL GENOMICS (JSFG)

Dear Author(s),

This is to inform you that the PDF proof of your paper for publication in *Journal of Structural and Functional Genomics* has been sent to you by e-mail. You may find, in addition, a Query Form detailing any questions we have regarding your article (as a separate attachment). Please **do not** attempt to edit the PDF file (including adding "post-it" type notes).

This PDF file has been produced automatically from an electronic database format; the original file is maintained by Kluwer Academic Publishers. Please note that certain details of page layout may still need to be amended before printing. However, the final, printed product will conform to our usual high standards for page layout and image resolution.

Corrections must be kept to a minimum. Please use the proofs solely for checking the typesetting and editing, as well as the completeness and correctness of the text, tables and figures. Changes to the article as accepted for publication will not be considered at this stage. Also, make sure that you answer any questions (see the query form) that have arisen during the preparation of your proofs.

These proofs will be read and corrected only by yourself. Kluwer Academic Publishers will not proofread the paper after it has been returned, and will not assume responsibility for any errors that the author has failed to correct in the proofs.

Your response, regardless of whether corrections are required or not, should reach us within 72 hours of the receipt of material. The e-mail address is: mail@kaplanc.co.uk

If, for any reason, you are unable to reply by e-mail, please mark the corrections and any other comments (including replies to the Query Form) on a printout of the PDF file and fax this to +44 (0)1524 32144 or mail to:

Kluwer Academic Publishers
Unit 3.3.4, Storey House
White Cross
Lancaster
LA1 1PE
United Kingdom

Your offprints are usually not available until three weeks after publication of the journal issue, and then they are sent out by surface mail. Consequently, it may take 2-3 months for offprints to reach authors outside Europe.

KLUWER ACADEMIC PUBLISHERS
Manufacturing Dept.

kluwer
the language of science

OFFPRINT ORDER

MS no: **2001-010**

PE: 5093362

Dr M Gerstein
Dept of Molecular Biophysics and Biochemistry
Yale University
266 Whitney Avenue
PO Box 208114
New Haven, CT 06520

Re: **Integration of genomic datasets to predict protein complexes in yeast**
R Jansen, N Lan, J Qian, M Gerstein

To be published in
Journal of Structural and Functional Genomics

Dear Dr M Gerstein
This is to let you know that the above paper has gone into production and will appear in due course. Offprints of your paper may be ordered by filling in and returning this form.

I would like to receive

50 offprints free of charge

_____ additional offprints without cover (minimum of 50 extra offprints)

Orders for offprints are only accepted if received with payment or accompanied by an official purchase order from your institution, failing which no additional offprints can be produced. Payment can be made by credit card, bank draft, personal cheque, international money order. Payment is accepted in any hard currency. Prices of additional offprints are mentioned on the enclosed price grid. Make cheques payable to 'Kluwer Academic Publishers - Dordrecht'.

[ ] I enclose payment to the amount of _____
OR
[ ] Please charge my credit card:
VISA [ ] Access [ ] American [ ] Bank Americard [ ] Diners Club [ ]
Eurocard Mastercard Express

Card no _____Expiry date_____

Signed _____

VAT number (if applicable) _____

Complete and return this form to:

Ms Linda Thomas
Assistant Editor
Kluwer Academic Publishers
PO Box 55
Lancaster LA1 1PE
United Kingdom
Fax: +44 (0)1524 32144

OFFPRINTS: Terms of Delivery

1. A minimum of 50 offprints may be ordered. Prices corresponding to the number of pages and quantities ordered are given below.
2. Only the corresponding author will receive the offprint order form. That author is therefore also responsible for any orders the co-authors wish to make. All orders for a particular paper should appear together on only one form.
3. Offprints are printed at the same time that the book or journal is printed. Thus, no alterations from the exact form in which the article appears in the book or journal are possible. Additionally this makes it impossible to fulfill any orders for offprints that are received after the book or journal is printed.
4. Offprints will be despatched within a short time after the appearance of the published paper.
5. No additional offprints can be printed or sent unless the order form is filled in, signed, returned with accompanying payment or official purchase order and received before the book or journal goes to press.
6. Any correspondence in connection with offprints must state the name of the book or journal, title and **code number** of paper and name(s) of the author(s). All communications should be sent to

   Kluwer Academic Publishers, PO Box, Lancaster LA1 1PE, United Kingdom

   Tel: +44 (0) 1524 34996; Fax: +44 (0) 1524 32144; E-mail: mail@kaplanc.co.uk

Prices of additional offprints are in EUROs

Number of copies	Number of pages					
	1–4	5–8	9–12	13–16	17–20	+4
50	108	177	248	318	388	71
100	153	238	323	407	492	85
150	200	300	400	500	599	100
200	246	361	476	591	706	115
250	292	422	551	681	810	130
300	339	483	628	772	917	145
350	385	544	703	863	1022	159
400	431	606	780	955	1129	175
450	478	667	857	1046	1235	189
500	524	728	932	1136	1340	204
+50	50	+62	+72	+92	+103	+15

Payment will be accepted in any convertible currency. Please check the rate of exchange with your bank.
If ordering from within The Netherlands please add 6% purchase tax (BTW) to the prices quoted above.

Customers within the EU must consider the following rules:

- **If you are in possession of a VAT identification number, please fill in the VAT number on the order form. You will not be charged VAT.**
- **If you do not have a VAT number, then please add the low VAT rate - applicable to your country - to the prices quoted above.**

kluwer
the language of science

PROOF CORRECTION MARKS TO BE USED BY AUTHORS

MARGINAL MARK	MEANING	CORRESPONDING MARK IN TEXT
⌀	Delete (take out)	/ or ⊢ Cross through
⌀	Delete and close-up	⌒ / ⌣ Above and below matter to be taken out
stet	Leave as printed (when matter has been crossed out by mistake)	····· Under matter to remain
caps.	Change to capital letters	≡ Under letters or words altered
l.c.	Change to lower case letters	Encircle letters altered
bold	Change to bold type	~~ Under matter altered
bold ital.	Change to bold italic type	≈ Under matter altered
ital.	Change to italics	— Under matter altered
rom.	Change to roman type	Encircle matter altered
×	Replace by similar but undamaged character or remove extraneous marks	Encircle letter to be altered
⅄	Insert (or substitute) superior figure or sign	⅄ or /
⅃	Insert (or substitute) inferior figure or sign	⅄ or /
⊂⊃	Insert (or substitute) hyphen	⅄ or /
EN ⊢⊣	Insert (or substitute) dash	⅄ or /
ⓛ	Insert (or substitute) solidus	⅄ or /
···	Insert (or substitute) ellipsis	⅄ or /
⌒	Close-up–delete space	⌒ Linking words or letters
#	Insert space	⏐ or ⊤ Between items
equal #	Make spacing equal	⏐ Between items
⊤	Reduce space	⏐ or ⊤ Between items
⊱	Insert space between lines or paragraphs	
←	Reduce space between lines or paragraphs	
⌐¬	Transpose	⌐¬ Between letters or words, numbered when necessary
5	Transpose lines	5
centre	Place in centre of line	⌋ ⌊ Around matter to be centred
⊢⊐	move to the left	⊐
⊏⊣	move to the right	⊏
NP	Begin a new paragraph	□ Before first word of new paragraph
run on	No fresh paragraph here	⟶ Between paragraphs
⅄	(Caret mark.) Insert matter indicated in margin	⅄
‘ ’ “ ”	Insert single/double quotes	⅄ ⅄

Remarks

To indicate a substitution, simply cross out the letters or words to be replaced, and write the correct letters or words in the margin. It is not necessary, nor even desirable, to use the marks for *delete* and *insert* when making a substitution. If there is more than one substitution in a line, place them in the correct order in the margin, and indicate the end of each correction with an oblique stroke / . Alternatively, continental location marks may be used, but these are to be placed in front of the corrections, not behind as in the case of the oblique stroke.

The typesetter treats *all* letters and words in the margin as insertions or substitutions, so – to avoid misunderstanding – any comments *not* intended to form part of the text should be encircled.

All alterations should be marked clearly so that there is no risk of misunderstanding; long additions or amendments should be typed on separate slips and attached. *Only really essential alterations should be made at proof stage.*

In addition to reading the proofs, please look through your edited manuscript to see if there are *any queries from the copy editor*, and if so, answer the queries *on the proofs.*

Journal of Structural and Functional Genomics **00:** 1–11, 2002.

Integration of genomic datasets to predict protein complexes in yeast

Ronald Jansen[1], Ning Lan[1], Jiang Qian[1] & Mark Gerstein[1, 2, *]
*[1]Department of Molecular Biophysics & Biochemistry and [2]Computer Science, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA; * Author for correspondence (tel: (203) 432−6105; fax: (360) 838−7861; e-mail: mark.gerstein@yale.edu)*

Received 15 January 2002; Accepted in revised form 23 May 2002

Key words: ???

Abstract

The ultimate goal of functional genomics is to define the function of all the genes in the genome of an organism. A large body of information of the biological roles of genes has been accumulated and aggregated in the past decades of research, both from traditional experiments detailing the role of individual genes and proteins, and from newer experimental strategies that aim to characterize gene function on a genomic scale.

It is clear that the goal of functional genomics can only be achieved by integrating information and data sources from the variety of these different experiments. Integration of different data is thus an important challenge for bioinformatics.

The integration of different data sources often helps to uncover non-obvious relationships between genes, but there are also two further benefits. First, it is likely that whenever information from multiple independent sources agrees, it should be more valid and reliable. Secondly, by looking at the union of multiple sources, one can cover larger parts of the genome. This is obvious for integrating results from multiple single gene or protein experiments, but also necessary for many of the results from genome-wide experiments since they are often confined to certain (although sizable) subsets of the genome.

In this paper, we explore an example of such a data integration procedure. We focus on the prediction of membership in protein complexes for individual genes. For this, we recruit six different data sources that include expression profiles, interaction data and essentiality and localization information. Each of these data sources individually contains some weakly predictive information with respect to protein complexes, but we show how this prediction can be improved by combining all of them. Supplementary information is available at http://bioinfo.mbb.yale.edu/integrate/interactions/.

Abbreviations: ???

Introduction

With the recent flux of genome sequences comes the challenge for functional genomics to ascribe biological information, including structure, localization, function and regulation, to every gene in the genome. Numerous experiments to study the genome, transcriptome or proteome of organisms have become commonplace, and new algorithms are being developed to help turn the rapidly increasing amount of whole-genome data into useful biological knowledge. For example, microarray experiments measure mRNA expression under various cellular conditions and are currently one of the most prominent experiment approaches [1–5]. The expression profile of a gene can shed light on its cellular function, and relate genes with similar or opposite functional roles. To measure gene function in terms of mutant phenotype, genome-wide deletion and transposon disruption strategies have been developed [6, 7]. Protein chips can directly assay the properties of proteins [8, 9]. Another major experimental area is the yeast two-hy-

brid assay which detects genome-wide protein–protein interactions and allows the construction of networks from which protein function and regulation can be inferred [10, 11]. In addition to these relatively new developments, there is of course also the large body of biological knowledge accumulated in the past decades of research.

However, relying on any one of these methods or data sources aloneis often not sufficient to unambiguously determine the function of uncharacterized genes. There are many examples of combining different genomic-scale data sources in the literature. The trivial case is the integration of *two* data sources. This is often the minimum amount of integration needed to interpret a genomic-scale experiment. This point might be so obvious that most researchers would not view it under the angle of data integration. For instance, previous efforts to interrelate information from two genomic datasets include analyzing expression data by a variety of supervised and unsupervised methods and comparing to functional categories, transcription-factor binding sites, protein families, protein–protein interactions, and protein abundance [12–26].

There have been considerably fewer attempts to integrate *more than two* types of whole-genome data. One example was the combination of expression correlations, phylogenetic profiles and patterns of domain fusion to predict protein function [27]. In another study, a Bayesian framework was used to integrate expression, essentiality, and sequence motif data for the prediction of protein subcellular localizations [28, 29].

There are several benefits of combining experimental and computational data sources. Often, one may be able to uncover non-obvious and potentially significant relationships, such as those between expression and chromosomal positioning or subcellular localization [29, 30].

Moreover, the integration of multiple sources obviously increases the range of the genome that can be characterized. This benefit of increasing coverage is obvious for integrating many of the experiments for individual genes or proteins, but is also valid for the combination of multiple genomic-scale experiments. Because of experimental limitations, it is in many cases difficult to conduct experiments that really include the complete genome. Thus, many genomic-scale experiments have been performed on sizable but only limited fractions of the genome.

When multiple experiments cover the same genes, then there are other benefits from combining data. Do the experiments agree, thus confirm each other and increase the confidence in the results, or did they yield conflicting information, thus leaving the result open for further investigation? In general, the combination of different data sources should help to increase the reliability of the interpretation of experimental results.

Of course, these last two goals of increasing coverage and reliability tend to be in conflict with one another. The reliability of information confirmed by independent sources usually increases, but in general the more sources are required to agree, the fewer the number of instances where this is the case. This is because the intersection between two datasets is always equal or less in size than the two data sources individually. Thus, one often has to find the right trade-off between coverage and reliability.

In this paper, we look at one particular example of data integration to discuss the issues mentioned above. Specifically, we focus on the prediction whether two yeast proteins are members of the same protein complex or not. We propose combining expression and interaction datasets and essentiality and subcellular localization data to this end. In order to judge whether the prediction is successful, we use the MIPS complexes catalog as the standard for known protein complexes [31]. Our study is preliminary but intended to show possible ways of combining new genome-wide datasets to ultimately determine all protein complexes. Similar ways of combining genome-wide datasets for predicting other kinds of biological information, such as biological functions or pathways could be possible as well.

What we are trying to do here is not so much characterizing or functionally defining individual genes, but rather *pairs* of genes or proteins that interact with one another in a complex.

There are a few reasons why we concentrate on protein complexes. Yeast-two hybrid data, one of the data types we use, can potentially be used to predict protein complexes. In addition, protein complexes also have a variety of nice properties that can be exploited for our data analysis. We start with the assumptions that:

1. The function of any protein complex depends on the function of its subunits; thus a complex is dysfunctional if one of its subunits is dysfunctional or missing; and

2. There are a variety of protein properties that should be shared by all subunits of a complex (for instance, if the complex has a particular biochemical function, then this most likely also provides a functional definition for its subunits).

Although these assumptions are rarely strictly met in reality, they provide some practical help for our task. Assumption #1 has implications for the prediction of protein complexes with expression data, as we will show below. Assumption #2 allows us to make use of essentiality and localization data for complex prediction.

In a previous publication, we have shown that the subunits of permanent protein complexes have a significant tendency to be coordinated in terms of their mRNA expression levels [21]. This can be explained by assumption #1). If the function of the complex were dependent on the presence of all its subunits, it would be energetically costly for the cell to express them in an uncoordinated and haphazard fashion.

Methods

Data sources

Many of the data sources we list in the following paragraphs might be only weakly predictive with respect to protein complexes and they may lead to many false positives and negatives if taken individually. However, we show later that combining the individual datasets can still lead to a relatively reliable prediction of protein complexes. Furthermore, it will become evident which data sources contribute most or least to the prediction.

Expression data
Two expression datasets were used: a cell-cycle experiment [32] and the Rosetta yeast compendium [33]. The two datasets represent different experimental methodologies and provide a reasonable sampling of the possible cellular states of yeast. The cell-cycle data contain expression profiles obtained from synchronized cells over the course of two cell cycles, whereas the Rosetta data contains genome-wide expression ratios for 300 stationary cell states, which are derived from 280 gene deletions and the 20 drug interaction experiments.

For the Rosetta data, we focused on those protein pairs whose correlation exceeds a certain threshold (0.52). This selects about 300 000 protein pairs from among the 18 000 000 theoretically possible.

For the time course data of the yeast cell cycle, we not only looked at regular correlations but also at correlations for time-shifted and inverted expression profiles. In this case, the threshold criterion was a match score of 13 [23].

These selection criteria are arbitrary in some sense, and other criteria (such as excluding genes that do not change at least two-fold in expression) are possible. However, our simple purpose here was to create datasets of protein pairs of manageable size that are likely related to protein complexes.

Predictive information of expression data
In this section, we would like to survey the ability of expression data to predict membership in protein complexes, and particularly address the following two questions:

1. To what extent can we predict that a protein belongs to a complex based on its expression correlations? Conversely, to what extent can we predict the expression correlation of a pair of proteins, given they are in a complex?
2. To what extent are pairs of proteins with highly correlated expression levels accounted for by relationships other than membership in a complex, e.g., being in the same metabolic pathway?

A simple way to analyze these questions is to look at the conditional probability $P(class|C)$ that two proteins are in the same particular class (e.g., functional class or complex) given that their expression profiles have a particular correlation C. We expect this conditional probability to increase with rising correlation C. Unfortunately, it is difficult to compare this variable between complexes differing in size, as we are much more likely to find two randomly selected proteins within a large complex than a small one. We, therefore, compare the odds ratio $P(complex|C)/P(complex)$ between different complexes. In this case, the probability of two random proteins being in the same protein complex $P(complex)$ functions as a normalization factor correcting for complex size. To better understand the meaning of this ratio, we can rewrite it applying Bayes' rule:

$$P(complex|C)/P(complex) = P(C|complex)/P(C)$$

We can see that the right-hand side of the equation represents the distribution of correlation coefficients

of the pairs with known biological interactions divided by the distribution of correlation coefficients of all possible pairs of genes in this genome.

Figure 1A shows the results of the odds-ratio calculations, addressing the first question above. As expected, the odds for finding a protein in a complex increase with higher levels of correlation. Note, however, this increase is much steeper for the permanent complexes than transient ones. This means that a highly correlated pair of genes has much greater odds of being in a permanent complex, by an order of magnitude or more. If we factor in that there are in general many more interactions in permanent complexes than transient ones, we can see that there is an overwhelmingly greater chance that a highly correlated pair of genes will be in a permanent complex than a transient one. Specifically, by adding up the interactions, we can see there are $\sim 13:1$ odds of finding a pair of proteins in a permanent complex as opposed to a transient one, independent of gene expression. However, if genes have an expression correlation close to 1, the odds rise to $\sim 1530:1$. Conversely, if the genes have a correlation close to -0.5, then the odds drop precipitously to $-1:9$. (Due to their size and great degree of correlation, the cytoplasmic ribosomes could potentially skew the results. Consequently, in the figure, we show the results for the permanent complexes, with and without the ribosome.) Overall, one can observe that in the high correlation coefficient region, the overall likelihood of belonging to a protein complex for two genes is much higher than expected because their odds ratios are much larger than 1. On the other hand, in the low correlation coefficient region, the likelihood of finding interactions is either close to or lower than expected according to their odds ratios. The likelihood of finding two genes belonging to a protein complex increases monotonically with the expression-profile correlation coefficient, which means there is some predictive information for protein complexes in the gene expression data.

Figure 1B addresses the second question, comparing the odds ratios for protein complexes with those for proteins belonging to the same metabolic pathway [31]. The observed odds ratios are similar to those for transient complexes. This indicates that the odds of highly correlated genes being in the same pathway are similar to those for being in a transient complex but substantially lower than for being in a permanent complex.

Interaction data

It is a straightforward idea to predict membership of protein complexes with existing interaction data. For this purpose, we looked two yeast two-hybrid datasets [10, 11].

The yeast two-hybrid data present in some sense different types of interactions from those among the groups of proteins unified in complexes. This is illustrated in Figure 2. Still, the yeast two-hybrid data can of course contribute to the prediction of protein complexes, although it is far from sufficient in itself and needs to be complemented with other data sources.

Essentiality data

Essentiality data comes from the MIPS database as well as from transposon and gene deletion experiments [6, 7, 31]. We look here at whether two proteins are either both essential or both non-essential as an indicator for membership in the same protein complex. If a complex is essential, then its subunits should also be essential if they are necessary for the function of the complex as a whole.

Localization data

The localization information we use comes from merging data from the MIPS, Swissprot, and YPD databases [28, 31, 34, 35]. If the localization is known, each protein is located in one of the five general compartments: N (nuclear), C (cytoplasm), M (mitochondria), E (extracellular environment or secretory pathway), T (transmembrane). If two proteins are in the same compartment, we use this as an indicator of potential membership of the same protein complex.

MIPS complexes catalog

The MIPS complexes catalog provides a complete list of the currently known protein complexes in yeast [31]. We extracted all possible protein pairs within the same complexes from the complexes catalog. We used this list in order to judge the performance of our prediction.

We systematically removed all those protein classes from the catalog that do not really represent complexes, but rather aggregated classes of related proteins. This left us overall with 8250 different protein pairs within the same complexes.

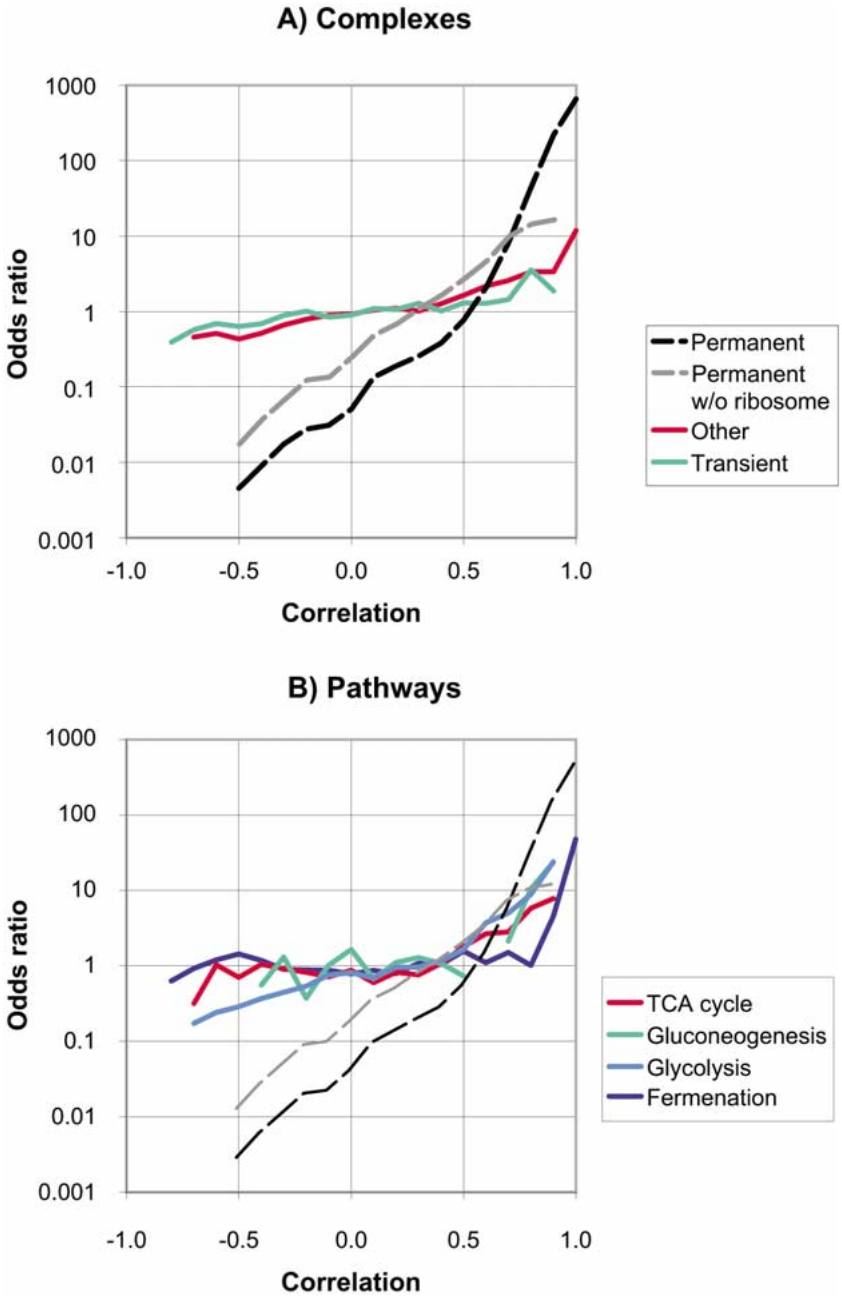

Figure 1. We show different plots of the odds ratio $P(class|C)/P(class)$, which is the ratio of the conditional probability of two proteins with a correlation C being in the same protein class to the probability of finding these two proteins in the same class independent of the correlation. (**A**) focuses on different complex classes. Two proteins are considered to be in the same class if they are both in the same complex. (**B**) shows the odds ratios for four representative pathways compared with those for permanent complexes from (**A**). In this case, proteins are considered to be in the same class if they are both participating in the same pathways (according to the MIPS functional catalog) or if they are directly interacting with one another by a genetic, physical or yeast two-hybrid interaction.

The definitions of permanent and transient complexes can be found in a previous publication [22]. Complexes with 10 or more subunits that are neither classified as permanent or transient are listed as 'other'.

In general, all odds ratios show a comparable significant increase as a function of the correlation C. However, the 'permanent' complexes show the greatest difference between odds ratios for high and low correlations.

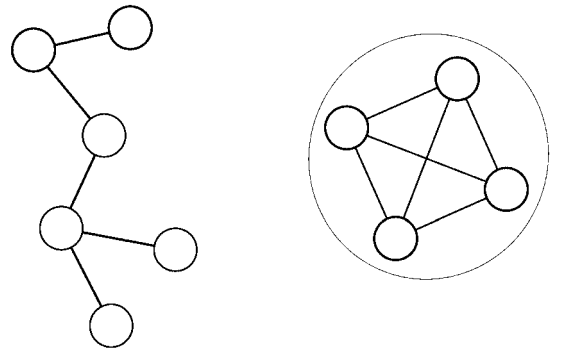

Figure 2. Conceptually, there are two different types of protein 'interactions'. First, there are binary-type interactions between pairs of proteins such as those measured by the yeast-two hybrid system (shown in the left part of the figure). These interactions are mostly of a physical nature, that is, the proteins are interacting with one another through a structural contact interface. Taken together, the collection of these interactions results in a whole network of genome-wide binary links between proteins. Secondly, there are interactions of whole groups of proteins that together form a structural complex (shown in the right part of the figure). Although not all of the subunits in a protein complex are in structural proximity and thus do not physically interact with one another, they form a coherent structural unit as a whole with common properties. For instance, if the protein complex localizes in a particular subcellular compartment, then all its subunits should be present in the same compartment as well. Thus, the subunits share certain properties, regardless of their structural proximity in the complex. The schematic shows the example of a protein complex composed of four subunits, with each link indicating a shared property between two subunits. All subunits are equally linked with the other subunits, thus, the resulting graph is complete with $(4^2 - 4)/2 = 6$ edges between the 4 nodes (proteins).

How to go about combining datasets

How should one go about combining these different data sources to improve prediction? The problem can be thought of as overlapping different protein–protein interaction networks (interactomes). Two different extremes can be imagined. For networks with individually low FP but high FN rates, the benefit of combining data comes from looking at the *union* of the disparate datasets. On the other hand, for networks with individually high FP and low FN rates, it is most useful to look at their *intersection* (see Figure 3).

Given that individual datasets have different FP and FN rates, they should be weighted differently. In general, there should be more effective rules for combining networks. Rather than building the union or intersection for all of the datasets at once, one should look at different combinations of unions and intersections among the datasets [36]. What kinds of data combination rules can one use in order to achieve simultaneously a low error rate and a high coverage of the prediction?

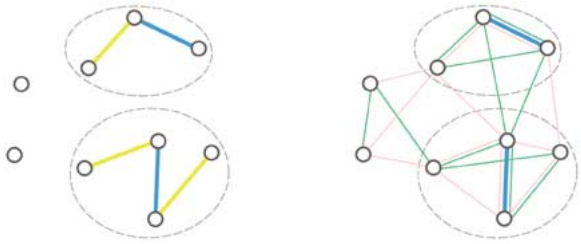

Figure 3. The integration of different datasets to predict membership in protein complexes can be visualized as overlapping different protein–protein interaction networks (interactomes). Two different extremes can be imagined. On the one hand, the networks might be associated with low FP but high FN rates (left). In this situation, the benefit of combining data comes from looking at the *union* of the disparate datasets. On the other hand, the individual networks might be associated with high FP but low FN rates (right). In this case, it is most useful to look at *intersections* of the different datasets. Circles represent proteins, links interactions, and dotted lines known associations.
Individual datasets should be weighted differently, given their different FP and FN rates. The data from some sources might be more reliable than from others. This is illustrated in the right hand panel, where the thicker lines correspond to lower FP rates.

Note that there needs to be some degree of independence or orthogonality between the datasets in order for the integration to work properly.

A more refined integration strategy than just intersections or unions of all datasets would be to first divide the data into all possible combinations of different subsets (see Figure 4).

Then one could go about determining the error rates for each subset by comparing the protein pairs in each subset with the MIPS complexes catalog as a standard. The error rate for a subset is defined as the fraction of false positives among the predicted interactions. We explain, in the results section below, why the error rate rather than the false positive rate [= FP/(FP + TN)] should be the crucial statistic.

Once this is done, the question remains which subsets to include for the best final prediction. It seems best to start with ordering all the subsets with respect to their error rates. One would pick the subset with the lowest error rate first and then successively add the subsets with the lowest remaining error rates. Each time, this would increase the coverage while increasing the error rate by the smallest amount. The process of including more and more subsets (i.e., accepting that the protein pairs in them are positives)

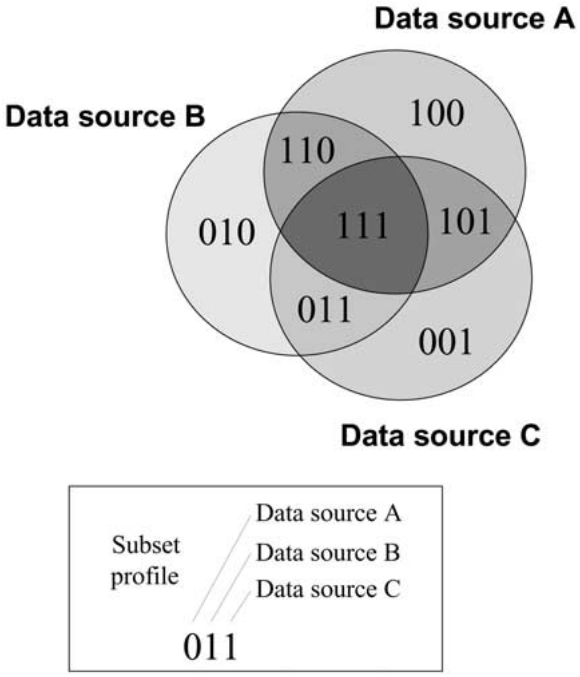

Figure 4. Hypothetical integration of three datasets. We define a binary 'subset profile' for each of the subsets of the Venn diagram. For instance, the profile '101' encompasses all data points that are present in dataset A and dataset C, but not in dataset B. The degree of intersection of a subset can be defined as the sum of the profile. For instance, for profile '101' the degree of intersection is 2, meaning that two data sources are intersecting. These definitions are useful when dealing with more than three data sources (see Figure 5). When integrating an increasing number of data sources, generally the error rates go down if more and more datasets agree with one another (although this is not necessarily the case). In other words, the error rates go down with increasing intersection among the several datasets. In the Venn diagram, darker shaded areas schematically indicate lower error rates. As the Venn diagram shows, there are overall 7 different subsets for 3 data sources. In general, for N data sources, there are $2^N - 1$ different subsets.

While focusing on the subsets of the data with higher degrees of intersection between the different data sources tends to reduce the error rate, it simultaneously reduces the coverage. The higher the degree of intersection of a subset, the fewer data points the subset usually contains.

Thus, an optimal strategy for combining multiple datasets would be to find a reasonable trade-off between the highest possible coverage and the lowest possible error rate. For the schematic example in the figure, one could for instance start by first considering the subset '111' (the intersection of datasets A, B, and C), thus choosing the subset the with the lowest error rate. Then, in order to increase the coverage, one could subsequently add the subsets with the lowest error rates, in this case the subsets '101' and '110', and so on. This would increase the coverage while simultaneously increasing the overall error rate at the lowest amounts. A practical example of this strategy is shown in Figure 5.

An open question is how to determine the error rate. In our example, we use the MIPS complexes catalog as the standard for protein complexes. We look at how the protein pairs in each subset profile compare with this standard and compute empirical error rates based on that.

should stop after a good compromise between coverage and error rate is reached.

Results

How much predictive information is there in the individual datasets?

Table 1 shows the six data sources we used in the prediction of whether two genes belong to the same protein complex. Before combining the data, we investigated to what extent the individual data sources overlap with the MIPS complexes catalog. For instance, for the expression data we asked how likely it is that two genes belong to the same complex based on whether their expression profiles exceed a certain similarity threshold (see Table 1 for details). Table 1 shows the resulting false-positive (FP) and false-negative (FN) rates if gene pairs are solely classified based on these threshold criteria, after comparing it with the MIPS complexes catalog. For both the cell cycle and the Rosetta data, the FP rates are 1.6% whereas the FN rates are well above 50%.

There are many protein pairs predicted to interact according to the individual data sources that are not in the MIPS complexes catalog. We define these as FP. This is, of course, because the MIPS complexes catalog is far from complete (not all protein complex interactions are known); thus the FP either represent protein pairs that do not interact in reality or new interactions not previously recorded in the MIPS complexes catalog. Thus, the FP are 'false' in the context of this classification and a machine-learning sense, rather than in a biological sense. However, the number of FPs we report here can be regarded as an upper bound of the real number of FPs if all protein–protein interactions were known. They also provide a numerical criterion to rank the data sets. If we assume that the protein–protein interactions from the MIPS catalog are a representative subset of all protein–protein interactions, then the ranking of the FP rates should not change very much if all interactions are known. Thus, we can rank the subsets in terms of their quality (see supplementary website).

For the essentiality and localization data, we asked how likely it is that two genes belong to the same complex if they both have the same essentiality and the same localization. Again, Table 1 shows the resulting FP and FN rates.

Table 1. The table shows the six data sources we used in the prediction of whether two genes belong to the same protein complex. Here we look at the data sources individually before combining them. The first two data sources are expression data, from the yeast cell-cycle time-course by Cho *et al.* [32] and from the Rosetta knockout experiments [33]. The third and fourth data sources are both from yeast two-hybrid experiments [10, 11]. The fifth and sixth data sources stem from information about the essentiality of genes [6, 7, 31] and the subcellular localization of their proteins [28, 31, 34, 35].

Data source	TP	FN	TN	FP	FN rate (%)	FP rate (%)	ε (%)
Cell cycle	3466	4784	17703707	296293	58.0%	1.6%	98.8
Rosetta	2772	5478	17720318	279682	66.4%	1.6%	99.0
Y2H (Uetz)	50	8200	17999109	891	99.4%	0.0%	94.7
Y2H (Ito)	78	8172	17995806	4194	99.1%	0.0%	98.2
Essentiality	2718	5532	15520460	2479540	67.1%	13.8%	99.9
Localization	7220	1030	17523304	476696	12.5%	2.6%	98.5

We have shown expression data to be predictive with respect to protein complexes, but so is the information about essentiality and subcellular localization. The reasoning is simply that if two genes belong to the same complex, then they should have the same or similar properties in terms of essentiality and localization of their protein products. If one subunit of a complex is essential, then the other subunits are often essential as well, and if a complex is present in a particular cellular compartment, then its subunits should most likely be present in the same compartment too.

We investigated to what extent we can predict whether two genes are in the same protein complex based on each of these data sources individually. For instance, for the expression data we asked how likely it is that two genes belong to the same complex based on whether their expression profiles exceed a certain similarity threshold. For the cell-cycle data, we looked at gene pairs for which either the regular correlation or the time-shifted and inverted correlations exceed a threshold with a match score of 13 [23]. For the Rosetta data we simply looked at gene pairs that exceeded the regular (Pearson) correlation of 0.52. Each of these criteria yields about 300000 protein pairs. As the standard to decide whether two genes really belong to the same complex, we used the complex catalog from the MIPS database. Table 1 shows the resulting FP and FN rates if gene pairs are solely classified based on these threshold criteria [FP rate = FP/(FP + TN) = 1 − sensitivity and FN rate = FN/(FN + TP) = 1 − specificity]. For both the cell cycle and the Rosetta data the FP rates are 1.6% whereas the FN rates are well above 50%.

For the essentiality and localization data, we asked how likely it is that two genes belong to the same complex if they both have the same essentiality and the same localization. Table 1 shows the resulting FP and FN rates. Note the low FN rate for the localization data, indicating that two proteins with different subcellular localizations are very likely not interacting in a complex, as expected.

At first, the low FP rates seem to be an encouraging result. However, we are facing the problem that there is only a low number of positive relative to negative gene pairs in the yeast genome. There are only 8250 protein pairs in all protein complexes according to the MIPS catalog, but the number of negative pairs is about 18000000 (= $6000^2/2$ given that there are about 6000 genes in the yeast genome). Thus, even a relatively low FP rate results in a relatively high absolute number of false positives. The lower part of the table, showing absolute numbers of false negatives and positives, indicates this. The error rate ε = FP/(TP + FP) represents the fraction of FP gene pairs among all positively predicted gene pairs (with TP being the number of true positives). For each data source the error rate is almost 100%. For an acceptable prediction the error rate should be at least lower than 50%.

At first, the low FP rates seem to be an encouraging result. However, we are facing the problem that there is only a low number of positive relative to negative gene pairs in the yeast genome. Recall that there are only 8250 protein pairs in all protein complexes according to the MIPS catalog, but the number of negative pairs is about 18000000 (= $6000^2/2$, given that there are about 6000 genes in the yeast genome). Thus, even a relatively low FP rate results in a relatively high absolute number of false positives. The lower part of Table 1, showing absolute numbers of false negatives and positives, indicates this. The error rate ε = FP/(TP + FP) represents the fraction of FP gene pairs among all positively predicted gene pairs (with TP being the number of true positives). For each data source the error rate is almost 100% (see lower part of Table 1). For an acceptable prediction the error rate should be at least lower than 50%. Again, we should mention here that the FPs are not necessarily 'false' in a biological sense. Thus, the error rates would be lower if all interactions were known.

Thus, an optimal combination of all these four data sources should not only aim to minimize the overall FP and FN rate, but also the error rate ε.

Combination of datasets to predict protein complexes

The application of our combination strategy to the six data sources is shown in Figure 5. With six data sources, there are $2^6 - 1$ possible subsets. Each of the subsets is represented as an open circle on the graph, with the abscissa representing the total error rate and

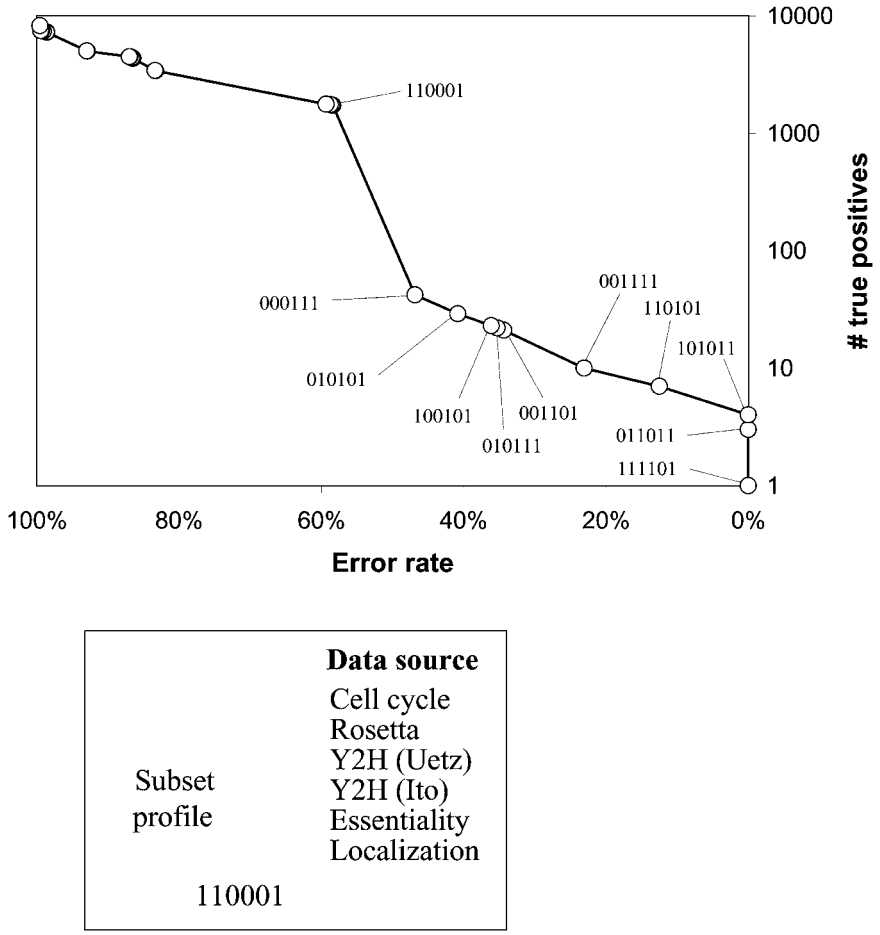

Figure 5. Here we show how the error rate and coverage change as we include more and more subsets for the prediction of membership in protein complexes. The subsets arise from the combination of the 6 data sources shown in Table 1. The abscissa shows the total error rate in decreasing order, whereas the ordinate shows the number of true positives, a measure of the coverage. We start in the lower right of the graph with the subset '111101' (the subset profile is explained in the legend at the bottom). There is only one protein pair in this subset, which is also present in the MIPS complexes catalog. Thus, we record 1 true positive and an error rate of 0%. Next we include the subsets '011011' and '101011', which increase the number of true positives to 4. The next subset with the lowest error rate among the remaining ones is '110101', which includes 3 true positives and 1 false positive. Thus, at this point, the total coverage would be 7 true positives, while the total error rate has increased to $1/(1+7) = 12.5\%$. The process of including successive subsets with the lowest error rate can be continued such that the coverage increases at the cost of a minimally increasing total error. The total error rate can be computed as $\varepsilon = \Sigma FP/(\Sigma FP + \Sigma TP)$. The first 11 of the 63 ($= 2^6 - 1$) subsets with the lowest error rates are labeled in the figure. Note that subsets with higher degrees of intersection generally have lower error rates, whereas subsets with lower degrees of intersection contribute more to coverage (although this relationship is not strict, given that the different data sources are of varying quality). In fact, several of the subsets with high degrees of intersection (>4), are empty. The subset with the highest degree of intersection '111111' is not shown because it did not contain any data. The subset '110001' – which is the intersection of the two expression data sets and the localization data – causes the coverage to increase strongly, but pushes the total error rate above 50%. In the extreme, when all subsets are included, both the coverage and the error rate are near 100%.

the ordinate showing the number of TPs, a measure for coverage. (If TP = 8250, then the coverage is 100% because all protein pairs from the MIPS complex catalog would have been detected.) The subsets are ordered with decreasing error rate from left to right. The successive inclusion of subsets would start in the lower right of the graph. Then more and more subsets would be added, moving along the graph in the upper left direction. The points associated with particular subsets show the total error rate and the total coverage if all subsets up to the current one were combined (i.e., accepting all the protein pairs in them as positives).

The total error rate can be computed as $\varepsilon = \Sigma FP/(\Sigma FP + \Sigma TP)$ where ΣFP and ΣTP are the sums of the numbers of all false and true positives in the subsets included.

Note that, in general, there is a tendency for the subsets with a higher degree of intersection to exhibit lower individual error rates, whereas the subsets with lower degrees of intersection often contribute more to the coverage.

Summary of results

In summary, Figure 5 clearly shows that one can find combinations of subsets of all the data sources that have much lower error rates than the data sources individually. Recall that these individual error rates were close to 100% (see Table 1).

In general, Figure 5 shows that, as expected, the error rates of subsets decrease the more agreement there is between the individual datasets. However, this relationship is not strict in the sense that some individual datasets carry more weight than others do. For instance, the subset '001101' has a lower error rate than '010111', although there are fewer individual datasets that predict an interaction.

The computation of error rates thus gives a numerical measure to weight the confidence in certain protein–protein interactions. The smaller the error rate of a subset, the larger is the weight that we can place on the interactions in the subset. Furthermore, some individual data sources seem to carry a weight of close to zero. This is the case for the essentiality data because there seems to be little difference in the error rates of subsets that contain the essentiality data and those that do not. In hindsight, the essentiality data did perhaps provide the least information with respect to protein complexes.

A trade-off between error rate and coverage has to be made. In our example, the error rates for many of the individual subsets are so high that a small coverage (with low error rates) seems advisable. For instance, if we confine ourselves to the 10 subsets with the lowest error rates (with a coverage of 42 TP), then the total error rate stays below 50%. A list of the protein pairs and error rates in these 10 subsets is available on our supplementary website.

Of particular interest is the subset '110001', in which the two expression experiments and the localization data would predict a protein pair within a complex. This subset adds only a small amount of additional error, but a large amount of additional coverage. An interesting aspect is that none of these three data sources were initially intended to detect protein–protein interactions *per se*. Given that the MIPS catalog of protein complexes is incomplete, the actual error rate of this subset should be lower. Thus, one may speculate that expression and localization experiments should provide a valuable tool for identifying protein complexes in organisms that have not been studied yet extensively.

Again, we note that the generated FPs are not necessarily FPs in the biological sense, especially if the error rates are low. There are overall 37 FPs in the subsets mentioned above. A further investigation of the results should focus on an analysis of the FPs, including further experiments.

Discussion

We have shown that the integration of different data sources can yield a combined dataset that has a substantially lower error rate than the individual datasets. The lower error rate comes at the cost of lower coverage, since those subsets of the data on which many of the independent data sources agree tend to be rare. For the example of predicting protein complexes, we have shown a procedure of how to identify these subsets of high quality.

Our procedure could be improved in many aspects. For instance, in the treatment of expression data, we arbitrarily chose particular correlation thresholds. The correlation thresholds could be optimized with respect to the final prediction. Many more datasets could of course be included in our analysis. For the special case of predicting complexes, one could also take the connectivity of the resulting interaction networks into account in addition to just looking at their overlap.

The MIPS complexes catalog reflects the currently known inventory of protein complexes in yeast, but this catalog is probably far from complete. This of course affects our analysis, in that FPs might give hints at where true protein complexes actually exist. This should be analyzed by further computational or experimental investigations.

The ideas proposed here could have two major impacts on functional genomics. First, our procedure could be used to identify new protein complexes in yeast. Second, they could be used to characterize protein complexes in newly sequenced organisms that have not been studied as extensively as yeast by tra-

ditional methods, but for which new genome-wide experiments are available.

Supplementary information

Supplementary information is available at http://bioinfo.mbb.yale.edu/integrate/interactions/.

References

1. Ermolaeva, O., Rastogi, M., Pruitt, K.D., Schuler, G.D., Bittner, M.L., Chen, Y., Simon, R., Meltzer, P., Trent, J.M., and Boguski, M.S. (1998) *Nat. Genet.* **20:** 19–23.
2. Gaasterland, T., and Bekiranov, S. (2000) *Nat. Genet.* **24:** 204–206.
3. Hegde, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Hughes, J.E., Snesrud, E., Lee, N., and Quackenbush, J. (2000) *Biotechniques*. **29:** 548–550.
4. Kim, S., Dougherty, E.R., Bittner, M.L., Chen, Y., Sivakumar, K., Meltzer, P., and Trent, J.M. (2000) *J. Biomed. Opt.* **5:** 411–424.
5. Shalon, D., Smith, S.J., and Brown, P.O. (1996) *Genome Res*. **6:** 639–645.
6. Ross-Macdonald, P., Coelho, P., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K., Sheehan, A., Symoniatis, D., Umansky, L., Heidtman, M., Nelson, F., Iwasaki, H., Hager, K., Gerstein, M., Miller, P., Roeder, G., and Snyder, M. (1999) *Nature*. **402:** 413–418.
7. Winzeler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H., Chu, A.M., Connelly, C., Davis, K., Dietrich, F., Dow, S.W., El Bakkoury, M., Foury, F., Friend, S.H., Gentalen, E., Giaever, G., Hegemann, J.H., Jones, T., Laub, M., Liao, H., Davis, R.W., and *et al.* (1999) *Science* **285:** 901–906.
8. Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R.A., Gerstein, M. and Snyder, M. (2001) *Science* **293:** 2101–2105.
9. 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.
10. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001) *Proc. Natl. Acad. Sci. USA* **98:** 4569–4574.
11. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J.M. (2000) *Nature* **403:** 623–627.
12. Ben-Dor, A., Shamir, R. and Yakhini, Z. (1999) *J. Comput. Biol.* **6:** 281–297.
13. Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., Jr., and Haussler, D. (2000) *Proc. Natl. Acad. Sci. USA* **97:** 262–267.
14. Bussemaker, H.J., Li, H. and Siggia, E.D. (2001) *Nat. Genet.* **27:** 167–171.
15. Ge, H., Liu, Z., Church, G.M., and Vidal, M. (2001) *Nat. Genet*. **29:** 482–486.
16. Gerstein, M., and Jansen, R. (2000) *Curr. Opin. Struct. Biol.* **10:** 574–584.
17. Greenbaum, D., Jansen, R., and Gerstein, M. (2002) *Bioinformatics* **18:** 1–12.
18. Greenbaum, D., Luscombe, N.M., Jansen, R., Qian, J., and Gerstein, M. (2001) *Genome Res*. **11:** 1463–1468.
19. Gygi, S.P., Rochon, Y., Franza, B.R., and Aebersold, R. (1999) *Mol. Cell. Biol.* **19:** 1720–1730.
20. Heyer, L.J., Kruglyak, S., and Yooseph, S. (1999) *Genome Res*. **9:** 1106–1115.
21. Jansen, R., and Gerstein, M. (2000) *Nucleic Acids Res.* **28:** 1481–1488.
22. Jansen, R., Greenbaum, D., and Gerstein, M. (2002) *Genome Res.* **12:** 37–46.
23. Qian, J., Dolled-Filhart, M., J., L., Yu, H., and Gerstein, M. (2001a) *J. Mol. Biol*. **314:** 1053–1066.
24. Qian, J., Stenger, B., Wilson, C.A., Lin, J., Jansen, R., Teichmann, S.A., Park, J., Krebs, W.G., Yu, H., Alexandrov, V., Echols, N., and Gerstein, M. (2001b) *Nucleic Acids Res*. **29:** 1750–1764.
25. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., and Golub, T.R. (1999) *Proc. Natl. Acad. Sci. USA* **96:** 2907–2912.
26. Toronen, P., Kolehmainen, M., Wong, G., and Castren, E. (1999) *FEBS Lett.* **451:** 142–146.
27. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D. (1999) *Nature* **402:** 83–86.
28. Drawid, A., and Gerstein, M. (2000) *J. Mol. Biol.* **301:** 1059–1075.
29. Drawid, A., Jansen, R., and Gerstein, M. (2000) *Trends Genet*. **16:** 426–430.
30. Cohen, B., Mitra, R., Hughes, J. and Church, G. (2000) *Nat. Genet.* **26:** 183–186.
31. Mewes, H.W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schuller, C., Stocker, S., and Weil, B. (2000) *Nucleic Acids Res*. **28:** 37–40.
32. Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., and Davis, R.W. (1998) *Mol. Cell.* **2:** 65–73.
33. Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H.Y., He, Y.D.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., and Friend, S.H. (2000) *Cell* **102:** 109–126.
34. Bairoch, A., and Apweiler, R. (2000) *Nucleic Acids Res.* **28:** 45–48.
35. Hodges, P.E., McKee, A.H., Davis, B.P., Payne, W.E., and Garrels, J.I. (1999) *Nucleic Acids Res.* **27:** 69–73.
36. Gerstein, M., Lan, N., and Jansen, R. (2002) *Science* **295:** 284–287.