56 Aravind, L. and Koonin, E.V. (2001) Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. *Genome Res.* 11, 1365–1374
57 Reeves, A.R. *et al.* (1998) Physical-genetic map of the erythromycin-producing organism *Saccharopolyspora erythraea. Microbiology* 144, 2151–2159
58 Redenbach, M. *et al.* (2000) Chromosome topology and genome size of selected actinomycetes species. *Antonie Van Leeuwenhoek* 78, 227–235
59 Volff, J.N. and Altenbuchner, J. (2000) A new beginning with new ends: linearisation of circular chromosomes during bacterial evolution. *FEMS Microbiol. Lett.* 186, 143–150
60 Redenbach, M. *et al.* (1996) A set of ordered cosmids and a detailed genetic and physical map for the 8 Mb *Streptomyces coelicolor* A3(2) chromosome. *Mol. Microbiol.* 21, 77–96
61 Friend, E.J. and Hopwood, D.A. (1971) The linkage map of *Streptomyces rimosus. J. Gen. Microbiol.* 68, 187–197
62 Louarn, J. *et al.* (1994) Hyperrecombination in the terminus region of the *Escherichia coli* chromosome: possible relation to nucleoid organization. *J. Bacteriol.* 176, 7524–7531
63 Bao, K. and Cohen, S.N. (2001) Terminal proteins essential for the replication of linear plasmids and chromosomes in *Streptomyces. Genes Dev.* 15, 1518–1527
64 Yang, C-C. *et al.* (2002) The terminal proteins of linear *Streptomyces* chromosomes and plasmids: A novel class of replication priming proteins. *Mol. Microbiol.* 43, 297–305
65 Lin, N.T. *et al.* (2001) Plasmids carrying cloned fragments of RF DNA from the filamentous phage ϕLf can be integrated into the host chromosome via site-specific integration and homologous recombination. *Mol. Genet. Genomics* 266, 425–435
66 Clerget, M. (1991) Site-specific recombination promoted by a short DNA segment of plasmid R1 and by a homologous segment in the terminus region of the *Escherichia coli* chromosome. *New Biol.* 3, 780–788
67 Barre, F.X. *et al.* (2001) Circles: the replication-recombination-chromosome segregation connection. *Proc. Natl. Acad. Sci. U. S. A.* 98, 8189–8195
68 Fraser, C.M. *et al.* (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi. Nature* 390, 580–586
69 Picardeau, M. *et al.* (1999) Physical mapping of an origin of bidirectional replication at the centre of the *Borrelia burgdorferi* linear chromosome. *Mol. Microbiol.* 32, 437–445
70 Goodner, B. *et al.* (2001) Genome sequence of the plant pathogen and biotechnology agent *Agrobacterium tumefaciens* C58. *Science* 294, 2323–2328
71 Wood, D.W. *et al.* (2001) The genome of the natural genetic engineer *Agrobacterium tumefaciens* C58. *Science* 294, 2317–2323
72 Yamasaki, M. *et al.* (2000) A complex insertion sequence cluster at a point of interaction between the linear plasmid SCP1 and the linear chromosome of *Streptomyces coelicolor* A3(2). *J. Bacteriol.* 182, 3104–3110
73 Yamasaki, M. *et al.* (2001) Integrated structures of the linear plasmid SCP1 in two bidirectional donor strains of *Streptomyces coelicolor* A3(2). *Mol. Gen. Genet.* 264, 634–642
74 Casjens, S. (1999) Evolution of the linear DNA replicons of the *Borrelia* spirochetes. *Curr. Opin. Microbiol.* 2, 529–534
75 Schardl, C.L. *et al.* (1984) Linearization of maize mitochondrial chromosomes by recombination with linear episomes. *Nature* 310, 292–296

Bridging structural biology and genomics: assessing protein interaction data with known complexes

Aled M. Edwards, Bart Kus, Ronald Jansen, Dov Greenbaum, Jack Greenblatt and Mark Gerstein

Currently, there is a major effort to map protein–protein interactions on a genome-wide scale. The utility of the resulting interaction networks will depend on the reliability of the experimental methods and the coverage of the approaches. Known macromolecular complexes provide a defined and objective set of protein interactions with which to compare biochemical and genetic data for validation. Here, we show that a significant fraction of the protein–protein interactions in genome-wide datasets, as well as many of the individual interactions reported in the literature, are inconsistent with the known 3D structures of three recent complexes (RNA polymerase II, Arp2/3 and the proteasome). Furthermore, comparison among genome-wide datasets, and between them and a larger (but less well resolved) group of 174 complexes, also shows marked inconsistencies. Finally, individual interaction datasets, being inherently noisy, are best used when integrated together, and we show how simple Bayesian approaches can combine them, significantly decreasing error rate.

Published online: 15 August 2002

Protein–protein interactions are one of the most important regulatory mechanisms in cells; they underlie intercellular communication, signal transduction and the regulation of gene expression. Indeed, most cellular processes are coordinated by specific protein interactions. The completion of the Human Genome Project led to the realization that the genome is composed of fewer protein-coding genes than had been previously believed [1,2]. It is now thought that the complexity underlying the biology of higher organisms could arise not from the number of their proteins, but rather from the combinatorial interactions among them [3].

Many large-scale protein-interaction datasets have been published, each attempting to completely characterize the 'interactome' – the set of all protein interactions in a cell [4–7]. (See supplementary

0168-9525/02/$ – see front matter PII: S0168-9525(02)02763-4

material website for a summary of the different techniques, http://genecensus.org/integrate/interactions.) These datasets, although extensive, miss many of the interactions, and report spurious interactions as well (e.g. Ref. [7]). To use these datasets to build robust and statistically significant protein-interaction networks, it will be essential to quantify the intrinsic error rates in these experiments. This will be challenging, because it is difficult to define a basis protein-interaction dataset with which to validate the experimentally identified interactions. Yet it is critical, because in yeast, for example, there are a possible 18 million protein interactions (given 6000 proteins), of which only a small fraction is relevant biologically. With so many potential interactions, even a very small false-positive rate can create a large amount of spurious interactions that swamp out the real ones.

Recently, some papers have attempted to evaluate the quality of these interaction datasets, by either integrating additional annotation (e.g. looking for similar expression patterns between two interacting proteins) or through comparison with reference interaction sets [8–11]. The conclusions from these studies were that the interaction datasets contain false positives and are missing many true protein interactions.

Our paper provides two complementary analyses of the reliability of protein-interaction datasets. First, we use information provided through X-ray crystallography of several macromolecular complexes to provide validated interactions that serve as a basis set for comparison with the interaction databases. Second, on a somewhat larger scale, we compared the interactions in the known complexes in the curated MIPS database, with those that had been determined using genome-wide approaches. (See supplementary material for more extensive discussion of MIPS complexes.)

Assessing interactions with the 3D structures of three complexes

Interactions discovered using structural biology of large complexes provide an excellent test set to assess the quality and coverage of protein-interaction datasets. Over the past two years, the structures of several large, conserved protein complexes have been determined, and from these structures, a list of verified protein–protein interactions can be extracted. The structures that we selected (RNA polymerase II, the proteasome and the Arp2/3 complex) were solved independently of the experimental interaction datasets, and the 3D structures were determined after most of the biochemical and genetic studies were published. Thus, the structures provided objective and verified interactions to assess the reliability of biochemical and genetic protein-interaction studies.

The 3D structures of RNA polymerase II, the proteasome and the Arp2/3 complex, which have 10, 14 and 7 subunits, respectively, were examined to identify stable protein–protein interaction interfaces, defined as more than 800 Å^2 of contacting surface area. This 800 Å^2 threshold was chosen based on the lower limits of the interaction interfaces of binary protein–protein complexes in the structural database (PDB) [12]; the 'average' protein-interaction interface is 1600 Å^2 [13].

For RNA polymerase II, of the 45 possible interactions that could exist among its ten subunits [(10 × 9)/2), 13 clear protein–protein interactions were found in the crystal structure [14,15]. Of the 21 possible interactions that could occur among the Arp2/3 subunits, six exist in the crystal structure [16]. Of the 91 possible interactions that could occur among the subunits in the proteasome, 14 were observed in the crystal structure [17]. The ribosome was eliminated as a candidate for our analysis because it is mostly held together by protein–RNA interactions, confounding attempts to verify all the protein–protein interactions.

An assumption in our analysis is that the protein interactions defined on the basis of the structural biology of stable macromolecular complexes provide an objective set of interactions. This assumption is supported by two observations. First, many macromolecular assemblies – particularly, RNA polymerase II, Arp2/3 and the proteasome – are very stable and can be purified to homogeneity without loss of subunits. In fact, the RNA polymerase II complex is even stable in high concentration of chaotropic agents, such as urea [18]. Therefore, the subunit–subunit interactions found in the crystal structure are unlikely to have formed randomly. Second, many crystallized proteins are enzymatically active. For example, it is known that the crystallized form of the active RNA polymerase II elongation complex has the same repertoire of specific protein interactions as does the native form [19].

False positives and negatives

For each of the three protein complexes, lists of biochemically and genetically defined interactions were extracted from genome datasets, as well as from the biochemical and genetic literature. The structural, biochemical and genetic lists were compared to assess overlap, consistency and the rates of false positives and false negatives. For clarification: a false positive (FP) is defined as an interaction that was documented in one of the datasets but did not exist in the crystal structure; a true positive (TP) is an interaction that was documented in one of the datasets and did, in fact, exist in the crystal structure; a false negative (FN) is an interaction that had been tested experimentally and failed to score as an interaction, but is known to exist in the crystal structure; and a true negative (TN) is an interaction that had been tested experimentally and failed to score as an interaction, and is known not to occur in the crystal structure.

Aled M. Edwards
Bart Kus
Jack Greenblatt
Banting and Best Dept of Medical Research, University of Toronto, C.H. Best Institute, 112 College St, Toronto, Ontario, Canada M5G 1L6.

Ronald Jansen
Dov Greenbaum
Mark Gerstein*
Dept of Molecular Biophysics and Biochemistry, Yale University, 266 Whitney Avenue, New Haven, CT 06520, USA.
*e-mail: mark.gerstein@yale.edu

Table 1. Agreement between the various interaction datasets in the literature and the crystal structure of RNA polymerase II[a]

	Dataset							Method			
	Miyao *et al.* [25] far-western (Rpb5 as probe)	Ishiguro *et al.* [26] cross linking	Ishiguro *et al.* [26] far-western (Rpb6/8/10/11/12 probe)	Acker *et al.* [20] GST-pulldown (X and GST-Y)	Acker *et al.* [20] GST-pulldown (GST-X and Y	Kimura *et al.* [21] pull down	Yasui *et al.* [24] far western	Combined	Pulldown	Far western	Cross linking
TP	2	10	6	6	7	2	1	**7**	15	9	10
FN	3	0	2	4	3	2	2	**6**	9	7	0
TN	6	3	17	14	16	2	2	**29**	32	25	3
FP	4	7	5	11	9	3	1	**3**	23	10	7
Coverage	15	20	30	35	35	9	6	**45**	35	20	36
False-negative rate FN/(FN + TP) [%]	60	0	25	40	30	50	67	**46**	38	44	0
False-positive rate FP/(TP + FP) [%]	67	41	45	65	56	60	50	**30**	61	53	41

[a] The left side gives statistics for the datasets individually, whereas we group the datasets by method on the right side. We show the count of true positive (TP), false negative (FN), true negative (TN) and false positive (FP), the coverage (that is, the number of unique subunit-subunit pairs the data covers), the false-negative rate (FN/(FN + TP)) as well as the false-positive rate defined as FP/(TP + FP) for each dataset. The table does not contain the Ulmasov and Larkin data referred to in the text since they both contain only one TP, which is already covered by the Yasui data. Note that for three FPs in the cross linking data, the subunits are actually in contact in the structure, but with less than our 800 Å^2. In addition, we show the corresponding statistics for a dataset that combines the individual datasets using a Bayesian procedure [40]. The integration of the various data sources both increases the coverage (45 subunit–subunit pairs for the combined data versus 6–35 pairs for the individual data) and lowers the false-positive rate (30% for the combined data versus 41–67% for the individual data).

If we think of each data source k as evidence e_k for or against a protein–protein interaction, we can compute a combined belief in an interaction I from:

$$O(I \mid e_1, e_2 \ldots e_N) = L(e_1, e_2 \ldots e_k \mid I) O(I)$$

where

$$O(I) = \frac{P(I)}{P(\neg I)}$$ are the prior odds of interaction I,

$$O(I \mid e_1, e_2 \ldots e_N) = \frac{P(I \mid e_1, e_2 \ldots e_N)}{P(\neg I \mid e_1, e_2 \ldots e_N)}$$ are the posterior odds of interaction I, and

$$L(e_1, e_2 \ldots e_k \mid I) = \frac{P(e_1, e_2 \ldots e_k \mid I)}{P(e_1, e_2 \ldots e_k \mid \neg I)}$$ is the likelihood ratio.

The prior odds are $O(I) = 13/(45 - 13)$ because there are 13 interactions among the 45 possible subunit-subunit pairs in the crystal structure. Assuming conditional independence of the evidence (meaning that each experiment is only dependent on whether there is an interaction and thereafter they are independent of the outcome of the other experiments), the likelihood ratio can be written as:

$$L(e_1, e_2 \ldots e_k \mid I) = \prod_{k=1}^{N} L(e_k \mid I)$$

The likelihood ratio $L(e_k|I)$ for experiment k can be computed from the experimental data.

For each method, we define the false-positive rate as the number of false-positive interactions reported per total reported interactions [FP/(FP + TP)], and the false-negative rate as the fraction of known interactions that are not identified [FN/(FN + TP)].

Comparing structural and biochemical interactions in the literature

Before the elucidation of the 3D structures for the three above-mentioned complexes, many of the subunits had been subjected to *in vitro* and *in vivo* binding experiments, chemical cross-linking and far-western analysis (see Table 1 for an overview of data relating to RNA polymerase II). The aim of these experiments, in the absence of crystallographic information, had been to deduce the protein-interaction map for a given complex. With 3D structures now in hand, we compared the subunit–subunit interactions defined by these experiments with the interactions revealed in the crystal structures.

In one series of experiments with RNA polymerase II subunits, every recombinant subunit was cloned into baculoviruses as either a glutathione-*S*-transferase (GST)-tagged or an untagged protein. Every combination of individual GST-tagged subunits and non-tagged subunits were co-expressed and precipitated, generating a comprehensive set of inter-subunit pairwise interactions [20–24]. This is a common and accepted strategy ('pull-down') to define binary protein–protein interactions. Remarkably, 61% of the interactions in these experiments were false positives. Some of the false-positive interactions were even 'validated' by several different biochemical approaches [20,24] and, in a few cases, spurious contact sites were also mapped in detail [24,25]. The false-negative rate was 38%.

Biochemical interactions were also extracted from a series of far-western experiments in which the RNA polymerase II subunits were resolved by denaturing gel electrophoresis, transferred to a solid support, renatured and over-laid with various purified subunits. Of the interactions in the RNA polymerase II crystal structure, 44% were not observed with this method, whereas 53% of the reported interactions were false positives [24,25].

Chemical cross-linking [26] proved a more effective method for identifying biochemical interactions between RNA polymerase II subunits. The cross-linking data contained no false negatives and ten true positives; there were seven false positives (41%), although for three of them we did find subunit–subunit contacts in the crystal structure with a contact area below our 800 Å^2 cutoff.

Cross-linking approaches were also used to study the Arp2/3 and proteasome complexes. For the Arp2/3 complex, 86% of the interactions found by cross-linking were present in the crystal structure [27,28], whereas the false-negative rate was 50%. For the proteasome, only five of the 14 interactions (36%) reported in the cross-linking study were not present in the crystal structure [29], whereas the false-negative rate was 79%. The relative success of cross-linking approaches (25–41% false-positive rate) suggests that this method could be preferable to others, particularly for relatively stable macromolecular complexes. Unfortunately, chemical cross-linking can be challenging to perform on a genomic scale, and the efficacy of cross-linking studies is dependent on the particular protein complexes and the juxtaposition of the appropriate amino acid residues that react with the cross-linking reagent.

Comparing structural and genetic interactions

The yeast two-hybrid (Y2H) method identifies interactions between two recombinant proteins expressed in yeast cells. Current estimates suggest that the false-positive rate could range between 47 and 91% in the genome-wide Y2H screens [30]. Our analysis of the structural information for RNA pol II, Arp2/3 and the proteasome was unable to contribute to the estimation of the false-positive rate in Y2H screens for two reasons. First, the structural information was derived from yeast complexes, and therefore we could not rule out the possibility that the interaction between two subunits could be bridged by other, endogenous yeast components. Second, the Y2H method could also detect transient interactions not revealed in the structure, but that might exist during complex assembly.

We were able to use the structural information to estimate the false-negative rates in Y2H screens. Our analysis was restricted to the Arp2/3 and proteasome complexes because the RNA polymerase II subunits were omitted in the Y2H, as they are known to score positively in the absence of a protein partner. The individual proteasomal subunits were screened against the rest of the yeast proteasome by Cagney and colleagues [31], and 12 interactions between the subunits were revealed. Of these 12 interactions, eight were also found in the crystal structure and have contact surfaces that exceeded 800 Å^2. Because the crystal structure revealed 14 interactions between subunits, this directed Y2H approach missed six out of 14 interactions (43%). When interactions among Arp2/3 subunits were tested directly using Y2H screens, 71% of subunit–subunit interactions present in the Arp2/3 crystal structure were not detected [32].

Comparison with genome-wide two-hybrid

The proteasome and Arp2/3 subunits were also analyzed as part of several genome-wide Y2H screens. In the first, carried out by Uetz and colleagues [6,33], five interactions involving proteasome subunits and other proteins were uncovered, but not one interaction between two known proteasomal subunits was found. This dataset also did not contain any interactions between Arp2/3 subunits. A subsequent two-hybrid screen carried out by Ito *et al.* [7] identified 30 interactions between proteasome subunits and non-proteasome proteins, but only one intra-complex interaction. This interaction was, in fact, present in the crystal structure. To reduce the extent of false positives in their dataset, Ito *et al.* created a 'core dataset' that was filtered by taking only data validated by more than three interaction sequence tags. This core dataset eliminated protein–protein interaction data that did not appear at least three times in their screen. This filtering process eliminated the single correct proteasome interaction from the dataset. The Ito *et al.* screen correctly identified one interaction between Arp2/3 subunits, and this interaction was maintained in the 'core dataset'.

In summary, we estimate that the false-negative rate in directed two-hybrid screens ranges from ~43 to 71%, and the rate was even higher in genome-scale applications of the Y2H method.

Large-scale pull-down experiments

Biochemical interactions can also be discovered using 'pull-down' experiments *in vivo*. In these experiments, a protein tag is engineered onto a specific subunit and the gene for the tagged subunit added back into the cell. The tagged protein is then purified from cell lysates using affinity chromatography, and the co-purifying proteins identified. For multi-protein complexes, this method does not provide information about binary protein interactions, but rather describes the collection of proteins that are stably associated with the tagged protein. Gavin *et al.* and Ho *et al.* reported the results from large-scale pull-down experiments in two recent publications [4,5]. For the three complexes we studied, the *in vivo* 'pull down' method was quite successful in identifying subunits that interact within the complex. In the case of RNA polymerase, two of the ten polymerase subunits were tagged and the co-purifying proteins identified. Half of the subunits known to interact directly with these subunits were detected, for a false-negative rate of 50%. When a subset of the Arp2/3 and the proteasome subunits were tagged, there were no false negatives; all interactions present in the crystal structure whose contacting surface are >800 Å^2 were found in these datasets. The *in vivo* 'pull-down' experiments seemed to have fewer false negatives than other methods, although the rate for RNA polymerase II was 50%.

In conclusion, we have used a small test set of structure-based interactions to assess the quality of several protein-interaction datasets, and have quantified significant sources of error. Can the results of our structural analysis of complexes be generalized? We cannot be sure. With regard to the analysis of the 3D structures of large complexes, it is possible that inter-subunit interactions distinct from those seen in the crystal might occur, but these complexes were selected to be quite stable. It is also likely that the discrepancies between the information in the crystal structure and the biochemical literature results from protocols used in high-throughput studies, which are acknowledged to generate unvalidated data in the interest of speed and efficiency. Even if these 3D complexes represent special cases, our analysis deserves some attention, because it is probably the first objective assessment of the protein-interaction datasets.

Broadening the comparisons to 174 complexes

Complexes in the MIPS database

Our structure-based analysis of the 31 proteins within the three complexes points to the potential for significant errors in large-scale protein-interaction datasets. To broaden our study, we looked at a larger set of complexes listed in the MIPS database [34]. This comprises complexes that are manually annotated from the yeast literature. From the MIPS complexes catalogue, we identified 174 protein complexes that contain 2–81 proteins each. The structures of the complexes in the MIPS complexes catalog are generally not known; consequently, it is impossible to perform 3D-structure-based analyses. However, we can still use them as a quality control for the genome-wide datasets, and their use broadens our analysis by orders of magnitude.

A theoretical minimum and maximum number of interactions within these complexes can be calculated (Fig. 1). For example, the maximum number of interactions would arise if each protein in these complexes interacted with every other in the complex (an unreasonable assumption for the very large complexes). Among the 174 protein complexes that we selected, there would be a theoretical maximum of 8250 interactions among all the proteins. The theoretical minimum number of interactions among the proteins in all the MIPS complexes would be 834, because the lowest possible number of interactions in a complex is the number of proteins itself minus one (when the proteins are arrayed like beads on a string, thus all proteins in the complex are connected to at least one and at most two other proteins). Our estimations also presume that the complexes are maintained by protein interactions and not by other molecules, such as RNA as in the case of the ribosome. Of course, the number of true interactions among these proteins somewhere in the middle of this range, probably closer to the minimum.

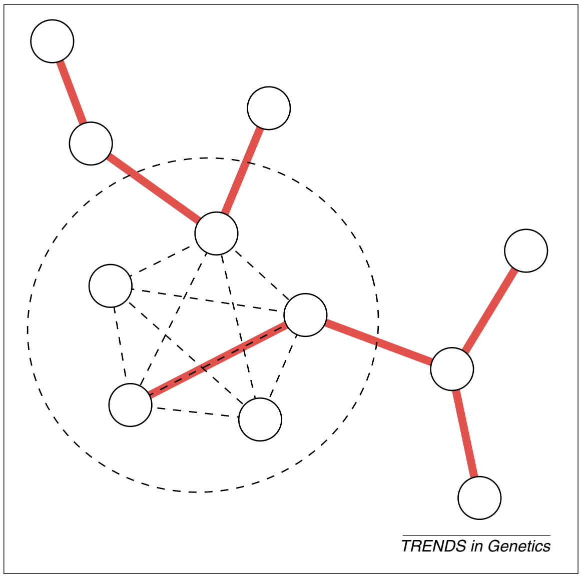

Fig. 1. We systematically compared the interaction datasets with the complexes. In general, we do not know the structure of the complexes in the MIPS catalog, so we simply counted which interactions from the genome-wide datasets are between proteins that are in the same MIPS complex. The proteins in the dashed circle all belong to the same protein complex, and the dashed lines between the proteins indicate all theoretically possible protein–protein interactions within this complex. The red lines indicate the interactions of a genome-wide dataset.

Overlap between MIPS complexes and genome-wide Y2H interaction datasets

In Table 2, we present a comprehensive overview of the overlap between the MIPS complexes and the other datasets. Each dataset is defined on a different subset of genes, which makes comparison between the sets difficult. Moreover, when we look at the intersections between the datasets we find that they are fairly small (Fig. 2). This arises for two reasons: (1) the different subsets of genes, and (2) the different interactions for the same sets of proteins in each dataset. Consequently, in the analysis we consider both how many interactions and how many genes are shared between the datasets and the complexes. This allows us to separate the two effects.

We can use the Uetz data [6] as an example. Within the 174 MIPS complexes, there are 871 distinct proteins, and the Uetz dataset overlaps with a subset of 179 proteins. Of these, there are 78 proteins involved in 50 known intra-complex

Table 2. The size of the genome-wide datasets and their overlap with the MIPS complexes catalog[a]

	Yeast two-hybrid			*in vivo* pull-down	
Dataset	**Uetz**	**Ito**	**Ito core**	**Ho**	**Gavin**
Ref.	**[6]**	**[7]**	**[7]**	**[5]**	**[4]**
No. proteins	1044	3278	786	1578	1361
No. proteins overlapping with MIPS	179	453	142	331	472
No. proteins linked to partner is in same MIPS complex	78	121	73	214	436
No. proteins not linked to MIPS partner	101	332	69	117	36
No. interactions	981	4393	754	25333	31304
No. interactions between proteins, both in MIPS catalog	64	177	51	2042	4718
No. interactions between proteins, same MIPS complex (TP)	50	78	46	464	1743
No. interactions between proteins, different MIPS complexes (FP or new)	14	99	5	1578	2975
No. MIPS protein pairs missed (FN)	52–283	250–1682	38–131	464	300
FN/(TP + FN) (false-negative rate)	51%–85%	76%–96%	45%–74%	50%	15%

[a]The overlap is first shown in terms of the number of proteins, then in terms of the number of interactions (or protein pairs). Note that we consider the maximum number of possible protein pairs between all subunits for the complexes in the MIPS catalog and the Ho and Gavin datasets. The number of interactions that overlap with the MIPS catalog is very small for each of the individual genome-wide datasets. This is partly due to the fact that the genome-wide datasets cover different proteins than those that the MIPS catalog contains. When we normalize for the protein set shared between the MIPS catalog and the genome-wide dataset in each case, the relative overlap becomes more substantial. For instance, of the 31304 interactions (protein pairs) in the Ho dataset, 4718 are between proteins that the MIPS catalog contains. Of these 4718 interactions, 1743 correspond to protein pairs that are within the same MIPS complex (true positives, TP), whereas 2975 are between proteins in different MIPS complexes (these either represent newly discovered interactions or false positives, FP). The Gavin dataset missed 300 protein pairs within the same MIPS complex (false negatives, FN). These statistics allow us to measure the sensitivity TP/(TP + FN) of each genome-wide dataset, that is, the fraction of protein pairs within the same MIPS complex that each genome-wide dataset recovers (normalized for the set of proteins shared with the MIPS catalog). The Gavin dataset recovers the largest amount of interactions (86%), followed by the Ho dataset (50%) and the Y2H datasets (Uetz 15%–49%, Ito 4%–24%). The sensitivity of the Ito core dataset (26%–55%) is higher than for the total Ito dataset.

interactions (i.e. where the observed Uetz interactions correspond to a protein-pair in a MIPS complex, generating a true positive). However, these 78 proteins are also involved in 14 interactions that connect different MIPS complexes, thus potentially representing newly discovered or false-positive interactions. The remaining 101 of the 179 proteins are only involved in interactions not found in any MIPS complex, again representing new interactions or false positives.

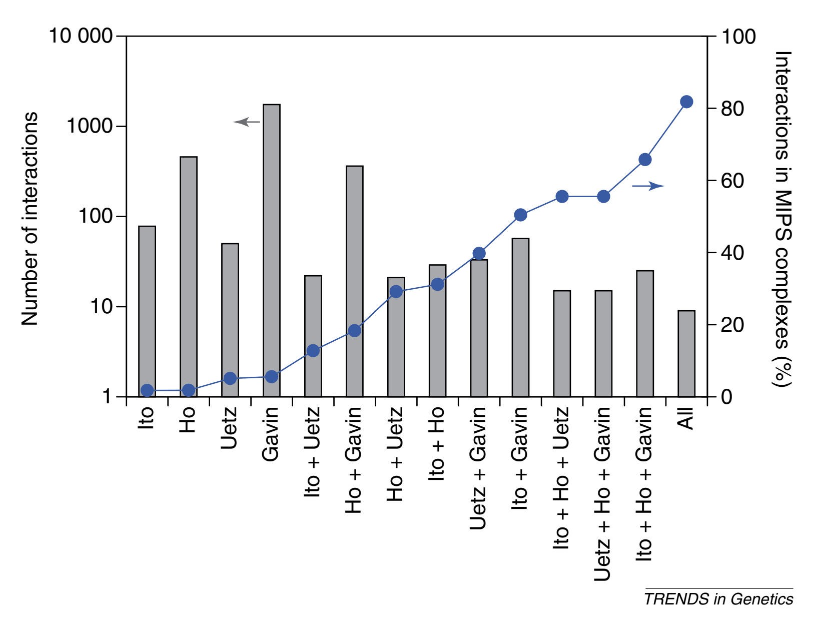

Fig. 2. The size of the different genome-wide datasets and their possible intersections and their consistency with the MIPS complexes catalog. The bars (relating to the left *y*-axis) indicate the number of interactions in each individual dataset and each possible intersection of the datasets (for instance, 'Ito + Uetz' contains only interactions that are both within Ito and Uetz). The blue line (relating to the right *y*-axis) shows what fraction of these interactions overlap with protein pairs within the same MIPS complex. We show this for both the individual datasets (left), pairwise intersections of datasets (middle) and higher order intersections (three or more datasets). As the degree of intersection among the datasets increases, the fraction of interactions within the same MIPS complex increases. The different datasets are complementary and cover more interactions than each dataset individually.

How many of the MIPS complex interactions do the Uetz data miss? This, of course, depends on how many interactions we count in the MIPS complexes. We can estimate that the 78 proteins that are both in the Uetz data and the MIPS complexes must be connected by between 102 and 333 existing interactions (using minimum and maximum numbers of interactions for complexes; see above and the website for a detailed derivation of these values). Thus, the Uetz data missed between 52 (= 102 – 50) and 283 (= 333 – 50) existing interactions, implying a false-negative rate between 51% (= 52/102) and 85% (= 283/333).

One can analyze the other two-hybrid datasets in a similar fashion. The Ito data has a false-negative rate of 76–96%; the corresponding values for the Ito 'core' data are 45–74%.

Overlap between MIPS complexes and genome-wide 'pull-down' interaction datasets

In contrast to the Y2H experiments where proteins are studied outside their normal cellular milieu, the comprehensive *in vivo* pull-down experiments capture the protein in its natural state. Consequently, one might expect the pull-down experiments to be somewhat more sensitive and miss fewer existing interactions.

The right panel in Table 2 shows the comparison between the two pull-down experiments (Gavin *et al.* [4] and Ho *et al.* [5]) with the MIPS complexes catalog, using a similar approach as for the Y2H. Overall, the false-negative rates of the Ho and Gavin datasets, are 50% and 15%, thus confirming our expectation. However, these datasets also contain 1578 (77% of 2042) and 2975 (63% of 4718) interactions between different MIPS complexes, respectively, thus representing either newly discovered or false-positive interactions.

Thus, it appears that the pull-down methods are very sensitive; they discover real interactions at a high rate, but the datasets are probably populated by false positives.

The relatively low false-negative rates in the *in vivo* pull-down experiments are remarkable, even with the caveat that the experimenters had the benefit of prior knowledge of the contents of the MIPS database. Why is this approach apparently superior to Y2H methods? First, MIPS is mostly populated by stable complexes, which are amenable to pull-down methods because they can remain intact for hours during affinity chromatography. Second, the pull-down experiments probe the complexes in an active state; presumably the proteins retain their structural integrity during the purification process. Why does the TAP-tagging approach used in the Gavin *et al.* paper appear to be superior to that used in the Ho *et al.* paper in reducing the false-negative rate? We speculate that it is because in the TAP-tagging approach, the engineered genes are introduced into the yeast genome using homologous recombination, and are expressed at normal levels from their natural promoter. The TAP-tagged proteins can therefore incorporate into the normal cellular environment. These features all contribute to the relative success of this method. By contrast, the method used by Ho and colleagues, which involved overexpression of the engineered protein, could have increased the level of proteins that were not associated specifically, and made the interpretation of the results more difficult.

Data integration

We have seen that both genome-wide and smaller conventional interaction datasets can be noisy and inaccurate. If the noise is not systematically biased, it should be possible to improve the accuracy of the datasets by integrating the information within them. The recent trend is to combine a variety of experimental information – including transcription co-expression, co-localization, essentiality information (i.e. whether a gene is essential for the viability of the cell) and functional annotations – to help predict and validate protein–protein interactions [8–10,35–39]. In general, data integration, combining multiple independent sources, should increase the degree to which the known complex interactions are found (i.e. increasing the coverage) while decreasing the number of errors in the data [10,35,36].

We attempted to quantify the improvements in merging datasets by using the structural information for RNA polymerase II. As described in Table 1, the false-positive rates in each of the RNA polymerase II datasets ranged from 41 to 67%. We combined the various interaction datasets for RNA polymerase II using a simple Bayesian network (Table 1 and supplementary material), where we essentially multiplied the probability of interaction across the various datasets. The combined data has a false-positive rate of 30%, which is lower than that for any of the individual datasets. Moreover, the combined data gives a complete coverage of the 45 possible subunit–subunit pairs in the RNA polymerase II structure, whereas the individual interaction studies only cover 6–35 of them. In other words, in contrast to the merged dataset, none of the interaction studies in the literature provide information on all of these possible interactions. (In particular, the sum of true positives and true negatives is 36 in the combined data versus only 23 or less for the individual datasets.)

Similarly, we can show the benefits of data integration for the genome-wide datasets (Fig. 2). First, for each genome-wide dataset (e.g. the Gavin TAP-tag set), we identified the fraction of interactions that contain information about the MIPS complexes. Then we looked at all possible pairwise intersections among the genome-wide sets (e.g. Gavin and Ho) and again looked at how well they found the MIPS complex interactions. Finally, we looked at triplet and quadruplet intersections amongst the datasets (e.g. Gavin and Ho and Uetz). One can see that as we progress from individual datasets to pairwise intersections and then to even higher-order overlaps, the accuracy increased (as measured by the fraction of interactions that represent protein pairs within the same MIPS complex). That is, in any individual dataset, only a small fraction of the interactions involved a partner within the same MIPS complex. However, in the final, merged dataset, for instance, the majority of the interactions were consistent with the MIPS data.

The intersections of the datasets tend to be rather small; however, together the datasets are complementary and increase the amount of interactions covered. These observations are consistent with the results from the focused analysis of RNA polymerase II.

A summary and a strategy

Our analysis has two main conclusions. First, biochemical and genetic methods, when applied on a small or large scale, have, and will introduce significant numbers of spurious interactions into the datasets of binary protein interactions. Genome-wide screens for individual protein–protein interactions have yet to uncover the bulk of protein–protein interactions. The validity of the interaction datasets can be improved with the use of structural information about protein complexes. Thus, it would be difficult to consider mining the current binary interaction data to model intracellular networks.

Second, the *in vivo* pull-down methods, particularly the TAP-tagging approach in yeast, appears to have a sufficiently low false-negative rate to warrant a comprehensive analysis of the yeast genome. In such an approach, each protein in every complex would be tagged and associated proteins identified. The false-positive results would be minimized by considering only the interactions common to all proteins in the complex.

However, it is not clear whether the relative success of the TAP-tagging approach will extend to higher eukaryotes. The TAP-tagging method is particularly powerful in yeast because the tagged gene is inserted into the proper location in the genome and under the normal transcriptional controls. Higher eukaryotes do not have efficient machinery for homologous recombination and therefore making TAP-tagged proteins on a genome-scale is impractical. Expression of the tagged protein in higher eukaryotes will probably have to be driven from a heterologous promoter, and the protein expressed in unnatural amounts. If this is the case, we can perhaps expect to get error rates equivalent to those obtained by Ho and colleagues (15–50% false negative and 63–77% false positive), who used this overexpression approach in yeast.

Our analysis validates current efforts to create databases that take into account the inherently statistical and error-prone nature of the current protein-interaction literature and genome-wide experiments. If this is done correctly, one can achieve tremendous benefit from systematically integrating different datasets. Moreover, the networks being described by the current interaction databases need to be described in some sort of statistical or probabilistic terms to be physically realistic. Second, it will be important to derive structures for large protein complexes to provide precise and accurate information on protein interactions. Indeed, one of the most enduring values of large-scale structural biology could be the identification of accurate protein–protein interactions using complex crystal structures.

Acknowledgements
The authors wish to acknowledge funding from the NIH Structural Genomics Initiative (P50 GM62413–01) (AME, MG), the Ontario Research and Development Challenge Fund (AME) and the Keck foundation (MG). AME is a CIHR Scientist. JG is a CIHR Distinguished Scientist and an International Scholar of the HHMI.

References

1 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
2 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
3 Claverie, J.M. (2001) Gene number. What if there are only 30,000 human genes? *Science* 291, 1255–1257
4 Gavin, A.C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147
5 Ho, Y. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183
6 Uetz, P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627
7 Ito, T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4569–4574
8 Jansen, R. *et al.* (2002) Relating whole-genome expression data with protein-protein interactions. *Genome Res.* 12, 37–46
9 Kemmeren, P. *et al.* (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell* 9, 1133–1143
10 von Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein protein interactions. *Nature* 417, 399–403
11 Aloy, P. and Russell, R.B. (2002) Interrogating protein interaction networks through structural biology. *Proc. Natl. Acad. Sci. U. S. A.* 99, 5896–5901
12 Berman, H.M. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242
13 Janin, J. *et al.* (1988) Surface, subunit interfaces and interior of oligomeric proteins. *J. Mol. Biol.* 204, 155–164
14 Cramer, P. *et al.* (2000) Architecture of RNA polymerase II and implications for the transcription mechanism. *Science* 288, 640–649
15 Cramer, P. *et al.* (2001) Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* 292, 1863–1876
16 Robinson, R.C. *et al.* (2001) Crystal structure of Arp2/3 complex. *Science* 294, 1679–1684
17 Groll, M. *et al.* (1997) Structure of 20S proteasome from yeast at 2.4 Å resolution. *Nature* 386, 463–471
18 Edwards, A.M. *et al.* (1991) Two dissociable subunits of yeast RNA polymerase II stimulate the initiation of transcription at a promoter *in vitro*. *J. Biol. Chem.* 266, 71–75
19 Gnatt, A.L. *et al.* (2001) Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 Å resolution. *Science* 292, 1876–1882
20 Acker, J. *et al.* (1997) Interactions between the human RNA polymerase II subunits. *J. Biol. Chem.* 272, 16815–16821
21 Kimura, M. and Ishihama, A. (2000) Involvement of multiple subunit-subunit contacts in the assembly of RNA polymerase II. *Nucleic Acids Res.* 28, 952–959
22 Larkin, R.M. and Guilfoyle, T.J. (1997) Reconstitution of yeast and *Arabidopsis* RNA polymerase alpha-like subunit heterodimers. *J. Biol. Chem.* 272, 12824–12830
23 Ulmasov, T. *et al.* (1996) Association between 36- and 13.6-kDa alpha-like subunits of *Arabidopsis thaliana* RNA polymerase II. *J. Biol. Chem.* 271, 5085–5094
24 Yasui, K. *et al.* (1998) Location of subunit-subunit contact sites on RNA polymerase II subunit 3 from the fission yeast *Schizosaccharomyces pombe*. *Biochemistry* 37, 5542–5548
25 Miyao, T. *et al.* (1996) Molecular assembly of RNA polymerase II from the fission yeast *Schizosaccharomyces pombe*: subunit-subunit contact network involving Rpb5. *Genes Cells* 1, 843–854
26 Ishiguro, A. *et al.* (1998) Two large subunits of the fission yeast RNA polymerase II provide platforms for the assembly of small subunits. *J. Mol. Biol.* 279, 703–712
27 Mullins, R.D. *et al.* (1997) Structure, subunit topology, and actin-binding activity of the Arp2/3 complex from Acanthamoeba. *J. Cell Biol.* 136, 331–343
28 Mullins, R.D. and Pollard, T.D. (1999) Structure and function of the Arp2/3 complex. *Curr. Opin. Struct. Biol.* 9, 244–249
29 Kopp, F. *et al.* (1997) Subunit arrangement in the human 20S proteasome. *Proc. Natl. Acad. Sci. U. S. A.* 94, 2939–2944
30 Mrowka, R. *et al.* (2001) Is there a bias in proteome research? *Genome Res.* 11, 1971–1973
31 Cagney, G. *et al.* (2001) Two-hybrid analysis of the *Saccharomyces cerevisiae* 26S proteasome. *Physiol. Genomics* 7, 27–34
32 Zhao, X. *et al.* (2001) Interactions among subunits of human Arp2/3 complex: p20-Arc as the hub. *Biochem. Biophys. Res. Commun.* 280, 513–517
33 Cho, R. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2, 65–73
34 Mewes, H.W. *et al.* (2000) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 28, 37–40
35 Gerstein, M. *et al.* (2002) Proteomics. Integrating interactomes. *Science* 295, 284–287
36 Jansen, R. *et al.* Integration of genomic datasets to predict protein complexes in yeast. *J. Struct. Funct. Genomics* (in press)
37 Deane, C.M. *et al.* Protein interactions: Two methods for assessment of the reliability of high-throughput observations. *Mol. Cell. Proteomics* (in press)
38 Ge, H. (2001) *et al.* Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.* 29, 482–486
39 Tong, A.H. *et al.* (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* 295, 321–324
40 Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann