

Making the most of high-throughput protein-interaction data

Robert Gentleman* and Wolfgang Huber[†]

Addresses: *Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA. [†]European Bioinformatics Institute, European Molecular Biology Laboratory, Cambridge CB10 1SD, UK.

Published: 2 November 2007

Genome Biology 2007, **8**:112 (doi:10.1186/gb-2007-8-10-112)

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2007/8/10/112>

© 2007 BioMed Central Ltd

Abstract

We review the estimation of coverage and error rate in high-throughput protein-protein interaction datasets and argue that reports of the low quality of such data are to a substantial extent based on misinterpretations. Probabilistic statistical models and methods can be used to estimate properties of interest and to make the best use of the available data.

Most protein functions involve their interaction with other molecules, often with other proteins in the assembly of operational complexes. A better understanding of protein interactions is fundamental to the study of biological systems. As many drugs act on proteins, it is also a prerequisite for understanding intended, and unintended, drug effects. Over the past few years a number of large-scale experiments have set out to map protein interactions systematically [1-15]. While there is interest in combining the resulting data, there appear to be substantial discrepancies between experiments, and evaluation studies have reported large error rates, lack of overlap and apparent contradictions between the different datasets [16-21].

The purpose of this article is to critically assess the methodology used to analyze protein-interaction datasets. When interpreting individual experiments or combining datasets from different experiments, we need to consider three questions: first, what do we want to know and which experiments provide data that can be used to answer our questions; second, which types of protein interactions were assayed and under what conditions; and third, what types of measurement errors may have occurred and what is their prevalence. In this article we will discuss how the formulation of appropriate statistical models can allow investigators to clearly identify and estimate quantities of interest.

We will consider two particular types of protein interactions: physical interactions, and interactions between members of

a protein complex - which we shall call 'complex membership interactions'. A physical interaction is a direct and specific contact between a pair of proteins [22]. We regard two proteins in a complex as having a physical interaction if they share an interaction surface. A complex membership interaction exists between proteins that are part of the same multiprotein complex and does not necessarily imply a physical interaction.

Sampling and coverage

The two most widely used experimental techniques for detecting protein-protein interactions are the yeast two-hybrid (Y2H) system [23] and affinity purification followed by mass spectrometry (AP-MS) [24]. The Y2H system assays whether proteins can physically interact with each other. Large-scale experiments are carried out in a colony-array format, in which each yeast colony expresses a defined pair of 'bait' and 'prey' proteins that can be scored for reporter gene activity - indicating interaction - in an automated manner [1,6,25]. The type of information obtained from a Y2H experiment is shown in Figure 1. In an AP-MS experiment, a tagged protein is expressed in yeast and then 'pulled down' from a cell extract, along with any proteins associated with it, by co-immunoprecipitation or by tandem affinity purification. The set of pulled-down proteins is identified by MS. In a laborious and expensive process, this procedure has been systematically applied to large sets of yeast proteins [7-11]. The tagged protein in AP-MS is also sometimes called

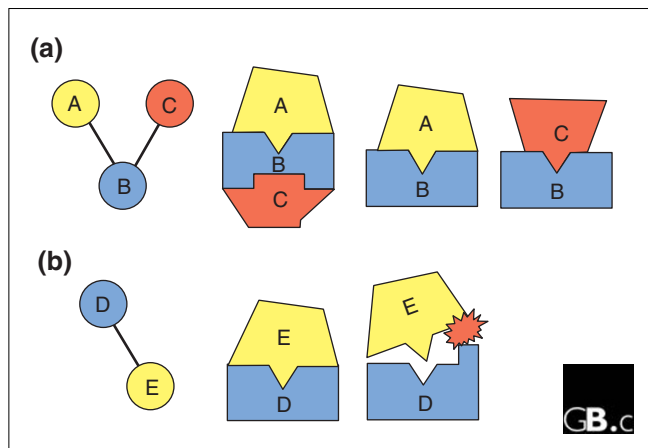


Figure 1
Interpreting results on direct physical interactions from Y2H experiments. **(a)** The observation of interactions A-B and B-C in a Y2H experiment does not indicate whether the two interactions can take place simultaneously (center) or whether they are exclusive of each other (right). **(b)** The ability of two proteins to interact may depend on post-translational modifications whose presence or absence may be actively regulated. Proteins D and E interact (center) in the absence of a certain post-translational modification (red shape), whose presence inhibits the interaction (right).

the bait and the proteins it pulls down the prey. The information on protein complexes given by Y2H and AP-MS experiments is compared in Figure 2.

An appreciation of the concepts of sampling and coverage is vital for interpreting the data from these types of experiments [26,27]. The term ‘sampling’ is used for experimental designs where only a subset of the population is interrogated. Representative sampling techniques are used in many fields of science, but they are not common in the generation of protein-interaction datasets, where sampling has often been guided by biological priorities. The ‘coverage’ summarizes which part of the total set of possible interactions has actually been tested. Even when genome-wide screening was intended [1,10,11], coverage was in fact well below 100%, and the success for each bait seems to depend on nonrandom biological, technological and economic factors. For example, Gavin *et al.* [10] used all 6,466 open reading frames (ORFs) that were at that time annotated in the *Saccharomyces cerevisiae* genome and obtained tandem affinity purifications for 1,993 of those. The remaining 4,473 (69%) failed at various stages, because, for example, the tagged protein failed to express or protein bands were not well separated by gel electrophoresis. Thus, neither the set of tested baits nor the set of tested prey in current experiments are random subsets of all proteins in the organism and in general, it is not valid to make inferences about the ‘population’, that is, the set of all physical interactions that take place in a cell under the conditions being studied, by assuming the available experimental data from a Y2H or AP-MS experiment to be a representative sample. We are not arguing that random sampling be used, as it would not

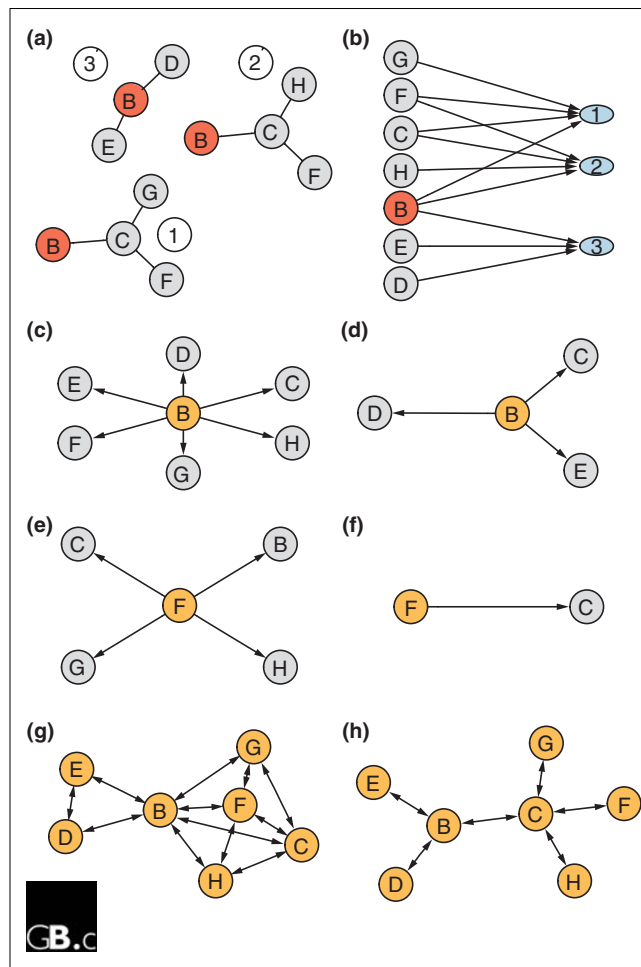


Figure 2
The manifestation of protein complexes in Y2H and AP-MS data. AP-MS experiments measure complex co-membership, and the fact that a prey is found by a certain bait means that there is either a direct physical interaction or an indirect physical interaction mediated by a protein complex. The set of proteins pulled down by a particular bait cannot therefore be equated with a single complex: if the bait is part of several different complexes, then the set of prey will be the union of all proteins in all complexes. **(a)** Protein B is involved in three different multiprotein complexes. In two of these it directly interacts with C, which itself can also interact with proteins F, G or H, whereas in the third complex, B interacts with D and E. **(b)** Assuming there are no other interactions under the conditions of the experiment, the bipartite graph between proteins B, ... H and complexes 1, 2, and 3 will look like this. **(c,d)** The result of a hypothetical AP-MS experiment with no false positives and no false negatives when (c) B is used as a bait and (e) F is used as a bait. **(e,f)** Result from a hypothetical Y2H experiment with a genome-wide set of preys and with no false positives and false negatives when (d) B is used as a bait and (f) F is used as a bait. **(g,h)** The results of (g) an ideal AP-MS experiment and (h) an ideal Y2H experiment if all proteins were used as baits. The Y2H data in (e,f,h) identifies the direct interactions, but it does not contain information on the number and architecture of the complexes. The maximal cliques identified by the AP-MS experiment in (g) correspond to the complexes in (a). However, the AP-MS data do not contain information on the topology of the direct interactions within each complex.

be appropriate in this setting, but rather that the data need to be interpreted more judiciously.

One problem in evaluating large-scale protein-interaction experiments is that the published data are often not sufficiently detailed to allow accurate description of the sets of baits and prey that were actually tested. As a proxy, we introduced the concept of 'viable baits' and 'viable prey' [28]. The first is the set of baits that were reported to have interacted with at least one prey, and the latter are those proteins reported to be found by at least one bait. Numbers for these can be unambiguously obtained from the reported data and provide surrogate measures for the tested baits and tested prey. The set of all pairs between viable bait and viable prey are the interactions that we are confident were experimentally tested and could, in principle, have been detected. The failure to detect an interaction between a viable bait and a viable prey is informative, whereas the absence of an observed interaction between an untested bait and prey is not. We note that the set of viable prey is a subset of the tested prey, and viable baits are a subset of the tested baits. This approach might introduce bias, because negative data from baits that were tested but found no prey, as well as from prey that were present but did not interact with any bait, are not recorded. On the other hand, presuming that combinations were tested, when in fact they were not, can also result in bias. Gilchrist *et al.* [29] used a randomization approach to estimate the size of the prey populations for the datasets in [7] and [8]. Their estimates are about double those of the number of viable prey.

Representation as graphs

Graph theory offers a convenient and useful set of terms and concepts to represent relationships between entities. Graphs most commonly represent binary relationships and these can be either directed or undirected. A further type of graph is needed to represent the membership of proteins in complexes: this relationship is not binary and requires a type of graph called a bipartite graph. Box 1 gives precise definitions of these concepts and an overview of how they apply to protein-interaction data.

Undirected graphs are often used as a model for physical interactions. True relationships are symmetric: if protein A interacts with B, then B interacts with A. The observed experimental data, however, often display asymmetry, which is a consequence of the experimental asymmetry between bait and prey. Protein A may identify protein B as an interactor when A is used as a prey, but B as a prey may not find A. To represent asymmetric data, we suggest using a directed-graph model. This is a point on which we diverge from much of the current practice. We argue that although the quantity of interest is an unknown undirected graph, it must be estimated from the observed data, which should be represented as a directed graph.

"All models are wrong, but some are useful." This maxim of George Box [30] reminds us that we should not expect these

models to adequately represent all possible aspects of protein interactions in a satisfactory way. For the current types of data and questions, graph models are useful. As the data and the questions that we ask become more sophisticated, more complicated models are likely to be needed. Some limitations of the graph models described here are related to their lack of resolution in time and space, failure to distinguish between different protein isoforms or post-translational modifications, and to the fact that experiments do not record interactions between individual protein molecules but between populations. It is the lack of such information that makes it difficult to use Y2H data to make inference about the composition of protein complexes (see Figure 1) or to use AP-MS data to identify the physical interactions of the proteins within a complex and their stoichiometry (see Figure 2).

Error statistics

Whether two proteins physically interact *in vivo* is not always simple to determine: the range of binding affinities of biologically relevant protein interactions spans many orders of magnitude [31], and interactions can be dynamic, transient and highly regulated. Nevertheless, the simple measurement model used to interpret the results of protein-interaction experiments presumes that for each pair of proteins, the question of whether or not they interact can be answered as either yes or no. The aim of making a measurement is to record the true, typically unknown, value of a physical quantity, but in practice there will be deviations - measurement errors. In such circumstances, statistical methods can be used to infer the true value of a quantity, given the data and some assumptions about how the measurement tool works. In this sense, the Y2H system or an AP-MS screen are simply measurement tools that provide imperfect data from which we make inferences about the true state of nature.

Standard definitions of various error statistics [32] are given in Box 2. We give them to enable a coherent dialog and to address some of the confusion in the literature. For example, a widely cited evaluation study by Edwards *et al.* [17] reported a "false positive rate" defined as $FP/(TP + FP)$: where FP is the number of false positives and TP the number of true positives. However, the more common name for this quantity is the 'false-discovery rate' (see Box 2). The difference between the false-positive rate, as usually defined by FP/N , and the false-discovery rate can be substantial, as their denominators are very different, N being the true tested non-interactions, given by $TN + FP$ (see Box 2). Incompatible terminology leads to confusion and makes comparison of error rates reported in different studies difficult.

Measurement errors can be decomposed into two components: stochastic and systematic errors. Stochastic errors are associated with random variability, whereas systematic errors are recurrent. Stochastic errors are simpler to

Box 1. The terminology of graphs

Undirected graphs

An undirected graph consists of a set of nodes V and a set of edges E and is denoted as $G = (V, E)$. Each element of the edge set E is an unordered pair (u, v) of nodes, and the two nodes in a pair are called 'adjacent'. The neighborhood of a node v is the set of nodes $N(v)$ to which it is adjacent, and its 'degree' $\delta(v)$ is the number of its neighbors, $\delta(v) = |N(v)|$. A subgraph S of a graph G contains a node set $V_S \subseteq V$ and an edge set $E_S = \{(u, v) \in E | u, v \in V_S\}$. The unordered pairs defining each edge $e \in E$ represent symmetric binary relationships between the elements of the node set. Undirected graphs can succinctly model physical protein interactions. The node set of a protein-protein interaction graph consists of all the individual proteins in the biological system of interest, and the edge set indicates which pairs of proteins physically interact.

Directed graphs

The definition of a directed graph builds upon that of undirected graphs, the only difference being that the edges are ordered. By convention, the direction of an edge (u, v) originates from u towards v . The edges (u, v) and (v, u) are distinct, and a graph may contain either one or both. The notion of degree in a directed graph is separated into two distinct concepts: 'indegree' and 'outdegree'. The outdegree, $\delta_o(v)$, of a node v is the number of directed edges that originate at v (out-edges). Its indegree, $\delta_i(v)$, is the number of edges that flow towards v (in-edges). Directed graphs can be used to represent Y2H data as well as AP-MS data. An edge $A \rightarrow B$ indicates that an interaction was tested with protein A as a bait and protein B when used as a prey. The result of the measurement is either positive or negative and can be represented as an edge attribute.

Bipartite graphs

Bipartite graphs or membership graphs are useful to represent the grouping of objects. They have two distinct types of nodes, and edges only connect a node of one type to a node of the other. For example, the proteins of a biological system could be the nodes of one type, its functional modules that of the other, and an edge in the bipartite graph represents membership of a protein in a module. Proteins can be members of multiple modules, and some proteins might not be assigned to any module.

One-mode graphs

Two graphs called one-mode graphs can be derived from a bipartite graph. If U and W are the node partitions of a bipartite graph G , then the edges in the one-mode graph on U (in respect of W) are determined by whether or not the two nodes both have edges in G to a common element of W (in respect of U). If A is the $|U| \times |W|$ adjacency matrix of the bipartite graph, then the one-mode graph for the node set U can be obtained by $A \otimes A^t$ and the one-mode graph for W by $A^t \otimes A$. The symbol \otimes represents matrix multiplication under Boolean algebra and the superscript t indicates matrix transposition. The one-mode graph of the proteins is the complex membership graph: two nodes are connected if they are members of the same complex. Similarly, the one-mode graph of the complexes is the complex overlap graph: two complexes are connected to each other in this graph if there is at least one protein that is a member of both.

Maximal cliques

A clique is a fully connected subgraph. A maximal clique is a clique that is not a proper subset of another clique.

address: they can be controlled by replication, can be eventually eliminated if the experiment is repeated many times, and they can often readily be described using probability models. Systematic errors give rise to bias: the quantity being measured is consistently different from the truth. Their identification is difficult, but if it can be done, they can be addressed either by improving the experimental procedures or by developing appropriate methods for post-experiment data processing.

Statistical models for the analysis of protein-interaction data

Statistical models can integrate the information from repeated or related measurements and quantify the (un)certainly that we have about the conclusions. Here we consider how statistical techniques have been applied to two distinct problems: estimating membership of a protein complex and the integration of data from different experiments (cross-experiment integration of data).

Box 2. Standard definitions of error terms

True positives (TP): Number of cases in which a true interaction is experimentally observed.

True negatives (TN): Number of cases in which two proteins do not interact (truly absent interaction); their interaction is tested but not observed.

False positives (FP): Number of cases in which two proteins do not interact, but an interaction is experimentally observed.

False negatives (FN): Number of cases in which a true interaction is experimentally tested and not observed.

True tested interactions (P): TP + FN

True tested non-interactions (N): TN + FP

False-positive rate (p_{FP}): Probability that a truly absent interaction is detected. It can be estimated by FP/N.

False-negative rate (p_{FN}): Probability that a true interaction is not detected. It can be estimated by FN/P.

Sensitivity: Probability that a true interaction is detected. It can be estimated by TP/P.

Specificity: Probability that a truly absent interaction is not detected, estimated by TN/N.

False-discovery rate (FDR): Informally, the expected value of FP/(TP + FP) [42].

Positive predictive value (PPV): Probability that an observed interaction is indeed true. It can be estimated by TP/(TP + FP).

Negative predictive value (NPV): Probability that an observed non-interaction is truly absent. It can be estimated by TN/(TN + FN).

See [32] for a more extensive discussion of these concepts. The probabilities are conditional on whether the interaction is tested.

Estimating membership of a protein complex

Russell and colleagues [10] have developed a heuristic that they term the 'socioaffinity index', A_{ij} . It quantifies the confidence that proteins i and j share complex membership, given a set of protein purifications each with its bait and a

number of prey. The score is the logarithm of the product of three odds-ratios. The first odds-ratio compares the frequency with which bait i pulled down prey j to the frequency that would be expected if prey came down randomly; the second is the corresponding value for bait j pulling down prey i ; and the third is the ratio of frequency of co-occurrence of i and j in a pull-down to what would be expected under random sampling. The authors then apply a customized clustering algorithm to the matrix A_{ij} to estimate sets of protein complexes from AP-MS data.

Scholten and colleagues took a different route [33,34]. They explicitly modeled the underlying bipartite graph of membership of proteins in protein complexes. They estimated the bipartite graph from the observed data using a penalized likelihood method. Their method explicitly differentiates between tested and untested edges in the data, and it deals with the possibility that some proteins can be members of multiple complexes and others may not be assignable to any.

Cross-experiment integration of data

Turning to the issue of the cross-experiment integration of data, Gilchrist and colleagues [29] described a statistical model for identifying stochastic errors in protein-protein interaction datasets that is based on the Binomial distribution. They assumed that there is a true underlying graph of protein interactions in the biological system under study and that multiple experimental runs are performed, each resulting in a set of observed edges. A true edge is observed with probability $1 - p_{FN}$ and missed with the false-negative probability p_{FN} . Similarly, a true non-edge is observed as an edge with false-positive probability p_{FP} and not observed with probability $1 - p_{FP}$. They assumed that all these stochastic events are independent of each other, and governed only by the two Binomial rates p_{FP} and p_{FN} . The statistical distribution of the number of observed edges S between two proteins, given n_t trials, and conditional on whether or not they truly interact, is then simply given by Binomial distributions:

$$S \mid \text{true edge} \sim \text{Bin}(n_t, 1 - p_{FN}) \quad (1)$$

$$S \mid \text{true non-edge} \sim \text{Bin}(n_t, p_{FP}) \quad (2)$$

From this, the authors constructed a maximum likelihood estimator of p_{FP} and p_{FN} , and a likelihood-ratio test to decide, for any pair of proteins, whether the data suggest an interaction between them.

Krogan and colleagues [11,35] took an approach that is similar in spirit to that of Gilchrist *et al.* [29]. Their formulation uses a Bayes factor that compares the probability of the observed data under the two possible alternatives, and a further component that represents the prior odds of an interaction. The use of a Bayes factor in this context is entirely appropriate, but given that the selection of baits is typically not a simple random sample from the population of potential baits, it is somewhat difficult to interpret the role of

Table 1**Overview of seven Y2H and five AP-MS experiments**

Reference	VB	CB	TB	VP	VBP	VBP/BP	TI	TI/VB	REC	UNR
Ito <i>et al.</i> [1]	1,522		6,604	2,493	773	0.51	4,524	3.0	75	803
Cagney <i>et al.</i> [2]	19		31	40	11	0.58	54	2.9	3	4
Tong <i>et al.</i> [3]	20		22	59	5	0.25	115	5.8	1	1
Hazbun <i>et al.</i> [4]	66		100	1,940	28	0.42	2,524	38	4	13
Zhao <i>et al.</i> [5]	1		1	90	0	0.00	90	90	0	0
Uetz <i>et al.</i> Experiment 1 [6]	508		6,604	630	142	0.28	952	1.9	10	47
Uetz <i>et al.</i> Experiment 2 [6]	139		192	400	36	0.26	524	3.8	18	7
Gavin <i>et al.</i> [7]	455	600	725	1,179	271	0.60	3,419	7.5	192	314
Ho <i>et al.</i> [8]	493	589	1,739	1,316	231	0.47	3,687	7.5	69	297
Krogan <i>et al.</i> [9]	153	165	165	483	151	0.99	1,132	7.4	89	157
Gavin <i>et al.</i> [10]	1,752	1,993	6,466	1,790	991	0.57	19,105	10.9	1,077	4,297
Krogan <i>et al.</i> [11]	2,264	2,357	4,562	5,323	2,226	0.98	63,360	28.0	1,969	34,363

VB, the number of viable baits; CB, the number of cloned (hybridized) baits, if available; TB, the total number of baits that the experimenters were initially aiming at; VP, the number of viable prey; VBP, the number of proteins observed as both bait and prey; TI, the total number of interactions observed; REC, the number of reciprocated interactions between proteins that were observed as both bait and prey; UNR, the number of unreciprocated interactions between proteins that were observed as both bait and prey. Not all of the experiments were genome-wide - some were focused on particular aspects of the cellular machinery [2-5,9]. Even in the so-called genome-wide studies [1,6-8,10,11], however, the viable baits cover only around a third of the yeast genes. This means that the largest part of interaction space by far, containing interactions between proteins not used as baits, was not sampled in any of these experiments. We can also see that TI/VB, the average number of interactions per viable bait, varies markedly between experiments. In the more focused studies, this will certainly be a result of different criteria for the selection of baits. In the genome-wide screens it may indicate the application of different, experiment-specific cutoffs.

the prior, and it seems some justification is needed. The two approaches [29,35] differ somewhat in how specific quantities, such as p_{FP} and p_{FN} , are estimated. An important difference is that Krogan and colleagues [35] were specifically interested in combining AP-MS datasets to solve the problem of identifying protein complexes.

Internal error rate estimation using reciprocity

The direction of an observed bait-prey interaction is informative for the estimation of error rates and the identification of systematic errors. If two proteins A and B are each tested both as bait and prey, then ideally we expect reciprocity in their interaction data: if they truly interact, bait A should find prey B and bait B should find prey A. If they truly do not interact, there should be no observed interaction in either direction. In real data there will be many pairs of proteins for which reciprocity does not hold, and these cases imply that either a false positive or a false negative measurement was made. Comparing the prevalence of reciprocally measured interactions among the reciprocally tested edges can tell us something about error rates, both stochastic and systematic.

As the set of reciprocally tested edges is usually not explicitly recorded, we have used the concept of viable baits and viable prey to produce Table 1, which gives the numbers of viable

bait and prey proteins, and based on this, the numbers of reciprocated and unreciprocated interaction measurements for several large-scale Y2H and AP-MS experiments. We can represent these data for each experiment as a directed subgraph G_{BP} , with nodes being the intersection of viable baits and viable prey, and with directed edges each representing an observed interaction of a bait with a prey. There are several experiments in which G_{BP} is sufficiently large for statistical analysis, and the usefulness of the reciprocity criterion can be used to measure the internal consistency of a dataset [28].

To identify proteins that are likely to be subject to systematic experimental error, we can compare their in-edges and out-edges (see Box 1) within the directed subgraph G_{BP} . Ideally, these edges should all reciprocate each other; if a certain protein has very many unreciprocated edges, this indicates that it is likely to be affected by a systematic error. To quantify this, the number of unreciprocated edges, n_{unr} , originating from or pointing to a particular protein can be compared with the number of reciprocated edges that it has and to the false-positive and false-negative rates p_{FP} and p_{FN} . Precise estimation of these rates is difficult, however, and a simple and effective criterion can instead be derived from considering symmetry.

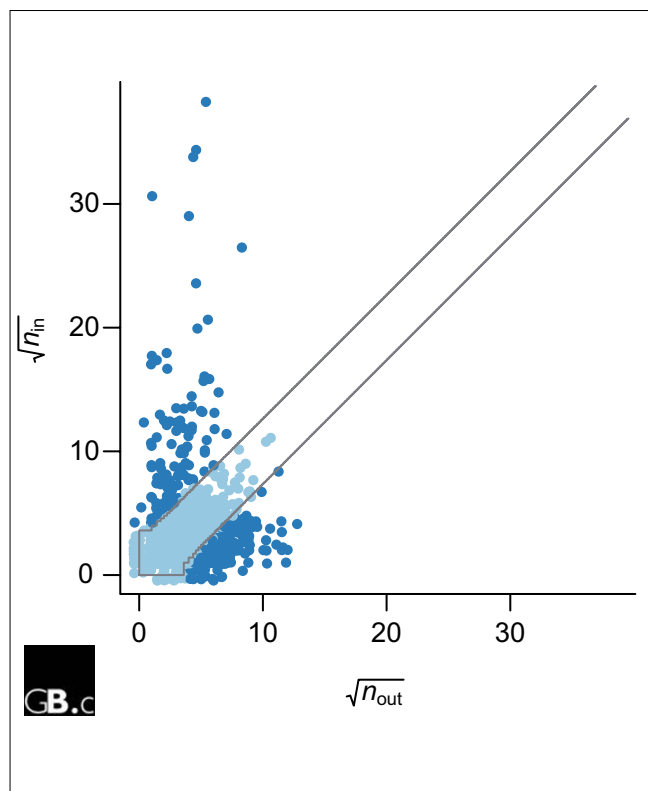


Figure 3

Scatterplot of n_{in} and n_{out} for the AP-MS data of Krogan *et al.* [11]. Each point in the plot corresponds to one protein. n_{in} is the number of times that the protein was found as a prey; n_{out} the number of prey it found when used as a bait. The two lines mark contours of probability $p = 10^{-4}$ according to the Binomial model in Equation (3). Outlying proteins (dark blue) show a significantly large difference between n_{in} and n_{out} , suggesting that at least one of them is wrong. For example, if $n_{out} \gg n_{in}$, one possible reason is that a protein is not expressed when used as prey or of such low abundance that it is outcompeted, but when tagged and expressed as a bait, it will identify and pull down its interaction partners as prey. Further validation experiments are needed to determine in each case whether the unreciprocated interactions correspond to false-positive or false-negative observations.

For a given number of unreciprocated edges, n_{unr} , if there are no systematic errors then the unreciprocated edges should be in-edges and out-edges in approximately equal numbers. If we denote their numbers by n_{in} and n_{out} , respectively, then $n_{in} + n_{out} = n_{unr}$, and we expect that

$$n_{in} \sim \text{Bin}(n_{unr}, 0.5) \quad (3)$$

If n_{in} and n_{out} are significantly different from each other, according to the Binomial distribution we would conclude that the protein behaved differently in the experiment when used as bait compared with prey, and would use this as an indication of systematic error affecting at least part of the data for that protein. An application of this criterion to the subgraph G_{BP} of the data of Krogan *et al.* [11] is shown in Figure 3.

Estimation of the properties of the interaction graph in this setting

There are two basic approaches to estimation: one is to estimate the true underlying graph, given the data and some modeling assumptions, then to calculate properties of interest from the estimated graph. The other is to directly estimate the quantities of interest without making an attempt to estimate the true underlying graph. For protein-interaction data we suggest that the latter is often preferable, as it can deal better with the low coverage of the datasets. As new methods and models for integrating datasets are developed it will be important to reassess the situation.

We distinguish between two different types of quantities to be estimated. The first type are single numeric values, such as degree, clustering coefficient or diameter. The second are more general structures, such as modules or subgraphs. The tools for estimation are more developed for numeric quantities than for modules, and there is agreement on the definitions of the different quantities. For modules, or cohesive subgroups, there is little agreement on what is being sought or how to find it.

The integration of data from different independent experiments

No single experiment has provided complete information on all interactions in a system of interest and so data from different experiments need to be integrated. Integration promises to increase coverage and reduce the effects of stochastic errors. Table 1 summarizes experiments done on the yeast protein interactome that are candidates for integration. The overlap between experiments is examined in Tables 2 and 3.

An essential step before integration of data is to assess their quality in terms of specificity, sensitivity and coverage. Such an assessment should provide reliable estimates of the false-positive and false-negative error rates. There are three main computational approaches: comparison to a benchmark or 'gold standard' data, within-experiment or internal validation, and between-experiment validation.

When direct physical interactions are being measured (for example, by Y2H), crystal structures of the interacting proteins can be used as the gold standard for the validity of the interaction. This was one of the approaches used in [17]. Only a handful of crystal structures of interacting proteins are known, however, and such data are still difficult and expensive to obtain. Some physical interactions and protein complexes have also been characterized through detailed biochemical investigations, and are collected in databases such as MIPS [36] and GO [37]. Circularity needs to be avoided, however; for example, the data from [7] and [9] are now reported as known complexes in some of the public protein complex databases.

Table 2

Pairwise comparison of Y2H datasets

References	Ito <i>et al.</i> [1]	Cagney <i>et al.</i> [2]	Tong <i>et al.</i> [3]	Hazbun <i>et al.</i> [4]	Zhao <i>et al.</i> [5]	Uetz <i>et al.</i> [6] Experiment 1	Uetz <i>et al.</i> [6] Experiment 2
[1]	-	9	7	24	1	224	47
[2]	28	-	0	0	0	7	3
[3]	34	0	-	0	0	4	7
[4]	856	14	25	-	0	15	12
[5]	43	1	2	38	-	0	0
[6] Experiment 1	388	14	22	272	15	-	36
[6] Experiment 2	200	9	26	204	13	108	-

The values above the diagonal give the number of viable baits in common between each pair of experiments, and the values below the diagonal give the number of viable prey in common. We see that the overlap between experiments in the sampled fractions of protein-interaction space is in all cases very small, given that thousands of interactions were assayed.

Table 3

Pairwise comparison of AP-MS datasets

References	Gavin <i>et al.</i> [7]	Ho <i>et al.</i> [8]	Krogan <i>et al.</i> [9]	Gavin <i>et al.</i> [10]	Krogan <i>et al.</i> [11]
[7]	-	82	51	442	334
[8]	516	-	25	222	286
[9]	299	246	-	121	151
[10]	1,143	717	371	-	1,128
[11]	1,149	1,277	478	1,732	-

As in Table 2 the values above the diagonal give the number of viable baits in common between each pair of experiments, and the values below the diagonal give the number of viable prey in common. Again, the overlap is very small. Consider the two largest experiments carried out so far: with a set of 2,264 viable baits and 5,323 viable prey, Krogan *et al.* [11] tested for the presence of at least 12 million complex membership interactions. Gavin *et al.* [10], with 1,752 viable baits and 1,790 viable prey, tested for at least 3.1 million interactions. However, even for these two datasets, the largest so far, the known overlap is only $1,128 \times 1,732 \approx 2.0$ million. One of the possible explanations for these low estimates of coverage and overlap is that our definitions of viable baits and viable prey are restrictive and that indeed a much larger space of interactions might have been tested. For example, Gilchrist *et al.* [29] estimated a value about twice ours for the number of tested prey in [7]. This situation will hopefully be alleviated as researchers report more complete data on which interactions were actually tested.

Within-experiment validation relies on internal properties of the data, such as redundancies or symmetries that are not used in the experiment, and that can therefore be used to validate the experimental results. One such property is reciprocity, as discussed above. Deviations from expectation can be used to estimate stochastic error rates, and they can also be used to identify individual proteins whose data appear to be subject to systematic artifacts (see Figure 3).

Reported replicate measurements can also be used to help validate experimental data and to estimate error rates. The basic idea is that if edges are tested multiple times under the same conditions, those that are found frequently can be termed true positives and can be used to estimate the false-negative rate from those cases when they were missed. Similarly, those that are seldom found can be deemed true negatives, and from the positive data points the false-positive rate can be estimated. This approach is complicated

by possible dependencies between the replicate measurements and by systematic errors that, if present, will affect all replicates. These complications may render the statistical model intractable. Further caution is warranted. Was the choice of replicates measures made *a priori* or because of anomalous results obtained during the experiment? Do they provide equal coverage of all important conditions and of all types of proteins that were studied?

Between-experiment comparisons rely on the experimental conditions being sufficiently similar to ensure that the measurements are made on the same underlying set of true interactions. However, as we see in Tables 2 and 3, in many cases there is relatively little overlap in bait selection and in observed prey. For two recent experiments with at least some overlap, a comparison was presented by [20]. These authors found a moderate overlap between the primary data, for example the proteins identified by each successful

bait, but a low overlap of the computed protein complexes by each group.

When integrating data from different experiments our recommendation is that validation to a gold standard and within-experiment validation should first be done on each experiment separately. Once the data are sufficiently well understood and as many of the systematic errors as possible have been resolved, integration becomes worthwhile. If there is little agreement on the existence of interactions for edges tested in different experiments, then one must question the prudence of their integration: it may be that the biological conditions were too different to allow their integration into a single meaningful dataset.

There is room for much more research here. Evidence in favor of, or against, experimentally detected interactions can often be obtained from other sources, such as data from other organisms, dependencies of different types of interactions on each other (for example, coexpression, colocalization and physical interaction), evolutionary conservation [38], protein structure [39] and amino-acid binding motifs [40]. The challenge is to ensure that the evidence is applicable and that it does bear relationship to the assay and system under study.

Our purpose in writing this article was to address the observation that the many different protein-interaction datasets available appear to have very little in common, and also to address reports that the data were inherently noisy and of low quality (for example [17,41]). Our investigations suggest that the data themselves, while problematic in some cases, are not the real issue, but rather there is often misinterpretation of the data, methods to address noisiness are often inadequate, and the lack of substantive comparisons between methods applied to the data has led to a situation where the data, rather than the methods, are treated with suspicion. As seen from Tables 2 and 3, low coverage, and not the false-positive rate, is responsible for the small amount of overlap between datasets.

The separation of errors into stochastic and systematic components is potentially of great benefit. Comparison of experimental data should be based on stochastic error rates. The identification of systematic errors can help to identify problems with the experimental techniques and hopefully suggest solutions to those problems. We believe that when more standard, and sound, statistical practices are adopted for preprocessing the data, it will be possible to estimate quantities of interest and to make substantial comparisons. An essential prerequisite is the adoption of standard methods for estimation of stochastic error rates and where possible the identification of systematic errors. Standardized preprocessing is also required in order to be able to synthesize different experimental datasets. Combining data requires attention to the differing error rates, and the

discounting of information from more variable experiments. Given the numbers in Tables 2 and 3, there is much to be gained by combining the different experimental datasets. We believe that the data, while noisy, are in fact very useful, and with appropriate preprocessing and statistical modeling they can provide deep insight into the functioning of cellular machineries.

Acknowledgements

We thank Richard Bourgon, Michael Boutros, Tony Chiang, Denise Scholtens and Lars Steinmetz for helpful comments on the manuscript. This work was supported by HFSP research grant RGP0022/2005 to W.H. and R.G.

References

- Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *Proc Natl Acad Sci USA* 2001, **98**:4569-4574.
- Cagney G, Uetz P, Fields S: **Two-hybrid analysis of the *Saccharomyces cerevisiae* 26S proteasome.** *Physiol Genomics* 2001, **7**: 27-34.
- Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B, Paoluzi S, et al.: **A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules.** *Science* 2002, **295**:321-324.
- Hazbun TR, Malmström L, Anderson S, Graczyk BJ, Fox B, Riffle M, Sundin BA, Aranda JD, McDonald WH, Chiu CH, et al.: **Assigning function to yeast proteins by integration of technologies.** *Mol Cell* 2003, **12**:1353-1365.
- Zhao R, Davey M, Hsu YC, Kaplanek P, Tong A, Parsons AB, Krogan N, Cagney G, Mai D, Greenblatt J, et al.: **Navigating the chaperone network: an integrative map of physical and genetic interactions mediated by the hsp90 chaperone.** *Cell* 2005, **120**:715-727.
- Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, et al.: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*.** *Nature* 2000, **403**:623-627.
- Gavin AC, Bösch M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon AM, Cruciat CM, et al.: **Functional organization of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415**:141-147.
- Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, et al.: **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry.** *Nature* 2002, **415**:180-183.
- Krogan NJ, Peng WT, Cagney G, Robinson MD, Haw R, Zhong G, Guo X, Zhang X, Canadien V, Richards DP, et al.: **High-definition macromolecular composition of yeast RNA-processing complexes.** *Mol Cell* 2004, **13**:225-239.
- Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dümpelfeld B, et al.: **Proteome survey reveals modularity of the yeast cell machinery.** *Nature* 2006, **440**:631-636.
- Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, et al.: **Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*.** *Nature* 2006, **440**:637-643.
- Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E, et al.: **A protein interaction map of *Drosophila melanogaster*.** *Science* 2003, **302**:1727-1736.
- Li S, Armstrong CM, Bertin N, Ge H, Milstein S, Boxem M, Vidalain PO, Han JD, Chesneau A, Hao T, et al.: **A map of the interactome network of the metazoan *C. elegans*.** *Science* 2004, **303**: 540-543.
- Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, et al.: **Towards a proteome-scale map of the human protein-protein interaction network.** *Nature* 2005, **437**:1173-1178.

15. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, et al.: **A human protein-protein interaction network: a resource for annotating the proteome.** *Cell* 2005, **122**:957-968.
16. Mrowka R, Patzak A, Herzel H: **Is there a bias in proteome research?** *Genome Res* 2001, **11**:1971-1973.
17. Edwards AM, Kus B, Jansen R, Greenbaum D, Greenblatt J, Gerstein M: **Bridging structural biology and genomics: assessing protein interaction data with known complexes.** *Trends Genet* 2002, **18**:529-536.
18. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions.** *Nature* 2002, **417**:399-403.
19. Goll J, Uetz P: **The elusive yeast interactome.** *Genome Biol* 2006, **7**:223.
20. Gagneur J, David L, Steinmetz LM: **Capturing cellular machines by systematic screens of protein complexes.** *Trends Microbiol* 2006, **14**:336-339.
21. Hart GT, Ramani AK, Marcotte EM: **How complete are current yeast and human protein-interaction networks?** *Genome Biol* 2006, **7**:120.
22. Jones S, Thornton JM: **Principles of protein-protein interactions.** *Proc Natl Acad Sci USA* 1996, **93**:13-20.
23. Fields S, Song O: **A novel genetic system to detect protein-protein interactions.** *Nature* 1989, **340**:245-246.
24. Kumar A, Snyder M: **Protein complexes take the bait.** *Nature* 2002, **415**:123-124.
25. Uetz P: **Two-hybrid arrays.** *Curr Opin Chem Biol* 2002, **6**:57-62.
26. Han JD, Dupuy D, Bertin N, Cusick ME, Vidal M: **Effect of sampling on topology predictions of protein-protein interaction networks.** *Nat Biotechnol* 2005, **23**:839-844.
27. Stumpf MPH, Wiuf C: **Sampling properties of random graphs: the degree distribution.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2005, **72**:036118.
28. Chiang T, Scholtens D, Sarkar D, Gentleman R, Huber W: **Coverage and error models of protein-protein interaction data by directed graph analysis.** *Genome Biol* 2007, **8**:R186.
29. Gilchrist MA, Salter LA, Wagner A: **A statistical framework for combining and interpreting proteomic datasets.** *Bioinformatics* 2004, **20**:689-700.
30. Box GEP, Draper NR: *Empirical Model-Building and Response Surfaces.* New York: Wiley; 1987.
31. Aloy P, Russell RB: **Structural systems biology: modelling protein interactions.** *Nat Rev Mol Cell Biol* 2006, **7**:188-197.
32. Kelsey JL, Whittemore AS, Evans AS, Thompson WD: **Methods in observational epidemiology.** In *Monographs in Epidemiology and Biostatistics*, New York: Oxford University Press; 1996.
33. Scholtens D, Gentleman R: **Making sense of high-throughput protein-protein interaction data.** *Stat Appl Genet Mol Biol* 2004, **3**:39.
34. Scholtens D, Vidal M, Gentleman R: **Local modeling of global interactome networks.** *Bioinformatics* 2005, **21**:3548-3557.
35. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ: **Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*.** *Mol Cell Proteomics* 2007, **6**:439-450.
36. Mewes HW, Frishman D, Mayer KF, Münsterkötter M, Noubibou O, Pagel P, Rattei T, Oesterheld M, Ruepp A, Stümpflen V: **MIPS: analysis and annotation of proteins from whole genomes in 2005.** *Nucleic Acids Res* 2006, **34(Database issue)**:D169-D172.
37. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, et al.: **The Gene Ontology (GO) database and informatics resource.** *Nucleic Acids Res* 2004, **32(Database issue)**:D258-D261.
38. Poyatos JF, Hurst LD: **How biologically relevant are interaction-based modules in protein networks?** *Genome Biol* 2004, **5**:R93.
39. Aloy P, Böttcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin AC, Bork P, Superti-Furga G, Serrano L, Russell RB: **Structure-based assembly of protein complexes in yeast.** *Science* 2004, **303**:2026-2029.
40. Neduva V, Russell RB: **Peptides mediating interaction networks: new leads at last.** *Curr Opin Biotechnol* 2006, **17**:465-471.
41. Chen J, Hsu W, Lee ML, Ng SK: **Increasing confidence of protein interactomes using network topological metrics.** *Bioinformatics* 2006, **22**:1998-2004.
42. Storey J: **A direct approach to false discovery rates.** *J R Stat Soc Ser B* 2002, **64**:479-498.