Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction

Ronald Jansen[1] and Mark Gerstein[2]*

The concept of 'protein function' is rather 'fuzzy' because it is often based on whimsical terms or contradictory nomenclature. This currently presents a challenge for functional genomics because precise definitions are essential for most computational approaches. Addressing this challenge, the notion of networks between biological entities (including molecular and genetic interaction networks as well as transcriptional regulatory relationships) potentially provides a unifying language suitable for the systematic description of protein function. Predicting the edges in protein networks requires reference sets of examples with known outcome (that is, 'gold standards'). Such reference sets should ideally include positive examples — as is now widely appreciated — but also, equally importantly, negative ones. Moreover, it is necessary to consider the expected relative occurrence of positives and negatives because this affects the misclassification rates of experiments and computational predictions. For instance, a reason why genome-wide, experimental protein–protein interaction networks have high inaccuracies is that the prior probability of finding interactions (positives) rather than non-interacting protein pairs (negatives) in unbiased screens is very small. These problems can be addressed by constructing well-defined sets of non-interacting proteins from subcellular localization data, which allows computing the probability of interactions based on evidence from multiple datasets.

Addresses
[1] Computational Biology Center, Memorial Sloan-Kettering Cancer Center, 307 East 63rd Street, 2nd floor, New York, New York 10021, USA
[2] Department of Molecular Biophysics and Biochemistry and Department of Computer Science, Yale University, Bass 432A, 266 Whitney Ave., New Haven, Connecticut 06520, USA
*e-mail: mark.gerstein@yale.edu

Current Opinion in Microbiology 2004, **7**:535–545

This review comes from a themed issue on
Genomics
Edited by Charles Boone and Philippe Glaser

Available online 15th September 2004

1369-5274/$ – see front matter

DOI 10.1016/j.mib.2004.08.012

Introduction

The availability of genome sequences for a range of prokaryotic and eukaryotic organisms has given us a comprehensive view of the parts list of genes encoded in these organisms, but the biological functions of many of these genes remain uncharacterized.

In recent years, several experimental methods have been developed to overcome this problem. They aim to systematically and globally characterize the basic properties of gene products and their interactions in these organisms, spawning a whole field of research termed 'functional genomics'. Functional genomics experiments are often unbiased screens of whole proteomes, on a large scale, rather than focusing on small-scale studies of individual proteins or groups of proteins (such as the members of an interesting pathway), as is more common in traditional and reductionist approaches of biological research. The hope is that functional genomics will allow us to gain a comprehensive understanding of the basic biology underlying cellular behavior.

Among the available techniques for globally characterizing genes and proteins are methods for the genome-wide measurement of transcription levels [1] and protein abundance [2], methods for determining deletion phenotypes of single genes [3] or combinations of them [4], global measurements of the subcellular localizations of proteins [5•,6•] as well as methods for measuring interactions between proteins [7–11] or between proteins and intergenic sequences in DNA [12–14].

Alongside these relatively new experimental approaches, a variety of computational techniques have become standard that are aimed at processing, managing and interpreting the large amounts of data that the experiments produce. Many of these computational techniques draw on methods developed for artificial intelligence, data-mining and statistical learning [15]. Machine-learning techniques, either of the unsupervised or supervised kind (depending on whether partial knowledge about the desired prediction outcome in reference datasets is used to train the algorithm or not), exploit statistical relationships between various types of functional genomics data and can be used to make computational predictions of protein properties [16].

In this review, we discuss several issues related to the successful application of such machine-learning algorithms. Among these is the problem of how to systematically define protein function. A more subtle issue is the proper definition of the reference sets. We will explain these issues using analysis of protein–protein interaction networks as an example.

Yeast as a model organism

The yeast *Saccharomyces cerevisiae* has become a central organism on which these experimental and computational

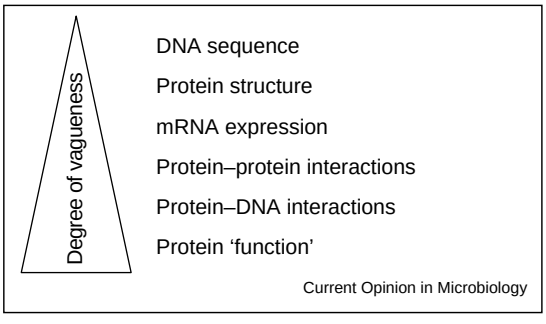
methods converge. Yeast is likely to be the first organism for which we will obtain a comprehensive description of most or all gene products based on functional genomics methods [17]. There are varied reasons for this, ranging from the technical advantages of yeast genetics, the relative simplicity of the single-cell organism (for instance, a relatively minor amount of gene splicing), a well-developed framework of existing functional annotations (such as MIPS [18] and Gene Ontology [19]), a strong tradition of yeast biology research as well as relatively high consistency among conventions and standards (for instance, yeast gene names are fairly systematically defined). The hope is that functional genomics techniques will eventually translate into a fundamental understanding of human biology and the causes of disease [13].

Uncertainty in functional genomics data

As we have moved from genome sequences to more advanced functional genomics data, it has become clear that they are associated with a considerable amount of uncertainty (Figure 1). The linear genome sequence is well-defined and, although there can be errors in DNA sequencing, quite reliable. Three-dimensional protein structures contain some more uncertainty. For instance, the exact position of the coordinates may be unknown; this is because of the limited resolution of X-ray crystallography or other structure determination methods or because of inherent changes in the protein structure that may involve motions of whole domains; the chain trace, however, is fairly certain.

On the next level, mRNA expression data from microarray experiments contains a much higher degree of experimental and biological uncertainty. The readout of such experiments is a continuous-valued, positive signal, often containing a high amount of error and noise. Protein–protein interaction and protein–DNA interaction screens contain a high amount of false positives and false negatives [20–23].

Figure 1



Uncertainty of functional genomics data. Whereas the genome sequences are considered to be fairly reliable (with high quality sequences having virtually 100% base accuracy), more advanced functional genomics datasets contain increasing amounts of uncertainty or vagueness. Protein 'function', which functional genomics aims to determine for many proteins on a large scale, is a concept that in itself is not very clearly defined.

There are a variety of statistical data analysis techniques that try to address these issues. A large part of the processing and analysis of microarray data are aimed at filtering the real, biological signal from noise that comes from non-biological sources; similar issues apply to the analysis of protein chip data [24,25].

The vague definition of 'function'

Finally, there is a degree of uncertainty related to the ultimate goal of functional genomics: the vague concept of protein 'function' itself. What do we mean by protein 'function'? The term somehow describes the 'biological process', 'cellular component' and 'molecular function' of a protein (as organized by Gene Ontology [19]). But how is this clearly defined? Gene names are sometimes abbreviations of a statement about the activity of a gene. But in other cases they simply reflect the way a gene was discovered historically. Sometimes arbitrary names are devoid of any biological meaning or, even worse, outright contradictory (Figure 2).

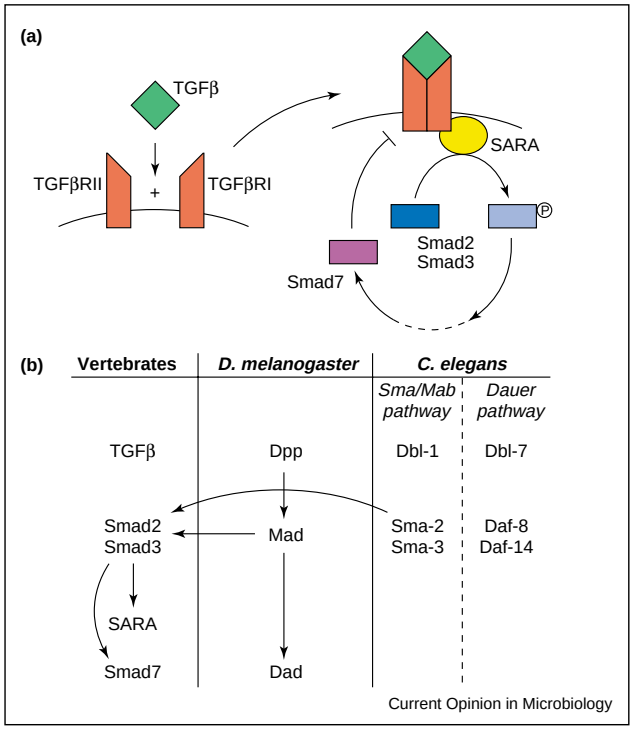
Another problem is that while it may be possible to find positive examples of proteins with the same or similar biological functions, it is difficult to find clear negative examples of two proteins that have absolutely no biological relationship at all.

Conceptual problems for predicting protein properties with machine-learning algorithms

Despite advances in artificial intelligence and data-mining, computational algorithms essentially require clear rules for processing data, and the absence of such clear rules negatively impacts the capability of such algorithms to predict protein properties. Supervised learning algorithms require that we have a subset of cases where we know the response variable (the protein property we would like to predict) as a function of the predictor variables (the collection of data from which the predictions are made). These cases are used to construct training and test sets (to train the algorithm and to cross-validate the prediction results). For a binary class prediction (where the prediction result is either 'positive' or 'negative'), it is necessary to have examples of both positives and negatives to construct the training and test sets [1,26]. This is an obvious statement, but how to define the negatives is not as obvious as it may seem.

Consider, for instance, how positive and negative examples are usually constructed when genomic data is used to predict the functional classes that proteins are associated with (such as those of the MIPS or the Gene Ontology functional classifications): the proteins of a desired functional class are labeled as positives and all other proteins as negatives. The prediction of multiple functional

Figure 2



The TGFβ pathway as an example of confusing and whimsical gene nomenclature. **(a)** In the canonical TGFβ pathway, the ligand TGFβ induces the formation of a heterodimer between two receptors (TGFβRII and TGFβRI), which then phosphorylate the intracellular Smad2 or Smad3 proteins. These proteins are recruited to the receptor complex by the anchoring protein SARA. Smad2 and Smad3 phosphorylation induces the expression of Smad7 (among others), which inhibits the receptor activity in a negative feedback loop. **(b)** The table shows the names of the pathway proteins in different organisms, with the arrows here indicating which names influenced the creation of others. TGFβ stands for 'transforming growth factor β′, a confusing description, because the response to TGFβ strongly depends on cellular conditions: initially discovered as a growth factor that enhances cell transformation (as the name suggests), it is now of great interest in cancer biology because it inhibits cell proliferation in epithelial cells. The term 'Smad' stands for 'Sma- and Mad- related protein' because sma-2 or sma-3 and Mad are homologues in *C. elegans* and *Drosophila*. 'Mad' is an abbreviation for 'mothers against Dpp'; this name was apparently chosen because Dpp (for 'decapentaplegic', the *Drosophila* equivalent of TGFβ) controls the activity of Mad. The Smad7 equivalent in *Drosophila* is called 'Dad' as in 'daughters against Dpp'. 'SARA' is short for 'Smad anchor for receptor activation'; note that it is different from the *Drosophila* protein named 'Sarah', a biblical reference to Abraham's wife, because the protein affects female fertility. Note another source of confusion: there are two pathways of the TGFβ family in *C. elegans* (the Dauer pathway and the Sma/Mab pathway); while the Dauer pathway is thought to be the closer relative of the canonical TGFβ pathway in vertebrates, it is the Sma/Mab pathway that contains the sma-2 or sma-3 proteins (from which the name 'Smad' was derived) [49,50]. There are many more examples of whimsical gene names in other pathways. For instance, the *Drosophila* genes *lush* and *cheapdate* reflect that their mutants exhibit a high attraction or increased sensitivity to ethanol [51,52]. Such naming is funny, but confusing when viewed in a genomic context and across organisms.

classes is then a succession of binary classifications for each individual class. Defining the negative class in this way is problematic, however, because the proteins in the negative class are often related to the members of the positive class by some biological process; it is difficult to draw clear boundaries between the classes. Such proteins tend to be predicted as belonging to the positive class by the algorithm. It is then difficult to assess the prediction results: did such predictions uncover interesting biological relationships or are they in fact genuine misclassifications of the algorithm? This cannot be answered without sufficient prior biological knowledge. An example of this situation is given in Figure 3. The difficulty of defining a 'negative' functional class is one of the root causes for the poor performance of machine-learning algorithms in the prediction of protein function. Such problems cannot be addressed by fine-tuning parameters of a given algorithm or looking for better machine-learning algorithms [27•]. The issue must be addressed with a more systematic definition of protein function.

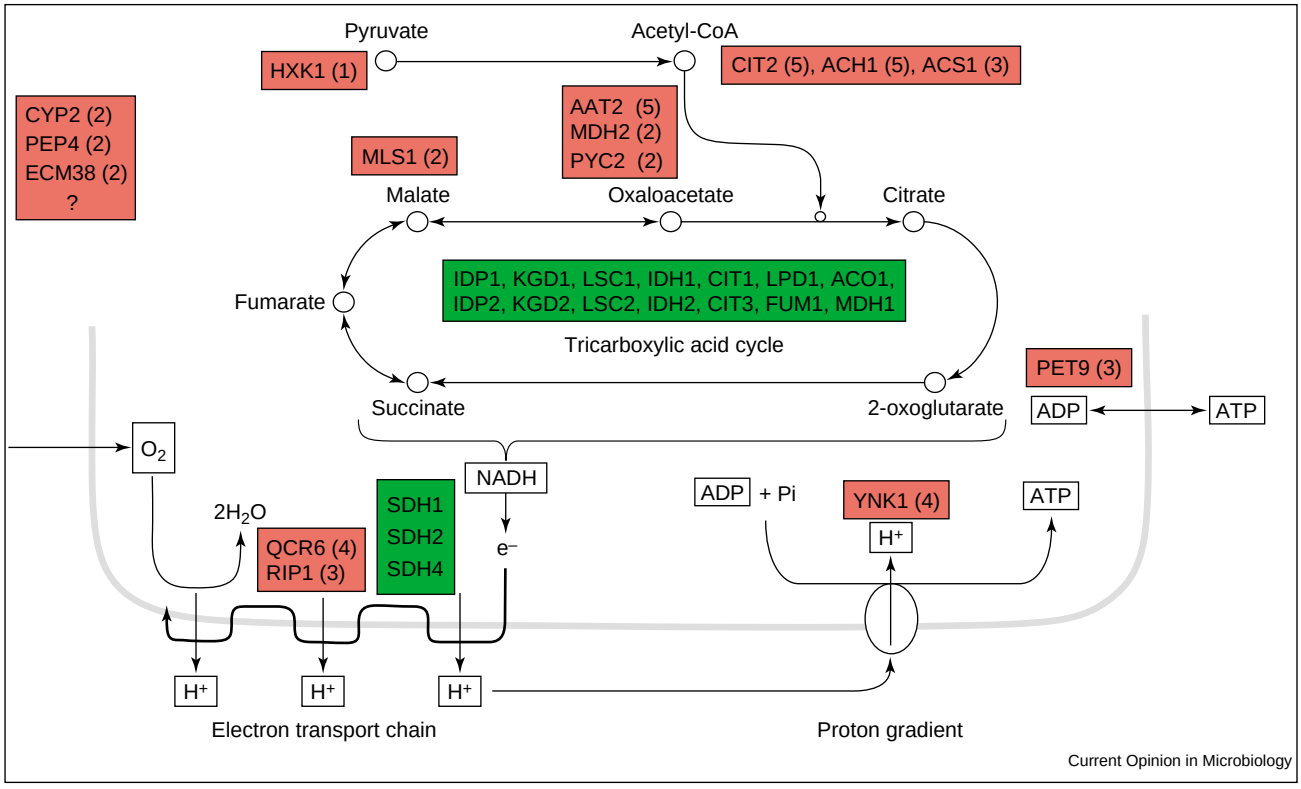
Networks of biological molecules

A network representation of relationships between proteins is potentially a unifying language that can both describe biological concepts of function and provide well-defined frameworks for computational analyses. For instance, molecular interaction networks naturally reflect the fact that many proteins engage in multiple biological processes, and the network distance between two molecular entities often correlates with varying degrees of functional similarity.

In many ways, functional genomics data can naturally be organized and represented in network structures, where the nodes represent molecular entities and the edges (quantifiable) relationships between them. Examples of such networks are protein–protein or genetic interaction networks [4] or networks describing transcriptional regulatory relationships [28,29]. Cellular pathways are in essence networks of interactions between biological molecules (proteins, DNA, RNA, metabolites etc.). Thus, a potentially more systematic way of defining protein function may be gained from analyzing the molecular interaction networks in that proteins are embedded [30,31••,32••,33•] (Figure 4). A complete exploration of such networks with functional genomics methods may lead to a systematic understanding of protein function.

The edges in a network (or the elements in a matrix describing the network) can, in principle, be deterministic or stochastic functions of time and varying cellular conditions, yielding a dynamic rather than static view of biological processes [34,35]. Currently, many of the networks derived from functional genomics data reflect information that is averaged rather than time- or context-specific. In addition, the uncertainty of biological information can be explicitly considered: given one or

Figure 3

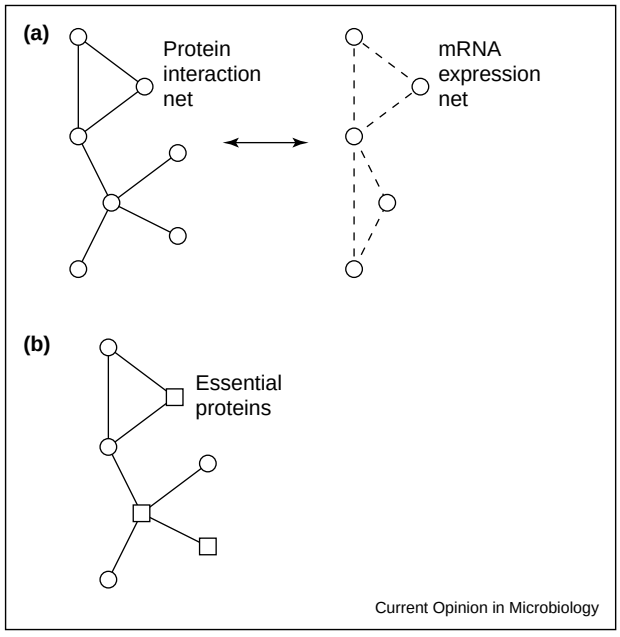


Poor performance of machine-learning algorithms as a result of the fuzzy definition of protein function. The results of a computational prediction of protein function by Mateos *et al*. [27•] are shown. The prediction is based on supervised learning with a neural network (multi-layer perceptron), and the prediction results were tested with threefold cross-validation. Yeast microarray expression data was used as input data. The goal was to predict membership of proteins in the classes of the MIPS functional categorization for yeast proteins [18]. The figure shows the prediction results related to the functional class 'tricarboxylic acid cycle' (TCA cycle) as defined by MIPS. This functional class represents a textbook example of a central metabolic pathway and was used to test the algorithm performance against a well-understood biological standard. The prediction task is essentially a binary classification: the proteins belonging to the class 'TCA cycle' are regarded as positives (that the algorithm is supposed to predict) and proteins outside this class as negatives. The proteins in the green box represent enzymes that the algorithm predicted correctly: they are indeed members of the MIPS class 'TCA cycle'. However, the algorithm also incorrectly predicted several proteins as belonging to the 'TCA cycle' class; these 'false positives' are shown in the red boxes. The prediction was repeated multiple times (with slightly different sets of proteins for training the algorithm), and the numbers in parentheses next to the false positives indicate how many times they occurred. An inspection of these false positives reveals that most of them participate in biological processes that are closely related to the TCA cycle, such as the glyoxylate cycle, oxidative phosphorylation and ATP synthesis. Only three of the false positives cannot be biologically associated with the TCA cycle at all (shown on the left with a question mark). This shows that many of the false positives are 'false' in a statistical sense, but not in a biological sense as they are closely related to the original functional class. The appearance of false positives, rather than being a mistake of the algorithm, results from the organization of the classification scheme that, although curated by human experts, contains rather arbitrary class boundaries around groups of proteins. Similar situations occur for the prediction of protein membership in other functional classes. In cases where we know little about the underlying biological processes, it is nearly impossible to decide whether a 'false positive' represents a mistake of the algorithm or an interesting biological prediction. This makes it difficult to rank the performance of different algorithms. Are more false positives better or worse? This demonstrates the computational problems that can arise from the vague definition of protein 'function'. At the root of these problems is the difficulty of defining when two proteins do not have any function in common or, in other words, what a 'negative' functional class is.

more datasets of evidence (ideally under well-defined cellular conditions), the probability that edges in the network are present or absent can be computed. Such probabilistic networks (with 'weighted' edges) may in the future lead to quantitative characterizations of protein function. In another context, edge weights can represent binding affinities associated with molecular interactions [30].

An important method for computing such probabilistic edge weights is based on comparing functional genomics data against trusted reference datasets ('gold standards'). The gold standards are collections of instances where we assume that the network edges are equal to either 1 ('positives') or 0 ('negatives'). Note that this definition of positive and negative instances in terms of presence or absence of edges is more precise than absence or presence

Figure 4



Analysis of interaction networks. Networks of interactions among proteins can be based on a variety of datasets, such as protein–protein interactions or co-expression relationships. In general, analysis of interaction networks involves two operations: **(a)** the comparison between different nets and **(b)** the statistical analysis of various edge and node properties. Here, the nodes represent proteins, and the edges relationships between them. In (b), an example of the distribution of essential proteins (squares) in the network is shown. A statistical analysis may be aimed at finding out what network properties they are associated with. The information gained from (a) and (b) can often be used for predictions of new protein properties or nets with the help of machine-learning algorithms [53–56].

of proteins in functional classes (as outlined above). Using the example of protein–protein interaction networks, we will discuss these issues in some more detail in the following sections.

Protein–protein interactions are well-defined and amenable to machine-learning approaches

Interactions between proteins represent an important sub-aspect of cellular pathways, and in recent years the biological research community has gathered a large amount of experimental information on them that is stored in publicly accessible databases such as BIND, DIP, MIPS and GRID [18,36–38].

Unlike protein function, protein–protein interactions are relatively clearly defined. Importantly, it is possible to define proteins that do not interact. For instance, inspection of crystal structures of multi-protein complexes allows determining which proteins have physical contacts with each other and which ones do not [21,39]. In a genomic context, it is possible to construct a list of non-interacting protein pairs from information on the preferential subcellular localizations of proteins. Although some proteins may translocate and thus engage in interactions with proteins outside of their primary compartment, subcellular localization measurements turn out to be robust negative controls of interactions in practice (Table 1). Pairs of proteins in different primary compartments are at least highly enriched with non-interacting pairs. Of 8250 interacting protein pairs that can be constructed from the MIPS complexes catalog (often used as a standard for interacting proteins) only 1.5% represent two proteins in different subcellular compartments (based on assigning proteins to four basic subcellular compartments) [40••]. Random shuffling of the subcellular compartments of the proteins in the MIPS complexes catalog increases this percentage to 64%, indicating that the majority of interactions are indeed between proteins in the same primary compartment.

Thus, the conceptual problems of defining protein functional classes (in particular, with respect to a 'negative class') do not exist in the context of protein–protein interactions where positive examples of interacting pro-

Table 1a

Hypothetical subcellular localization measurements with 95% sensitivity and specificity

	Cytoplasm	Nucleus	Sum
Number of proteins measured in each compartment with s = 95%	1400	900	2300
P/N	1.56	0.64	–
PPV = s/(s + (1 − s)N/P)	96.7%	92.4%	–
Number of correct compartments	1354	832	2186
Number of incorrect compartments	46	68	114

Data on the subcellular localization of proteins can be used to construct protein pairs in which the two proteins are in different compartments [5•,6•]. Such pairs are good approximations for proteins that do not interact ('negative' interactions). While screening for protein–protein interactions produces many false positives (see Figure 5), the list of negatives constructed from the localization data are fairly robust against uncertainty in the underlying localization data, as demonstrated in this example calculation. Assume that we measure the subcellular localization of 2300 proteins (with 1400 measurements resulting in 'cytoplasm' and 900 'nucleus') with an experiment that has sensitivity and specificity of $s = 95\%$. For the cytoplasmic measurements, for instance, this results in a *PPV* of 96.7% (with $P/N = 1400/900 = 1.56$) or 1354 correct measurements (see Figures 5 and 6 for definition of PPV).

Table 1b

Percentage of constructed negatives that are actually interacting ('correct').

Number of protein pairs with different measured localization (list of 'constructed negatives')	630 000
Number of correctly 'constructed negatives'	564 827
Number of protein pairs of the list of 'constructed negatives' that are actually interacting	424
Number of 'constructed negatives' that are actually interacting/number of interacting protein pairs	**9.7%**

From these 2300 proteins we can construct 630000 (= 1354 × 832/2) protein pairs where both proteins were measured in different compartments (list of 'constructed negatives'). Of these 630000 pairs 564827 are indeed correctly 'constructed negatives', whereas the remaining ones represent protein pairs that are actually in the same compartment because of the experimental error. How many of these do now correspond to actually interacting protein pairs? Assume that the prior probability of a protein–protein interaction between any random protein pair is 1/600 (meaning that the 2300 proteins form about 4400 interactions among each other). Given the PPV calculated in Table 1a, this results in 424 protein–protein interaction pairs (or 9.7%) being incorrectly assigned to the negatives list. More details of this calculation can be found at http://www.cbio.mskcc.org/~jansen/comb/.

Table 1c

'Correct' constructed negatives as a function of sensitivity and specificity (s = 90%–100%).

s	Number of 'constructed negatives' that are actually interacting/number of interacting protein pairs
100%	0.0%
99%	2.0%
98%	3.9%
97%	5.8%
96%	7.6%
95%	9.4%
94%	11.2%
93%	12.9%
92%	14.5%
91%	16.2%
90%	17.7%

The fraction of constructed negatives that are actually positives (calculated in Tables 1a and 1b) as a function of different values of the sensitivity and specificity parameter *s* in the interval between 90% and 100%. The fraction of falsely constructed negatives never falls below 18%; compare this with the measurements of positive protein–protein interactions, where the fraction of false positives is much higher, namely between 84 and 98% for *s* = 95% (see Figure 6).

teins and negative examples of non-interacting proteins are relatively easily available. Protein–protein interaction networks are therefore well suited for computational analyses and prediction approaches.

The expected relative occurrence of positive and negatives affects misclassification rates of experiments

While protein–protein interactions are clearly defined and are well amenable to computational algorithms, several retroactive analyses of published protein–protein interaction datasets have shown that they contain a high amount of false positives and false negatives [20–23]. Protein–protein interactions are often measured under non-physiological conditions, leading to artifacts in the experimental results. However, it is important to realize that one of the main reasons for the poor accuracy of many protein–protein interaction screens is that the expected number of negatives (non-interacting protein pairs) is several orders of magnitude higher than the number of positives (interacting protein pairs). A consequence of this imbalance is that even experiments with sensitivity and specificity close to 100% may produce a large absolute number of false positive predictions; in this situation, the absolute number of false positives may be larger than that of true positives (Figures 5 and 6) [40••,41].

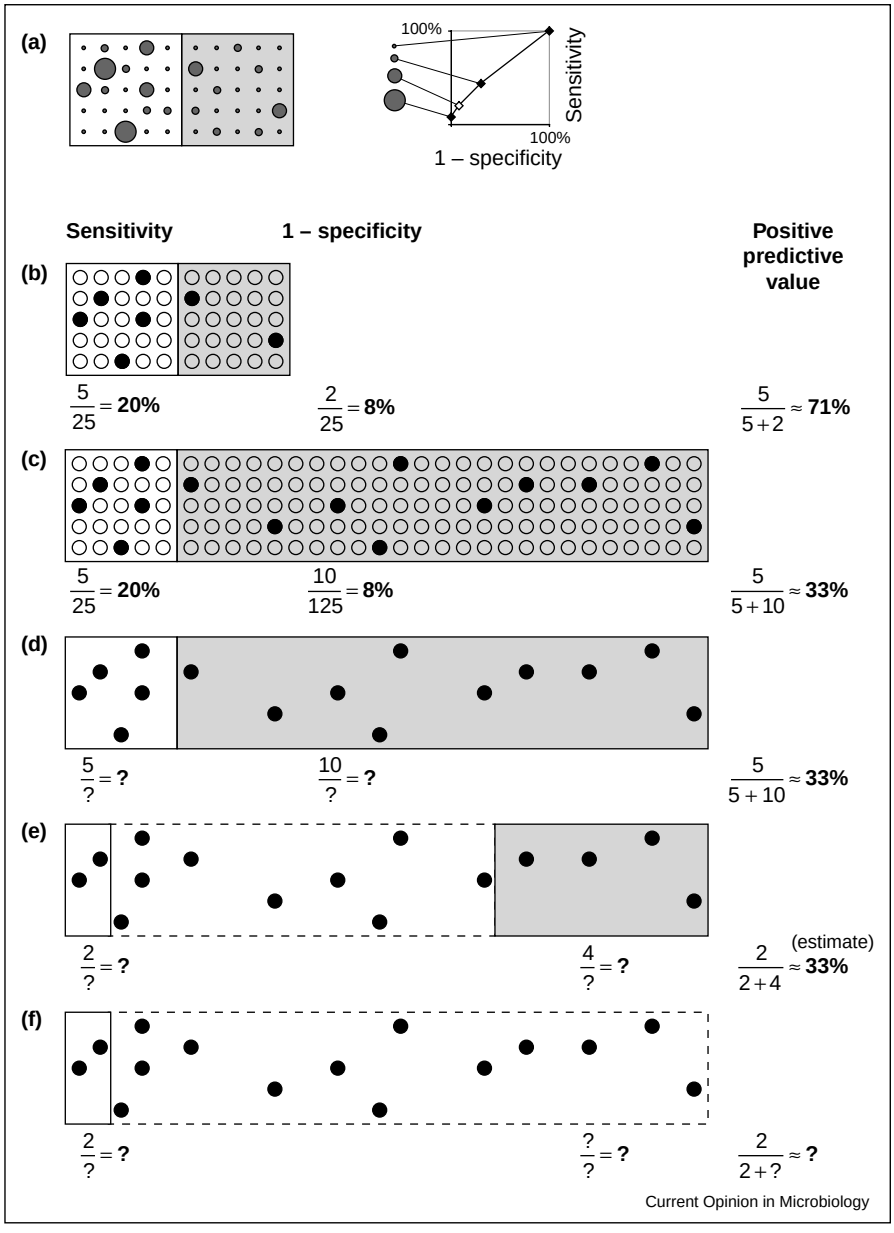
Scalability to higher organisms

The imbalance between a small amount of positives and a large amount of negatives is partially a result of experimental strategies that involve testing pairwise relationships between proteins. For instance, the ~6000 yeast proteins give rise to about 18M distinct protein pairs (and in the case of protein–protein interactions, most of these are negatives). Thus, looking at pairs (or even higher order combinations such as triplets) vastly increases the space of possibilities and often leads to a situation with a very low prior probability of finding a positive among them. This also presents a problem for scaling up screening approaches to higher organisms. For instance, the roughly 30 000 human genes give rise to about 450M pairs — not considering different splice variants and posttranslational modifications that imply an even higher number. It is therefore likely that unbiased screens of pairwise interactions in the human proteome are a lot more prone to false positives than those in smaller model organisms. By contrast, protein–protein interactions studies focusing on smaller-scale systems are usually more reliable; they are often based on prior biological information that makes interactions more likely (Figure 6).

Combining multiple protein–protein interaction datasets

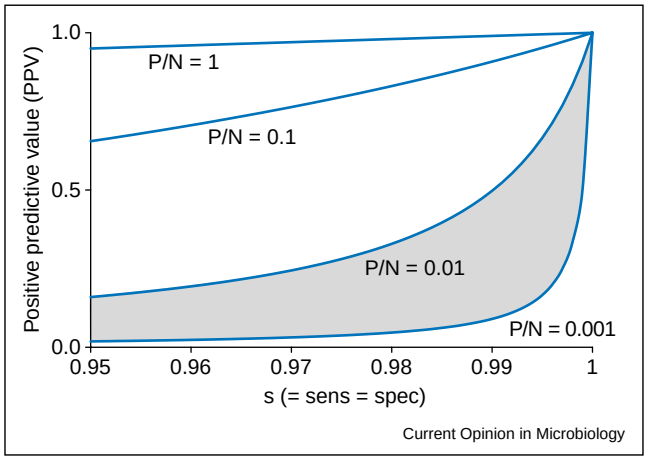
One way to address the issue of the high amount of false positives in the large-scale protein–protein interaction datasets is to analyze the topology of interaction networks for certain network motifs that are associated with more reliable interactions [42,43]. Another approach is the combination of multiple interaction datasets and additional evidence from other genomic data sources that support the existence of protein–protein interactions. Several research groups have developed methods for combining datasets [16,40••,44–48]. Combination of multiple interaction datasets cross-validated against well-defined reference sets of positives and negatives allows estimating probabilities that protein–protein interactions occur (represented as weights on the network edges); the

Figure 5



Positive and negative reference sets are necessary for assessing functional genomics data. **(a)** A hypothetical array experiment (for instance, to detect proteins with kinase activity or to find protein–protein interactions, where the circle sizes represent continuous-valued experimental readouts). The left half represents a set of real positives (that the experiment is designed to detect), whereas the gray colored right half represents a set of real negatives (to which the experiment ideally should be insensitive). A usual first step in the analysis of the experiment is the transformation of the continuous-valued readouts into binary values (positive or negative experimental outcomes) by setting a threshold. When the readout exceeds this threshold, the outcome is classified as positive, and negative otherwise. The level of the threshold fixes sensitivity and specificity of the experiment (see definition below). The ROC curve on the right shows how sensitivity and specificity depend on the threshold (as represented by the varying circle sizes). **(b)** The experimental continuous-valued readout shown in (a) transformed into a binary-valued outcome, where filled circles represent positive and empty circles negative results. The 'sensitivity' is the fraction of real positives that were classified as positive (black dots in the left half); the chosen threshold results in a sensitivity of 20%. '1 – specificity' is defined as the fraction of negatives that were incorrectly classified as positive (black dots in the right half). A third and often most important statistic is the 'positive predictive value' (PPV), defined as the fraction of real positives among the positive experimental outcomes. Note that the number of real positives is equal to the number of real negatives in this example, which is often typical of experiments that are performed on small scale. **(c)** In functional genomics experiments, the number of real negatives often outweighs the number of real positives. For instance, when the same experimental method as in (a) and (b) is applied to a situation where the set of real negatives is five times larger, sensitivity and specificity of the experiment stay the same, but the positive predictive value (PPV) drops from 71% to 33%. In other words, the positive classifications of the experiment are now false more than twice as often as before. **(d)** Many published experimental datasets unfortunately do not contain information

Figure 6



One statistic to characterize experimental methods is their specificity (see definition in Figure 5). However, the positive predictive value (*PPV*) of an experiment is often more informative, especially when the prior probability of finding a positive is very low. Expressed as a function of sensitivity (*sens*) and specificity (*spec*) and the number of real positives (*P*) and negatives (*N*), the *PPV* is:

$$PPV = \frac{P \cdot sens}{P \cdot sens + N \cdot (1 - spec)}$$

If we arbitrarily assume, for demonstration purposes, that sensitivity and specificity are equal, with $s \equiv sens = spec$, we obtain:

$$PPV = \frac{s}{s + (N/P)(1 - s)}$$

The graph shows the *PPV* in a range of relatively high sensitivity and specificity ($s > 95\%$) for different values of *P/N* (a measure of the prior probability). When $P/N = 1$, we obtain $PPV = s$ (that is, the *PPV* is equal to the specificity in this special case). For low values of *P/N*, however, the *PPV* falls off dramatically when *s* is only slightly less than 100%. For instance, the *PPV* is less than 10% when $s = 99\%$ and $P/N = 0.001$. In small-scale studies, the prior probability is often higher than in genome-wide screens of randomly chosen proteins. For instance, a protein–protein interaction experiment may be performed, on a small scale, to find out how the members of a multi-protein complex interact with each other to form a macromolecular structure, or to find out whether proteins in a given pathway have interactions. The chance of finding interactions among such groups of proteins is a lot higher than among randomly picked proteins that may have completely different biological functions. By contrast, the prior probability of finding a positive may be very low in unbiased, genomic screens, implying a low *P/N* ratio. In genomic screens that involve testing pairwise effects between two proteins this problem is even more serious. Typical *P/N* ratios for protein–protein interaction screens in model organisms are between 1/100 to 1/1000 range, resulting in a *PPV* range indicated in gray: small deviations from 100% sensitivity and specificity lead to large drops of *PPV*. In yeast, the ~6000 yeast proteins allow for about 18M potential interactions between them, however, the actual number of interactions is less than 100 000 by various estimates, implying a *P/N* ratio of 1/200 or less. Similar trends can be observed in lethality screens of deletion mutants of single genes or genes in combination. Roughly 17% of yeast proteins (1 out of 6) are lethal when knocked out [3], but of the about 1.3M double knockouts that can be formed from the remaining genes, only about 0.8% (or 100 000) are lethal. Initial studies indicate that the fraction of lethal triplet knockouts is even lower than that [4].

resulting interactomes are probabilistic in nature [40••] (Table 2).

Subcellular localization data as one of the most informative 'protein–protein interaction' datasets

We would like to stress that well-defined reference sets of negatives (non-interacting protein pairs) are just as important as those for positives. Key statistics that characterize an experiment, including specificity and positive predictive value, cannot be computed when a negative reference dataset is not available. In the absence of a negative reference dataset, we simply do not know whether prediction of a previously unknown interaction represents an interesting new discovery (true positive) or a false positive. An interesting point is that the 'negative' interactions that can be derived from the subcellular localization data tend to be more robust against experimental errors than the positive protein–protein interaction data itself [5•,6•] (Table 1). One reason for this is that localization, unlike interactions, is essentially a property of individual proteins rather than protein pairs; this leads to a better balance between the occurrence of positive (proteins in a given compartment, such as 'nucleus') and negative cases (proteins outside of compartment 'nucleus').

(Figure 5 Legend Continued) about negative outcomes (represented by the absence of empty circles). This makes it impossible to determine sensitivity and specificity of the experimental method in a retroactive analysis. For instance, we do not know whether an experiment 'missed' a known positive because the outcome of the experiment was incorrect or whether this instance simply was not covered by the experiment. (Note that many 'genome-wide' experiments practically cover only a large subset of the genome). The PPV, however, can still be computed because it depends only on the correctly or incorrectly *predicted positives*. Therefore it can also be used to rank alternative experimental methods. **(e)** Usually, only subsets or 'reference sets' of the real positives and negatives are known (the reason for conducting the experiment in the first place). This makes it impossible to determine, for all experimental outcomes, whether they are correct or incorrect (area surrounded by dashes). An estimate of the PPV, however, can still be computed from the overlap of the predicted positives with the reference sets. If negative experimental outcomes are reported, sensitivity and specificity can be also computed. Note that the estimated PPV is only close to the actual PPV if the relative occurrence of positives and negatives in the reference sets is similar to that of the real positives and negatives. In practice, we need to estimate what this relative occurrence is. **(f)** When there is no well-defined reference set of negatives, it is impossible to estimate the PPV. We simply do not know whether a positive experimental outcome that does not overlap with the positive reference set is an interesting new discovery (a positive that was previously unknown) or a false positive (an experimental error). A practical example of this situation was shown in Figure 3. Overall, the example illustrates that having a well-defined reference set of real negatives (in addition to having a set of real positives) is essential for assessing experimental methods. Whether an experiment is suitable for genome-wide screens depends not only on sensitivity and specificity of the experiment, but also, and most importantly, on the PPV. The PPV depends on the underlying occurrence of the feature the experimental method is designed to detect or, in other words, the prior probability of finding the feature by chance. This is further quantitatively described in Figure 6.

Table 2a

Likelihood ratio	No. of protein pairs		mRNA expression	MIPS functional similarity	GO biological process similarity
	PIP	Essentiality			
0 – 10	18524971	8130528	18610532	5874302	3125819
10 – 100	214686	0	65268	287503	20467
100 – 1000	25404	0	678	0	0
>= 1000	8067	0	0	0	0

Table 2b

Likelihood ratio	No. of protein pairs				
	PIE	Gavin	Ho	Uetz	Ito
0 – 10	27377	0	0	0	0
10 – 100	30074	31304	25333	0	4393
100 – 1000	2038	0	0	981	0
>= 1000	129	0	0	0	0

Demonstrates the beneficial effects of combining multiple protein–protein interaction datasets and supporting evidence from other genomic datasets. The likelihood ratio *L* (first column) describes how likely a protein–protein interaction is to occur based on the evidence from multiple data sources. If the random odds of finding a protein–protein interaction are given by O_{prior}, then, given the evidence in the data, the posterior odds O_{post} are: $O_{post} = L\, O_{prior}$. The likelihood ratios are essentially computed by comparing the datasets against empirical reference sets of well-defined positive interactions and non-interacting proteins ('negative' interactions). The likelihood ratio increases if the overlap with the positive references is higher and with the negative references smaller. Table 2a shows the results from combining four genomic feature datasets (data on whether proteins are essential or not, mRNA expression correlations and data on the functional similarity between protein pairs in the MIPS functional classification and the GO biological process classification). The four datasets were combined into a 'predicted probabilistic interactome' (PIP). For instance, there are 678 protein pairs in the mRNA expression correlation dataset that have a likelihood ratio between 100 and 1000. However, there are 25404 protein pairs in the same likelihood ratio range in the PIP. Table 2b shows similar results from combining four high-throughput protein–protein interaction datasets into an 'experimental probabilistic interactome' (PIE). Both the PIP and the PIE have many more protein pairs in higher likelihood ratio ranges than the individual datasets [40••].

With this in mind, the subcellular localization data may, in a provocative way, be called the most informative dataset for determining protein-protein interactions. Most experimental interaction screens focus on finding positive interactions. Yet, it would be useful to accompany such experiments with measurements on the localizations of the involved proteins, providing a reliable and necessary negative control. Similar designs, aimed at explicitly measuring negatives, may be useful for other types of functional genomics experiments as well.

Conclusion and perspectives

To fully leverage the results from functional genomics experiments with computational means, it is necessary to define protein function in a systematic way. The fuzzy concept of protein function is one reason for poor performance of machine-learning algorithms. The representation of functional genomics data in networks of relationships between proteins and other biological molecules is a potential way to address this challenge. In the context of protein–protein interaction networks, for instance, it is possible to clearly define both positive and (equally importantly) negative reference examples. Such 'gold standards' allow the computation of edge weights in the network that represent the probabilities of interactions in a given context. When expanding an experimental method for detecting a protein feature from small-scale systems to large-scale, genome-wide screens, it is necessary to understand that the expected occurrence of the feature may be much different in the genome as a whole than in the small-scale system. A low prior probability of occurrence (which is characteristic for protein–protein interactions, for instance) can result in a high absolute number of false positive results even for experimental methods that are fairly reliable in a situation where positives and negatives are balanced.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Wu LF, Hughes TR, Davierwala AP, Robinson MD, Stoughton R, Altschuler SJ: **Large-scale prediction of *Saccharomyces cerevisiae* gene function using overlapping transcriptional clusters**. *Nat Genet* 2002, **31**:255-265.

2. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS: **Global analysis of protein expression in yeast**. *Nature* 2003, **425**:737-741.

3. Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H *et al.*: **Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis**. *Science* 1999, **285**:901-906.

4. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M *et al.*: **Global mapping of the yeast genetic interaction network**. *Science* 2004, **303**:808-813.

5. • Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, Piccirillo S, Umansky L, Drawid A, Jansen R, Liu Y *et al.*: **Subcellular localization of the yeast proteome**. *Genes Dev* 2002, **16**:707-719.
See annotation [6•]

6. • Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK: **Global analysis of protein localization in budding yeast**. *Nature* 2003, **425**:686-691.
With regard to analysis of interaction networks, these two datasets [5•,6•] of genome-wide subcellular localization of proteins can be used as robust, negative controls for protein–protein interactions.

7. 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.

8. 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.

9. 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.

10. Gavin AC, Bosche 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.

11. 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.

12. Horak CE, Luscombe NM, Qian J, Bertone P, Piccirrillo S, Gerstein M, Snyder M: **Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae***. *Genes Dev* 2002, **16**:3017-3033.

13. Euskirchen G, Royce TE, Bertone P, Martone R, Rinn JL, Nelson FK, Sayward F, Luscombe NM, Miller P, Gerstein M *et al.*: **CREB binds to multiple loci on human chromosome 22**. *Mol Cell Biol* 2004, **24**:3804-3814.

14. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I *et al.*: **Transcriptional regulatory networks in *Saccharomyces cerevisiae***. *Science* 2002, **298**:799-804.

15. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H: **Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data**. *Bioinformatics* 2003, **19**:1636-1643.

16. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D: **A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*)**. *Proc Natl Acad Sci USA* 2003, **100**:8348-8353.

17. Bader GD, Heilbut A, Andrews B, Tyers M, Hughes T, Boone C: **Functional genomics and proteomics: charting a multidimensional map of the yeast cell**. *Trends Cell Biol* 2003, **13**:344-356.

18. Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, Pagel P, Strack N, Stumpflen V *et al.*: **MIPS: analysis and annotation of proteins from whole genomes**. *Nucleic Acids Res* 2004, **32 Database issue**:D41-D44.

19. 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.

20. 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.

21. 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.

22. Deane CM, Salwinski L, Xenarios I, Eisenberg D: **Protein interactions: two methods for assessment of the reliability of high throughput observations**. *Mol Cell Proteomics* 2002, **1**:349-356.

23. Deng M, Sun F, Chen T: **Assessment of the reliability of protein–protein interactions and protein function prediction**. *Pac Symp Biocomput* 2003:140-151.

24. Luscombe NM, Royce TE, Bertone P, Echols N, Horak CE, Chang JT, Snyder M, Gerstein M: **ExpressYourself: A modular platform for processing and visualizing microarray data**. *Nucleic Acids Res* 2003, **31**:3477-3482.

25. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T *et al.*: **Global analysis of protein activities using proteome chips**. *Science* 2001, **293**:2101-2105.

26. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines**. *Proc Natl Acad Sci USA* 2000, **97**:262-267.

27. • Mateos A, Dopazo J, Jansen R, Tu Y, Gerstein M, Stolovitzky G: **Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons**. *Genome Res* 2002, **12**:1703-1715.

This paper discusses various problems that arise for machine-learning algorithms from the fuzzy concept of protein 'function', including the difficulty of defining class boundaries and sharing and intermingling between classes.

28. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA *et al.*: **Computational discovery of gene modules and regulatory networks**. *Nat Biotechnol* 2003, **21**:1337-1342.

29. Stuart JM, Segal E, Koller D, Kim SK: **A gene-coexpression network for global discovery of conserved genetic modules**. *Science* 2003, **302**:249-255.

30. Lan N, Jansen R, Gerstein M: **Towards a systematic definition of protein function that scales to the genome level: defining function in terms of interactions**. *Proc IEEE* 2002, **90**:1848-1858.

31. •• Deng M, Tu Z, Sun F, Chen T: **Mapping Gene Ontology to proteins based on protein–protein interaction data**. *Bioinformatics* 2004, **20**:895-902.

See annotation for [32••]

32. •• Letovsky S, Kasif S: **Predicting protein function from protein–protein interaction data: a probabilistic approach**. *Bioinformatics* 2003, **19 (Suppl 1)**:i197-i204.

Approaches for probabilistically inferring Gene Ontology functional classes for proteins in interaction networks based on Markov random field propagation algorithms are described here [31••,32••].

33. • Vazquez A, Flammini A, Maritan A, Vespignani A: **Global protein function prediction from protein–protein interaction networks**. *Nat Biotechnol* 2003, **21**:697-700.

An algorithm for inferring functional classes for proteins in interaction networks, representing a special case of the approaches by Deng *et al*. [31••] and Letovsky *et al*. [32••] based on minimizing interactions between functional classes.

34. Jansen R, Greenbaum D, Gerstein M: **Relating whole-genome expression data with protein-protein interactions**. *Genome Res* 2002, **11**:37-46.

35. Fraser AG, Marcotte EM: **A probabilistic view of gene function**. *Nat Genet* 2004, **36**:559-564.

36. Bader GD, Betel D, Hogue CW: **BIND: the Biomolecular Interaction Network Database**. *Nucleic Acids Res* 2003, **31**:248-250.

37. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D: **The Database of Interacting Proteins: 2004 update**. *Nucleic Acids Res* 2004, **32 Database issue**:D449-D451.

38. Breitkreuz BJ, Stark C, Tyers M: **The GRID: the General Repository for Interaction Datasets**. *Genome Biol* 2003, **4**:R23.

39. Aloy P, Russell RB: **Interrogating protein interaction networks through structural biology**. *Proc Natl Acad Sci USA* 2002, **30**:5896-5901.

40. •• Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data**. *Science* 2003, **302**:449-453.

This study shows how well-defined, empirical reference sets of interacting and non-interacting proteins can be used to assess and combine protein–protein interaction datasets. Non-interaction datasets are used to predict interactions.

41. Manly KF, Nettleton D, Hwang JT: **Genomics, prior probability, and statistical tests of multiple hypotheses**. *Genome Res* 2004, **14**:997-1001.

42. Bader JS, Chaudhuri A, Rothberg JM, Chant J: **Gaining confidence in high-throughput protein interaction networks**. *Nat Biotechnol* 2004, **22**:78-85.

43. Goldberg DS, Roth FP: **Assessing experimentally derived interactions in a small world**. *Proc Natl Acad Sci USA* 2003, **100**:4372-4376.

44. Iossifov I, Krauthammer M, Friedman C, Hatzivassiloglou V, Bader JS, White KP, Rzhetsky A: **Probabilistic inference of molecular networks from noisy data sources**. *Bioinformatics* 2004, **20**:1205-1213.

45. Zhang LV, Wong SL, King OD, Roth FP: **Predicting co-complexed protein pairs using genomic and proteomic data integration**. *BMC Bioinformatics* 2004, **5**:38.

46. Strong M, Mallick P, Pellegrini M, Thompson MJ, Eisenberg D: **Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach**. *Genome Biol* 2003, **4**:R59.

47. Asthana S, King OD, Gibbons FD, Roth FP: **Predicting protein complex membership using probabilistic network reliability**. *Genome Res* 2004, **14**:1170-1175.

48. Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D: **Prolinks: a database of protein functional linkages derived from coevolution**. *Genome Biol* 2004, **5**:R35.

49. Savage-Dunn C: **Targets of TGF beta-related signaling in *Caenorhabditis elegans***. *Cytokine Growth Factor Rev* 2001, **12**:305-312.

50. Tewari M, Hu PJ, Ahn JS, Ayivi-Guedehoussou N, Vidalain PO, Li S, Milstein S, Armstrong CM, Boxem M, Butler MD *et al.*: **Systematic interactome mapping and genetic perturbation analysis of a *C. elegans* TGF-beta signaling network**. *Mol Cell* 2004, **13**:469-482.

51. Kim MS, Repp A, Smith DP: **LUSH odorant-binding protein mediates chemosensory responses to alcohols in *Drosophila melanogaster***. *Genetics* 1998, **150**:711-721.

52. Moore MS, DeZazzo J, Luk AY, Tully T, Singh CM, Heberlein U: **Ethanol intoxication in *Drosophila*: Genetic and pharmacological evidence for regulation by the cAMP signaling pathway**. *Cell* 1998, **93**:997-1007.

53. Yu H, Greenbaum D, Xin Lu H, Zhu X, Gerstein M: **Genomic analysis of essentiality within protein networks**. *Trends Genet* 2004, **20**:227-231.

54. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M: **Annotation transfer between genomes: protein–protein interologs and protein-DNA regulogs**. *Genome Res* 2004, **14**:1107-1118.

55. Yu H, Zhu X, Greenbaum D, Karro J, Gerstein M: **TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics**. *Nucleic Acids Res* 2004, **32**:328-337.

56. Pereira-Leal JB, Enright AJ, Ouzounis CA: **Detection of functional modules from protein interaction networks**. *Proteins* 2004, **54**:49-57.