Protein–Protein Interactions and Genetic Disease

Jeffrey R Brender, *Radiation Biology Branch, Center for Cancer Research, National Cancer Institute-NIH, Bethesda, Maryland, USA*

Yang Zhang, *Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, USA*

Advanced article

Article Contents

Online posting date: 31st October 2017

The advent of high-throughput experiments to measure protein–protein interactions has created a flood of proteomic information parallel to the influx of data caused by the advent of next-generation sequencing technologies. The creation of whole organism protein interaction maps has opened new avenues of predicting genetic disease. Advances in network science now allow the association of genes with disease directly from the characteristics of the protein map without reference to the characteristics of the gene itself. However, none of these techniques has reached the 'black box' level and require careful consideration of the systematic errors in both the underlying experimental data and the computational methods to give reliable results. Here, we review the main methods to characterise protein interactions *in vitro* and *in vivo*, the methods by which protein networks are constructed and the characteristics of the major protein interaction databases, and the techniques used to predict the functional impact of mutations on protein interaction networks.

Introduction

The completion of the Human Genome Project has provided a tentative list of the estimated 25 000 genes that regulate human biology. This landmark project has raised as many questions as it has answered. One of the most fundamental questions in biology is predicting the impact of an alteration of a protein on the phenotype of an individual. This alteration can take many forms, from a change in expression level, the natural variation of protein sequence due to single nucleotide polymorphisms (SNPs), to a drug that alters enzymatic activity. Despite our increasing knowledge of the mechanisms of individual proteins, this question has remained essentially unanswered because the reductionist approach associated with traditional mechanistic studies ignores the context in which proteins act. Few proteins carry out their tasks in isolation. Instead, the action of each protein is controlled by its interactions with others. A major focus in the post-Genome era has been the shift from the 'who' to the 'how', from cataloging of individual genes to piecing together the networks that control the collective activity of the cell.

eLS subject area: Genetics & Disease

How to cite:
Brender, Jeffrey R and Zhang, Yang (October 2017) Protein–Protein Interactions and Genetic Disease. In: eLS. John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0026856

Deducing the Impact of Disease-associated SNPs on Protein–Protein Interactions

Since nearly every process in a cell is dependent on protein–protein interactions (PPIs), mutations hindering these interactions have severe consequences for the associated cellular function. A central problem in medical genetics and personalised medicine is to predict the effect of SNPs to establish a link between the genetic heritage of a person and their susceptibility to disease. Functionally important residues are more likely to be conserved than nonessential ones. SIFT (Ng and Henikoff, 2001) and PROVEAN (Choi *et al*., 2012) (**Table 1**) use this property to estimate the tolerance of a gene to a SNP by forming a sequence profile from a multiple sequence alignment of related proteins. A SNP occurring below a cut-off probability in the profile is deemed to be deleterious. This simple procedure can detect ∼78% of disease-associated SNPs (Choi *et al*., 2012). While Sequence-based methods such as SIFT and PROVEAN are effective in identifying disease-associated SNPs, they provide little information on how the effect manifests. In the absence of a functional assay, evaluating the effect of a mutation on a protein interaction pair usually involves measuring the effect of the mutation on free energy of binding ($\Delta\Delta G$). This naturally raises the question of how changes in binding thermodynamics correlate with changes in function. There is relatively a strong correlation (∼0.7) between the degree of sequence conservation

Table 1 Computational tools for the prediction of a mutation's effect on PPI binding affinity and function

Prediction of deleterious mutations	
SIFT	Detects SNPs that affect protein function by sequence homology http://sift.jcvi.org/
PROVEAN	Detects SNPs with functional impact by the effect of mutation on the global alignment score http://provean.jcvi.org
Prediction of changes in binding affinity upon mutation	
FoldX	Predicts changes in binding affinity and stability through an empirical, physics-based force field with minimal backbone movement http://foldxsuite.crg.eu
Rosetta	Multipurpose software for docking, protein design, structure prediction and more. Allows more extensive backbone movements than other methods https://www.rosettacommons.org/manuals/archive/rosetta3.4_user_guide/df/dc8/_interface_analyzer_doc.html
ZEMU	Extends the FoldX force field to allow larger backbone movements for multiscale modelling http://simtk.org/projects/zemu
BindProfX	Predicts changes in binding affinity through the sequence profiles of structurally related interfaces http://zhanglab.ccmb.med.umich.edu/BindProfX

at structurally similar sites in protein interfaces and $\Delta\Delta G$ values (Xiong *et al*., 2017). This suggests that most protein–protein interfaces are near optimal with small tolerances for changes in binding affinity, provided we assume that there is a tight relationship between sequence conservation and protein function. More quantitatively, comparison of $\Delta\Delta G$ values from the OMIM (online Mendelian inheritance in man) data set of rare mutations that cause hereditary disorders (Amberger *et al*., 2009) and common SNPs that are presumably neutral from the HAPMAP database suggests that a $\Delta\Delta G$ of 1.5 kcal or higher corresponding to a 12-fold change in binding affinity is sufficient to disrupt function for most proteins (Berliner *et al*., 2014). **See also**: **Protein Interaction and Genetic Disease**; **Impact of Missense Variants on Protein–Protein Interactions**

Experimental measurement of PPI binding affinities

The most direct way of measuring the effect of a mutation on binding thermodynamics is to compare the heat evolved during binding of the mutant protein using titration of the binding partner to the heat using isothermal binding calorimetry (ITC) (Velazquez-Campoy *et al*., 2015). ITC provides information not readily available by other means, such as decomposition of the binding energetics into enthalpic and entropic terms. ITC is a technically demanding technique that is prone to errors if the procedure is not followed correctly. However, it is not actually necessary to directly detect the binding event to measure binding. Any signal that changes linearly in response to a change in the bound fraction can serve as a proxy for protein binding. Surface-based technologies such as those used in the Biacore and Octet Red platforms are particularly useful due to their relative sensitivity and ease at which they are scaled up to immobilised protein arrays. Although the physics of some of these methods is fairly complicated, most are based on a change in the optical properties of light when a protein establishes an interaction with another protein immobilised on a surface. The surface plasmon resonance (SPR) technique uses the change in the refractive index that occurs upon a protein binding to the sensor surface. The SPR effect only manifests when the incident angle of light is close to the resonance angle for coupling between the incident beam and the surface plasmons in the conducting surface of the sensor. This angle is dependent on the thickness of the film, which allows the measurement of binding indirectly by monitoring of the resonant angle. In biolayer interferometry, an optical interference pattern is created whose maximum wavelength shifts as the optical path length changes due to binding to sensor surface. In general, surface-based methods have the advantage of determining the kinetic rates k_{off} and k_{on} by monitoring the disappearance of the signal as the protein dissociates from the surface during a wash cycle after the injection stops (Nikolovska-Coleska, 2015). Knowledge of association rates is helpful when studying interactions that are under kinetic control rather than thermodynamic control, such as when several proteins compete for the same receptor binding site (Zhao and Beckett, 2008). **See also**: **A Biophysical Toolkit for Molecular Interactions**

Computational prediction of PPI binding affinities

Regardless of the method of detection, $\Delta\Delta G$ is measured by comparing the binding of a recombinantly expressed and purified WT protein against a mutant prepared by site-directed mutagenesis. The need for expression, purification and site-directed mutagenesis can be time-consuming in many cases and difficult to scale up to the proteome level. Considerable effort has therefore been devoted towards developing computational methods for this task. Most of these approaches rely on physics-based methods that attempt to faithfully model the interactions determining protein–protein binding affinity on the atomic level. FoldX, one of the most successful methods, models the system using an empirical force field built from the measurements of the transfer energies of amino acids from water to hydrophobic solvents and from protein engineering double-mutant cycles (Schymkowitz *et al*., 2005). Because many of the terms are dependent on the precise distances between atoms, FoldX and other physics-based methods need the model of both wild type (WT) and mutant complexes to be accurate on the atomic scale to be effective. The need for an

accurate structure in these methods poses a problem for many proteins, as the crystal structure of the mutant protein is available for relatively few proteins and in many cases the crystal structure of the WT complex is absent as well. FoldX also uses a fixed backbone approximation with minimal relaxation of the backbone after mutation, making the assumption that the mutation has only a minor impact on the structure. In many cases, this assumption is valid, but in other instances, particularly for the substitution of large amino acids for small ones, the complex undergoes a substantial change in conformation. To accommodate these types of changes, protein design programs such as Rosetta can be used to relax the complex (Kortemme *et al*., 2004). However, this global relaxation can cause the protein structure to drift away from the experimentally verified conformation. Multiscale modelling with local flexibility only around a small region near the mutation can be used to compensate for this drift (Dourado and Flores, 2014).

A major obstacle of such approaches is the need for the reconstruction of the full atomic model for every mutant complex, which both limits the accuracy of the approach (since the position of the side chains is difficult to model) and reduces the computational speed and the range of applications (since rebuilding the full atomic model is generally the most time-consuming step). In addition, using a more exact physical representation of the molecular structure and interactions has proved to be less accurate in many cases than using simpler models due to inherent inaccuracy of each term in the force field. As such, alternative methods have been proposed that use reduced representations of the protein structure that do not require the creation of full atomic models. Structurally similar interfaces are expected to serve similar roles regardless of their evolutionary relationship. The likelihood that the mutated residue is found in a structural profile formed from interfaces that are structurally similar to the target can serve as a surrogate for the effect of mutations on binding affinity (Brender and Zhang, 2015). The structural profile method gives results comparable to, and in many cases superior to, more involved physics-based calculations (Xiong *et al*., 2017; Brender and Zhang, 2015). An additional advantage is that the structural profile method works at the residue level and is not reliant on the detailed atomic structure of the complex. The ability to ignore atomic details is a powerful feature since the structure of only ~6% of protein complexes has been determined experimentally (Szilagyi and Zhang, 2014).

Building a Protein–Protein Interaction Network

All the methods mentioned above measure the biophysical effect that a mutation or potential drug has on an individual PPI. The feedback and redundancy built into the PPI network can reroute the flux through the network around defective areas. To fully understand the mutational effects or the targeted silencing of a protein by a drug, it is necessary to understand the interaction network that it is embedded in.

If some of the interaction partners of the protein are known, the techniques in the previous section can be used to build a local map of the network around the target protein. If none of the partners are known, the search must start with a wide net to capture as many potential interactions as possible. This type of search immediately runs into a problem. There have been estimated to be as many as ~650 000 PPIs in the cell (Stumpf *et al*., 2008). It is impossible to screen such a large number of possible interactions *in vitro* by recombinantly expressing each protein and then testing every individual protein against each other. Rather than expressing and test each protein individually, researchers have turned to high-throughput *in vivo* experiments that allow the simultaneous detection of all interactions on a genomic scale.

The workhorse of protein–protein interaction studies: the yeast two-hybrid assay

The most common method for high-throughput mapping of protein interactions is the yeast two-hybrid (Y2H) assay. The Y2H method detects the physical interaction of two proteins indirectly through the downstream activation of a separate reporter gene (Van Criekinge and Beyaert, 1999). The Y2H screen is based on the observation that the activating and the DNA (deoxyribonucleic acid) binding functions of eukaryotic transcription factors are localised into two spatially distinct protein domains (**Figure 1**). Because the two domains fold independently and associate with each other by noncovalent interactions, the transcription factor can be split into two fragments and still activate transcription when the fragments are brought into physical proximity by binding. **See also**: **Two-Hybrid and Related Systems**

This modularity allows two proteins of interest to be tagged with different domains of a fragmented transcription factor. Binding of the two domains forms a functional transcription unit. Once assembled, the transcription factor can then bind a promoter element upstream of the reporter genes to activate their transcription. The exact nature of the detection depends on the reporter gene selected. Regardless of the detection method, the yeast is modified to be deficient in at least two distinct pathways, one for the 'bait' plasmid containing the DNA binding domain of the transcription factor and one of the binding partners and the other for the 'prey' plasmid containing the activation domain and the other binding partner. The prey plasmid is constructed from a cDNA library of ORFs (open reading frame), allowing the screening of thousands of potential interactions of tagged prey proteins against a single bait. Yeast colonies are then grown sequentially on three separate media: a complete growth media experiment that allows yeast deficient in both pathways to grow to ensure the yeast is viable, a selection media experiment lacking leucine and tryptophan that serves as a positive control to ensure both bait and prey plasmids have been successfully incorporated, and a final reporter media experiment that serves as the readout for the experiment. In an auxotrophic selection experiment, the reporter gene encodes a gene essential for growth under specific conditions. Usually, this gene is the transcription factor for HIS3, which encodes the enzyme catalysing the sixth step in biosynthetic pathway of histidine and allows selection in histidine-deficient media. Alternatively, a reporter gene can be used that causes a physical change in the cell that allows colonies with active transcription factors to be clearly identified, such as LacZ (blue colonies in the presence of X-gal) or green fluorescence protein (glowing colonies, useful

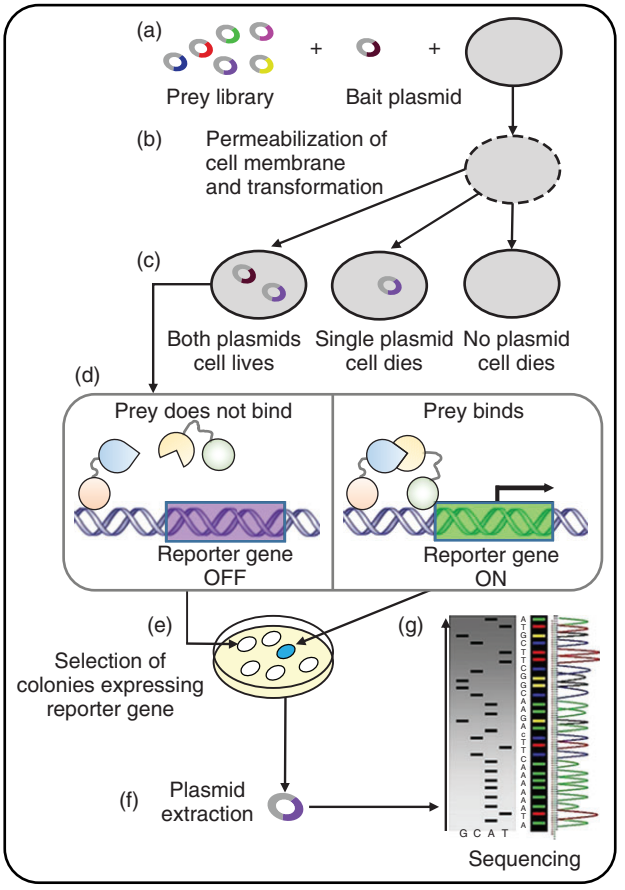

Figure 1 **The yeast two-hybrid assay.** (a) The yeast two-hybrid assay begins with the construction of a prey plasmid library. Each prey plasmid encodes a protein fused to the transcription factor activation domain along with a selection marker to detect the successful incorporation of the plasmid into the cell. A bait plasmid is also constructed encoding the protein of interest fused to the DNA (deoxyribonucleic acid) binding domain of the transcription factor along with a second orthogonal selection marker. (b) The yeast cells are then permeabilised to allow entry of the plasmid and transformation of the yeast genome. (c) Once transformed, the yeasts are grown in media-deficient pathway. (d) Binding of the prey protein to the bait protein brings the activation domain into proximity of the reporter gene and activates transcription. (e) Colonies showing transcription of the reporter gene are selected. (f) The plasmids from the active colonies are extracted and (g) the DNA corresponding to the prey protein sequenced.

for fluorescence-activated cell sorting). In either case, the plasmids of the selected colonies are isolated and sequenced to give the interaction partners of the 'bait' protein.

Limitations of the yeast two-hybrid assay

The Y2H system is found to be widely used because it is fast and requires little hands on time, does not require elaborate instrumentation and is relatively inexpensive in terms of reagents compared with other techniques. The Y2H system also excels at detecting low-abundance protein complexes, as the transcriptional control is under the exogenous reporter gene. For this reason, the coverage of the Y2H assay is higher than most other methods that rely on native protein transcription. Offsetting these advantages are a number of limitations. The Y2H assay is particularly prone to false positives, which some estimates put as high as 40% of the total number of hits (Vidalain *et al*., 2004). False positives in the PPI assays can be divided into 'technical' and 'biological' false positives (Vidalain *et al*., 2004). A biological false positive is a PPI pair that interacts *in vitro* but is not expressed biologically at the same time or place. For transcription to occur, both the prey and bait proteins must be in the nucleus, which is accomplished by the addition of a nuclear localisation signal to the bait and prey protein sequences. Proteins that are in different subcellular compartments and do not interact under normal conditions may find themselves in close proximity due to their artificial localisation in the nucleus.

Technical false positives often arise due to auto-activation of the transcription complex by the bait protein in the absence of the prey protein. About 5% of protein sequences will have sufficiently high transcriptional activity to give a false positive in the absence of prey binding (Walhout and Vidal, 1999). This is not so much a problem when a single protein is used as bait to map out all its interaction partners; however, when the procedure is iterated over a bait library to generate a complete interaction map, the frequency of self-activators can overwhelm the number of true interactions unless precautions are taken to eliminate self-activated bait proteins from the pool (Walhout and Vidal, 1999). Incorrect folding of the bait or prey protein may generate a hydrophobic, sticky interface that attracts many proteins to its surface nonspecifically, generating enough of a transient interaction to activate transcription.

The Y2H assay also generates false negatives. For some proteins, particularly extracellular proteins and membrane proteins (Stynen *et al*., 2012), the protein does not translocate to the nucleus even when a nuclear localisation signal is artificially attached. These proteins will not be detected since the Y2H assay requires nuclear localisation for a functional transcription factor to be made. Finally, the Y2H assay detects binary interactions where the two proteins interact physically through a binding interface. Indirect interactions where the two proteins are part of the same protein complex but are not in physical contact are not detected by the Y2H assay.

Co-affinity purification to capture intact protein complexes

Affinity purification was developed in part to overcome these disadvantages (Gingras *et al*., 2007). In the most common approach, tandem affinity purification (TAP), the target protein is tagged with protein A, which has strong affinity for IgG antibodies. When the IgG antibodies are immobilised on a bead, the high affinity of protein A for the antibody results in the target protein being immobilised as well. All proteins that form complexes with the target protein will also be attracted to the bead and immobilised whereas nonbound proteins are washed away. Cleavage at the protein A site exposes a second tag that is used for a second round of purification to eliminate proteins that nonspecifically bind to the column. The associated proteins are then digested into peptides by proteases that can then be sequenced by mass spectroscopy.

Affinity purification is complementary in many respects to the Y2H assay. Although the Y2H assay only detects binary protein interactions in which each partner is in physical contact, affinity purification also detects proteins indirectly associated with the target through membership in a common protein complex. Affinity purification has a number of disadvantages that balance these advantages (Gingras *et al*., 2007). Affinity purification is an *in vitro* technique using cellular lysates. All cellular localisation is lost during the lysing process, which can give rise to false biological positives, as proteins that are normally segregated in different cellular compartments are artificially brought together. TAP is normally done with an endogenous promoter (there is a TAP variant, high-throughput mass spectrometric protein complex identification (HMS-PCI), that uses recombinant overexpression) (Ho *et al*., 2002). Using an endogenous promoter has the advantage that expression levels are close to the physiological range but has the disadvantage that complexes from low-abundance proteins will tend to be systematically missed (Ivanic *et al*., 2009). Finally, the considerable time involved in purification means that many transient complexes with high k_{off} rates are not detected as the complex dissociates and is lost in the washing procedure (Gingras *et al*., 2007). **See also**: **Tandem Affinity Purification (TAP) Tags**

Increasing accuracy by combining multiple lines of evidence

The complementary nature of affinity purification and Y2H suggests that the results from each assay can be combined to form a more accurate PPI network with potentially wider coverage (**Figure 2**). Using the strict intersection of Y2H and TAP positives significantly reduces the false-positive rate at the expense of a larger percentage of false negatives. This concept of increasing accuracy by combining the predictions of orthogonal techniques can be expanded beyond combining experimental measurements of physical interaction to other indirect measures pointing to the existence of two interacting proteins. Most of these methods are relatively of high accuracy but low coverage: they find few positives in comparison with the high-throughput experimental methods, but the ones that they do find are usually of high confidence.

See also: **Interaction Networks of Proteins**; **Primer on Protein – Protein Interaction Maps**

- *Literature curation***.** The easiest way to find additional interactions is to comb the literature for previously published results. Literature curation can take either of two forms. One method uses expert human curators to search the literature for PPIs, typically looking for high accuracy, low-throughput experiments such as ITC. The other method uses automatic text mining algorithms to infer associations based on statistically significant co-occurrences of gene names from natural language processing of PUBMED abstracts (Papanikolaou *et al*., 2015). Human curation is generally of high accuracy/low coverage, whereas text mining gives many hits of low confidence.
- *Coevolution***.** Random mutations in the protein–protein interface are expected to decrease binding affinity. In many cases, the gradual loss of affinity due to random mutations can be overcome by a compensating mutation at the other side of the interface. The evolution of the two interacting proteins will then be tied together, each protein evolving in response to the other. The *Mirrortree* method constructs phylogenetic trees from the multiple sequence alignment of each protein and calculates the distance between protein sequences within each tree. A high correlation between the distance matrices is a sign of coevolution and marks a likely PPI (Pazos and Valencia, 2001). The method has low coverage but high accuracy. The results are almost completely orthogonal to Y2H and TAP predictions due to its fundamentally different basis (Juan *et al*., 2008).
- *Phylogenetic profiling***.** A PPI cannot be established if one of the partners does not exist. One of the earliest methods for bioinformatic detection of PPIs used the simultaneous appearance and disappearance of protein pairs during evolution as a probe for PPIs (Pellegrini *et al*., 1999). A high correlation across genomes is indicative of a PPI.
- *Orthologous transfer***.** Highly related sequences derived from a common ancestor (orthologs) are likely to participate in the same set of interactions. If a PPI has been identified in the PPI network of one species, it can be transferred to another species with high accuracy if the sequence identity is 80% or higher (Yu *et al*., 2004).

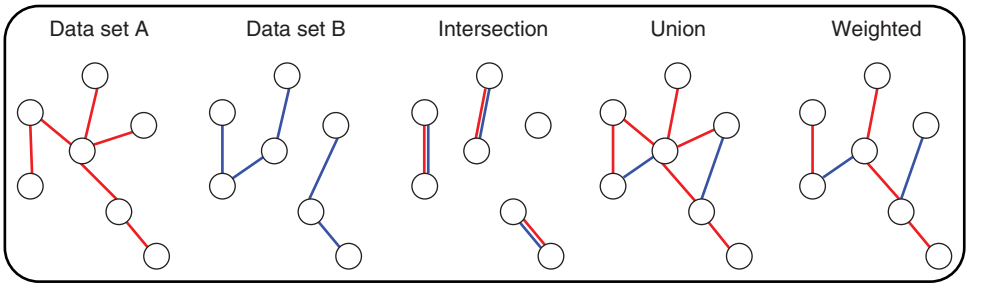

Figure 2 **Integration of different data sets into a protein–protein network.** Accuracy can be increased by considering only the strict intersection of the data sets where a positive PPI (protein–protein interaction) exists in each data set. Alternatively, the coverage can be extended by considering the union of the data sets where a PPI is considered to exist if it is found in either data set. Weighted integration counts only the PPIs in each data set considered to be the most reliable through the consultation of an outside gold standard database.

- *Gene coexpression*. Proteins that participate in a PPI are likely to be needed simultaneously by the cell, which will lead to a correlation in expression levels if the dominant factor for expression is cellular demand for that specific complex (Fraser *et al*., 2004).
- *Synthetic lethality*. An interaction can sometimes be inferred when two genes, which when mutated separately have minimal impact on the phenotype, are lethal when mutated together. Mutations of this type infer a functional relationship between the two genes, although not necessarily a physical interaction. The two genes can be either on alternate but redundant biochemical pathways or are redundant interaction partners in a protein complex with a third protein (Talavera *et al*., 2013).
- *Inverse docking*. The surfaces of interacting proteins are usually complementary in both shape and composition. If the structure of each protein is known individually, the surface complementarity of interacting partners can be used to detect PPIs through molecular docking (Wass *et al*., 2011; Zhang *et al*., 2012). In a benchmark test, 34% of complexes could be distinguished at the 95% confidence level by rigid body docking and 64% at the 80% confidence level (Wass *et al*., 2011). Rigid body docking has the drawback that the structure of the monomers making up the complex must be known beforehand, which is only true in about 30% of cases (Szilagyi and Zhang, 2014). Rigid body docking is also less efficient for flexible proteins that undergo conformational changes upon binding. As an alternative, it is possible to search through the structural library of known complexes for related sequences, based on the well-tested assumption that similar sequences will usually generate similar structures because the number of likely conformations is limited. PrePPI, PRISM and SPRING check each monomer for a match to one of the subunits in structural library of complexes (**Table 2**). If a hit is obtained, there is a high likelihood that a PPI exists between the two proteins.

Constructing a PPI network using multiple sources of evidence requires some method of integrating the different data types. The method of integration can have a significant impact on the final model (**Figure 2**). Accuracy can be maximized by requiring the same interaction to be detected in all data sets for the PPI to be considered to exist. Alternatively, coverage can be increased by considering a PPI to exist if it is present in any of the data sets. Neither method considers the difference in accuracy among different experiments or the bias each experiment has towards some types of proteins. Accounting for these differences requires an additional source of information. Intuitively, another PPI map where all the interactions within the network are known with high reliability (a gold standard training set) can give us information about the relative reliability of a technique for a specific interaction. Formally, we seek the conditional probability that an interaction exists given a set of experimental observations ($x_1, x_2, \ldots, x_n$), $p(\text{TRUE}|\, x_1, x_2, \ldots, x_n)$, where the observations can be either categorical or numeric variables. This conditional probability cannot be inferred directly from the training data. Indirectly, it can be calculated easily by Bayes theorem using the fraction of positive interactions among all those tested in the training set, $p(\text{TRUE})$, and the fraction of experiments with the experimental value x_i among the proteins known to interact within the training set, $p(x_i|\,\text{TRUE})$ and those reliably known to not interact $p(x_i|\,\text{FALSE})$ (Jansen *et al*., 2003):

$$\frac{p(\text{TRUE}\, x_1, x_2, \ldots, x_n)}{p(\text{FALSE}\, x_1, x_2, \ldots, x_n)} = \frac{p(\text{TRUE}) \prod_{i=1}^{n} p(x_i\,|\,\text{TRUE})}{p(\text{FALSE}) \prod_{i=1}^{n} p(x_i\,|\,\text{FALSE})} \quad (1)$$

The naïve Bayes method is reliable when the data from each experiment is statistically independent from each other and high-quality positive and negative interaction training sets exists. If some of the data from different experiments share the same systematic errors, the redundancy leads to an inappropriately high weighting of the correlated data sets relative to others since they are effectively counted twice. The naïve Bayes method also requires a reliable source of true positives and true negatives for each technique being integrated. Obtaining the negative set of protein pairs reliably known to not interact $p(x_i|\,\text{FALSE})$ poses a problem (Jansen and Gerstein, 2004), as negative examples are rarely recorded in the literature. One method is to use proteins with different subcellular localisation, as proteins in different cellular compartments cannot interact (Jansen *et al*., 2003; Jansen and Gerstein, 2004). This method has the drawback that it introduces biases into the integration process since different

Table 2 Computational tools for the prediction of PPIs by sequence or structure

SPRING	Detects PPIs by the similarity of the target sequence to one of the monomers in library of protein complexes. The sequence similarity is calculated by threading, which besides sequence identity, also takes into account the site-specific structural and chemical environment http://zhanglab.ccmb.med.umich.edu/spring
COTH	Similar to SPRING, except the sequence of both monomers is threaded simultaneously through a structural library. Useful when the target protein undergoes a large conformational change upon binding http://zhanglab.ccmb.med.umich.edu/COTH
PRISM	Detects PPIs through interface similarity. Requires the structure of the monomers to be known or predicted beforehand http://cosbi.ku.edu.tr/prism
PrePPI	Similar to SPRING, except sequence similarity is used instead of threading. The inverse docking score is combined with five other data sources through a Bayesian classifier to give a consensus score. Also contains a queryable database similar to STRING http://honig.c2b2.columbia.edu/preppi

Table 3 Popular protein–protein interaction databases (statistics are as of May 2017)

Database	Feature	Proteins	Interactions	Species
HPRD[a]	Physical interactions between human proteins	30 047	41 327	1
IntAct[b]	Database of experimental evidence for physical interactions between proteins	98 289	720 711	7
HINT[c]	Database of high-confidence results for physical interactions between proteins	NA	387 615	12
MIPS[d]	Manually curated database of high accuracy, manually performed experiments	982	1859	3
BIOGRID[e]	Database of experimental evidence of both physical and functional associations among proteins	65 958	1 137 230	62
STRING[f]	BIOGRID and IntAct entries along with computational predictions and orthologous transfer from other organisms. Confidence scores are based on a Bayesian network	9 643 763	1 380 838 440	2031

[a]HPRD: http://www.hprd.org/
[b]IntAct: http://www.ebi.ac.uk/intact/
[c]HINT: http://hint.yulab.org
[d]MIPS: http://mips.helmholtz-muenchen.de/proj/ppi/
[e]BIOGRID: https://thebiogrid.org/
[f]STRING: https://string-db.org/

experimental techniques have different sensitivities to proteins in different cellular compartments.

Protein–protein interaction databases

There are several databases that aim to collect the results of PPI screening from the literature (**Table 3**). The databases mainly differ in the lines of evidence used to construct the database, with the main division between those that define an 'interaction' strictly as physical binding between proteins and others that define a PPI more loosely as being any functional association between proteins, for example proteins that share a common substrate. Among the former, the IntAcT (Hermjakob *et al*., 2004) and Human Protein Reference Databases (HPRD) (Peri *et al*., 2003) consider only binding interactions of proteins verified either through direct detection of a binary interaction or verified comembership in a protein complex. The HINT (*h*igh-quality *int*eractomes) database is even more strict in only considering interactions that have been verified by two orthogonal assays (for high-throughput assays) or two separate publications for manually performed experiments (Das and Yu, 2012). Finally, the Mammalian Protein–Protein Interaction Database (MIPS) focuses exclusively on high accuracy, manually performed experiments (Pagel *et al*., 2005). Because these databases are considered to have fewer false positives than those that consider genetic information as inference for a physical interaction, they often serve as the 'gold standard' for accuracy for testing new PPI prediction methods.

The second set of databases is more inclusive in the definition of a PPI. In addition to physical interactions, the BIOGRID database also includes genetic evidence in the form of synthetic lethality experiments as evidence for a PPI (Chatr-aryamontri *et al*., 2017). Because mutations of different proteins in congruent pathways give a positive result in this experiment, synthetic lethality experiments are evidence for a functional association but not necessarily a physical interaction. The STRING database is the most inclusive of all the above. In addition to all the interactions in the BIOGRID and IntAcT databases, STRING also includes predictions from the orthologous transfer of interactions across species as well as predictions of functional association from coexpression, text mining of PUBMED abstracts and phylogenetic profiling (Szklarczyk *et al*., 2017). STRING combines these data sources to give a confidence score based on a Bayesian network (**Figure 3**). Regardless of the database, successful interpretation of a PPI map requires careful attention to data sources and a solid understanding of the underlying methodology.

Predicting Disease-associated Genes through Network Topology

Since the condition of the cell is ultimately regulated by the state of the cellular interaction network and not by the action of any single protein, identifying key control points has been an important goal in drug discovery. Identifying the network of PPI partners is not enough to predict the effect of disrupting one of the nodes of the network. Evolution has evolved multiple levels of redundancy and feedback mechanisms that divert the flux along alternate pathways in case of loss of one of the network nodes. In principle, the effect of a removal of node can be determined by solving all the relevant differential equations that describe the biochemical flux through that node. In practice, uncertainties in the measurements place limitations on the reliability of the constructed model. This is especially true for protein interaction networks constructed from the Y2H assay, which yields qualitative rather than quantitative information. **See also**: **Protein-Protein Interactions: The Structural Foundation of Life Complexity**

An alternative approach is to avoid considerations of the detailed flux between proteins and instead concentrate on the

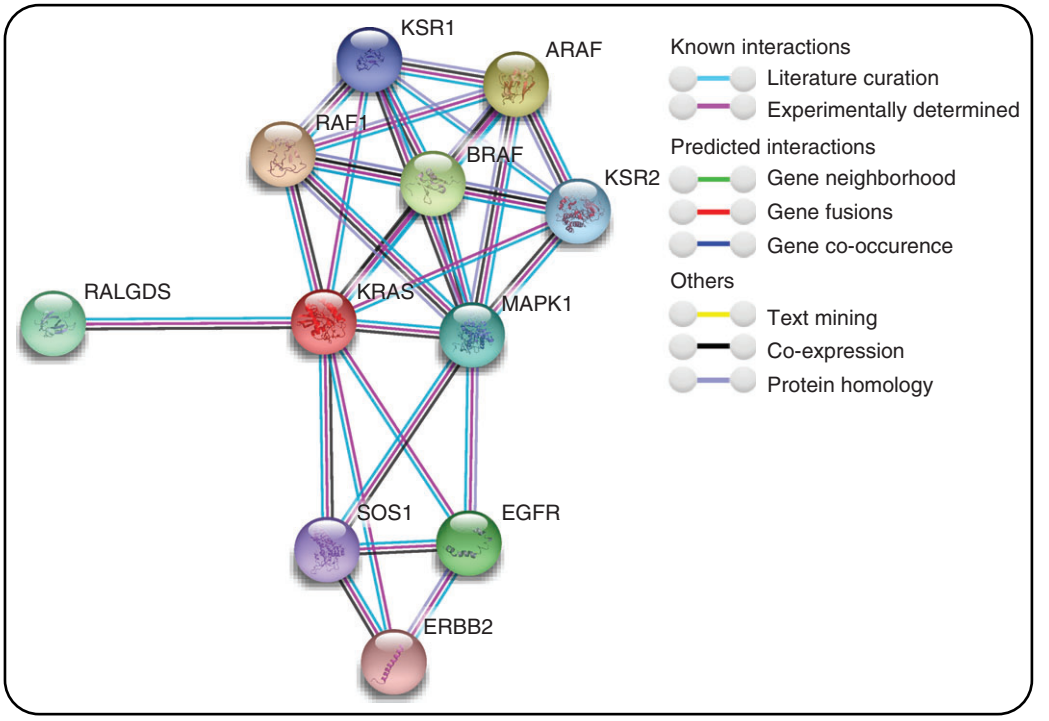

Figure 3 Modified screenshot of an example query from the STRING database. Example of a protein network from the STRING database using the KRas protein, an oncogene implicated in the development of many cancers. The colour of the lines connecting the protein nodes indicates the particular lines of evidence used in establishing a functional association whereas the distance between the nodes is a measure of the confidence of the interaction as established by the Bayesian scoring system. Predicted GO pathways are also available (not shown).

manner in which the proteins are connected with another; focusing on the overall topology of the network instead of the details of the interactions between individual nodes. To quantitatively define a network topology, we must first define some basic vocabulary from the abstract theory of networks and graphs. Following the conventions of mathematical graph theory, proteins in a network are called *nodes* and the interactions between protein are called *edges*. The most fundamental property of an individual node in a graph is its *connectivity* or *degree*, which is simply defined as the number of direct connections (or edges) it makes to other nodes in the network. The *degree distribution* of a network is the probability distribution of finding a node with a degree of exactly k.

Consideration of the path length between nodes gives us another set of higher order measurements. The typical separation between two nodes in the network can be measured by considering the average number of steps along the shortest paths for all possible pairs of network nodes. This gives us the *characteristic* or *average path length* of the network. The distance between the two nodes with the longest minimum path distance is the *network diameter*. All the above are measures of the global network topology. To measure local network properties, two more measures of centrality and cliquishness, or the tendency of linked nodes in a network to group together and share similar links with other nodes, can be defined. Cliquishness can be defined by the number of connections or edges between node neighbours divided by the theoretical maximum of these connections: $k(k-1)/2$. The *clustering coefficient* is the average of this number over all nodes in the network. The final metric measures the centrality or importance of nodes, following the assumption that an important node will lie on a high proportion of paths between other nodes in the network. A centrality measure, betweenness, can be defined as $BC(n) = \sum_{s \neq t \neq n} \frac{\sigma(s,tn)}{\sigma(s,t)}$, where $\sigma(s,t)$ is the number of shortest paths from node s to node t to measure the traffic through a particular PPI.

One of the goals of constructing a PPI network is to enable predictions that cannot be performed by looking at individual interactions alone. Morbidity prediction or the identification of genes associated with a specific, usually hereditary, disease is one of the most basic tasks in bioinformatics. It is also where the network-based approach has shown the greatest success. Although approximately 10% of known genes have some level of association with any disease (Amberger *et al*., 2009), the probability of a given gene being associated with a specific disease is much lower. The most conceptually simple method for network-based morbidity prediction relies on the transfer of gene function annotations to nearby nodes in the network. If a dysfunction in a specific gene is known to cause a defect, there is a high probability that a defect in genes within the same pathway will produce a similar dysfunction. This simple direct neighbour approach is capable of enriching the chances of successful morbid gene prediction 10-fold compared with positional genetic linkage analysis alone (Oti *et al*., 2006). The 10-fold enrichment

factor can be increased to ~25-fold by replacing direct neighbour linkage by a proximity metric based on the diffusion time from a random walk to all genes known to be associated with that disease (Kohler *et al*., 2008).

Measures of the local network topology can provide additional complementary information on the prediction of morbid genes, although this is currently controversial. Disease-associated genes are possibly slightly more likely to be network hubs (have a high node degree k) than other genes (Tu *et al*., 2006), although this effect is skewed towards cancer genes, which tend to affect the highly regulated growth and survival networks. The higher node degree of morbid genes is most evident when literature sources are used to construct the network, which suggests that a publication bias towards disease-associated genes may be partly responsible for the effect and some studies find that no significant difference in node degree exists between morbid and non-morbid genes.

Drug Discovery Targeting Protein–Protein Interactions

One of the ultimate goals of researching PPIs is to use the knowledge of the PPI network to improve drug discovery. Modern drug design follows the 'magic bullet' hypothesis: a disease can be controlled by the alteration of the activity of a single protein (Nolan, 2007). The success of this theory depends on the degree to which the cell compensates for the loss or alteration of any single node or edge within the network. A robust network or functional module is tolerant to such changes whereas a fragile one is sensitive to small perturbations. The robustness of the network is dependent on the underlying architecture. Due to the influence of mutation, biological networks have evolved to be tolerant to the removal of random nodes. This tolerance is manifested in the degree distribution of the network. The node degree k in PPI networks follow, at least approximately, a power law distribution known as scale free: $P(k) \propto k^{-\gamma}$ where the scaling constant γ is between 2 and 3. While the power law distribution means that many proteins within the network have few connections to other proteins and are therefore less likely to be fatal if mutated, it also means that the loss of any of the few hub proteins with high node degree will be catastrophic for the cell (Albert *et al*., 2000). The existence of such hub proteins has implications for drug design. For drugs meant for infectious diseases (Raman *et al*., 2008) and cancer (Mitsopoulos *et al*., 2015) where the goal is to destroy the cell, hub proteins with high betweenness measures are an obvious choice, provided that such sites can be differentially targeted from normal human cells. In particular, network articulation points where removal of the node severs the PPI network into separate graphs are ~twofold overrepresented in cancer drug targets (p-value = 0.0003) (Mitsopoulos *et al*., 2015).

Targeting a PPI assumes a drug can be created that blocks the PPI. This assumption is often not valid with current drug design techniques. Most protein–protein interfaces are large, flat and relatively featureless compared with the binding pockets of the receptors and ion channels that comprise the majority of most drug targets today (**Figure 4**). The size of the PPI surface is an obstacle for the development of small-molecule drugs against PPIs, as small-molecule drugs by definition can cover only a limited area of the interaction surface. This is compounded by the fact that PPI surfaces are sometimes disjoint with the interaction occurring in two spatially distinct regions.

For a long time, PPIs have been considered essentially undruggable by small molecules. PPI have only been targeted indirectly, for example by down-regulating the expression of the interaction partners. This concept of PPIs being undruggable by small molecules is slowly changing with the recognition that, although PPI surfaces appear to be large and featureless, the actual interaction is dominated by a few hot spot residues that contribute most of the binding energy (Zerbe *et al*., 2012; Wells and McClendon, 2007). In some favourable cases, a small molecule can be found that targets these hot spots on the PPI surface and disrupts the protein–protein interface (Wells and McClendon, 2007). Although only one drug specifically targeting PPI interfaces has currently reached approval status (the blood thinner Tirofiban), there are several dozen others in development (Scott *et al*., 2016).

The other option is to simply use larger molecules. Small molecules have been preferred in drug development because

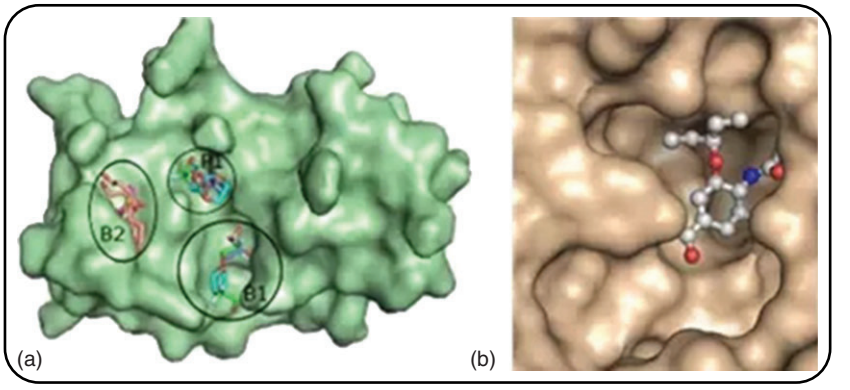

Figure 4 Comparison of PPI interfaces and small molecule binding pockets. (a) A large and shallow PPI with three potential hot spots for small molecule binding. (b) Enzyme binding pocket. Note the smaller size and greater depth of the enzyme binding pocket in comparison with the PPI interface. Adapted from Zerbe (2012) © American Chemical Society.

of their ability to cross the cell membranes, their ease of manufacture and the possibility of oral delivery, but protein-based biologics have steadily gained ground in recent years as the low-lying fruit is exhausted and drug makers reach for more difficult target. Although proteins such as antibodies cannot cross the cell membrane and will generally not reach intracellular targets, peptides large enough to disrupt PPIs can be tagged to penetrate the cell membrane (Johnson *et al*., 2011). Ordinary peptides have the disadvantage that they are easily cleaved in plasma, but techniques exist to extend this half-life. Macrocycles are another potential avenue for exploration (Driggers *et al*., 2008).

Acknowledgements

This work was supported in part by the National Institute of General Medical Sciences (GM083107, GM116960) and the National Science Foundation (DBI1564756).

Glossary

Degree The number of links from a node in a network.

Degree distribution The probability distribution of node degrees within a network.

Druggable A protein that has an active site or interface for which a molecule with high affinity can be designed. If the characteristics of the interface or active site are such that it is not currently possible to create such a molecule, the protein is considered undruggable.

Functional association A relationship involving a mutual influence among two proteins that may or may not involve a physical interaction, for example two proteins that share a common substrate or a common membership in a biological pathway.

Morbid gene A gene known to have some functional association with a disease.

Network node A connection point within a network. For protein–protein interaction networks, proteins serve as network nodes.

Network topology How the nodes are connected within a network.

Scale-free network A network in which the number of links from a node follows a power law distribution. Scale-free networks are robust to random mutations but are more prone to catastrophic failures than networks with a random distribution of the number of links from a node.

References

Albert R, Jeong H and Barabasi AL (2000) Error and attack tolerance of complex networks. *Nature* **406**: 378–382.

Amberger J, Bocchini CA, Scott AF and Hamosh A (2009) McKusick's online mendelian inheritance in man (OMIM (R)). *Nucleic Acids Research* **37**: D793–D796.

Berliner N, Teyra J, Colak R, Lopez SG and Kim PM (2014) Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation. *Plos One* **9**: e107353.

Brender JR and Zhang Y (2015) Predicting the effect of mutations on protein-protein binding interactions through structure-based interface profiles. *PLoS Computational Biology* **11** (10): e1004494.

Chatr-aryamontri A, Oughtred R, Boucher L, *et al.* (2017) The BioGRID interaction database: 2017 update. *Nucleic Acids Research* **45**: D369–D379.

Choi Y, Sims GE, Murphy S, Miller JR and Chan AP (2012) Predicting the functional effect of amino acid substitutions and indels. *Plos One* **7**: e46688.

Das J and Yu H (2012) HINT: high-quality protein interactomes and their applications in understanding human disease. *BMC Systems Biology* **6**: 92.

Dourado DFAR and Flores SC (2014) A multiscale approach to predicting affinity changes in protein-protein interfaces. *Proteins* **82**: 2681–2690.

Driggers EM, Hale SP, Lee J and Terrett NK (2008) The exploration of macrocycles for drug discovery–an underexploited structural class. *Nature Reviews. Drug Discovery* **7**: 608–624.

Fraser HB, Hirsh AE, Wall DP and Eisen MB (2004) Coevolution of gene expression among interacting proteins. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 9033–9038.

Gingras AC, Gstaiger M, Raught B and Aebersold R (2007) Analysis of protein complexes using mass spectrometry. *Nauture Reviews Molecular Cell Biology* **8**: 645–654.

Hermjakob H, Montecchi-Palazzi L, Lewington C, *et al.* (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Research* **32**: D452–D455.

Ho Y, Gruhler A, Heilbut A, *et al.* (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**: 180–183.

Ivanic J, Yu X, Wallqvist A and Reifman J (2009) Influence of protein abundance on high-throughput protein-protein interaction detection. *Plos One* **4**: e5815.

Jansen R, Yu H, Greenbaum D, *et al.* (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**: 449–453.

Jansen R and Gerstein M (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Current Opinion in Microbiology* **7**: 535–545.

Johnson RM, Harrison SD and Maclean D (2011) Therapeutic applications of cell-penetrating peptides. *Methods in Molecular Biology* **683**: 535–551.

Juan D, Pazos F and Valencia A (2008) High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 934–939.

Kohler S, Bauer S, Horn D and Robinson PN (2008) Walking the interactome for prioritization of candidate disease genes. *American Journal of Human Genetics* **82**: 949–958.

Kortemme T, Kim DE and Baker D (2004) Computational alanine scanning of protein-protein interfaces. *Science's STKE* **2004**: pl2.

Mitsopoulos C, Schierz AC, Workman P and Al-Lazikani B (2015) Distinctive behaviors of druggable proteins in cellular networks. *PLoS Computational Biology* **11**: e1004597.

Ng PC and Henikoff S (2001) Predicting deleterious amino acid substitutions. *Genome Research* **11**: 863–874.

Nikolovska-Coleska Z (2015) Studying protein-protein interactions using surface plasmon resonance. *Methods in Molecular Biology* **1278**: 109–138.

Nolan GP (2007) What's wrong with drug screening today. *Nature Chemical Biology* **3**: 187–191.

Oti M, Snel B, Huynen MA and Brunner HG (2006) Predicting disease genes using protein-protein interactions. *Journal of Medical Genetics* **43**: 691–698.

Pagel P, Kovac S, Oesterheld M, *et al.* (2005) The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**: 832–834.

Papanikolaou N, Pavlopoulos GA, Theodosiou T and Iliopoulos I (2015) Protein-protein interaction predictions using text mining methods. *Methods* **74**: 47–53.

Pazos F and Valencia A (2001) Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Engineering* **14**: 609–614.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D and Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 4285–4288.

Peri S, Navarro JD, Amanchy R, *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research* **13**: 2363–2371.

Raman K, Yeturu K and Chandra N (2008) targetTB: a target identification pipeline for Mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. *BMC Systems Biology* **2**: 109.

Schymkowitz J, Borg J, Stricher F, *et al.* (2005) The FoldX web server: an online force field. *Nucleic Acids Research* **33**: W382–W388.

Scott DE, Bayly AR, Abell C and Skidmore J (2016) Small molecules, big targets: drug discovery faces the protein-protein interaction challenge. *Nature Reviews. Drug Discovery* **15**: 533–550.

Stumpf MP, Thorne T, de Silva E, *et al.* (2008) Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 6959–6964.

Stynen B, Tournu H, Tavernier J and Van Dijck P (2012) Diversity in genetic *in vivo* methods for protein-protein interaction studies: from the yeast two-hybrid system to the mammalian split-luciferase system. *Microbiology and Molecular Biology Reviews* **76**: 331–382.

Szilagyi A and Zhang Y (2014) Template-based structure modeling of protein-protein interactions. *Current Opinion in Structural Biology* **24**: 10–23.

Szklarczyk D, Morris JH, Cook H, *et al.* (2017) The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Research* **45**: D362–D368.

Talavera D, Robertson DL and Lovell SC (2013) The role of protein interactions in mediating essentiality and synthetic lethality. *Plos One* **8**: e62866.

Tu ZD, Wang L, Xu M, *et al.* (2006) Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics* **7**: 31.

Van Criekinge W and Beyaert R (1999) Yeast two-hybrid: state of the art. *Biologival Procedures Online* **2**: 1–38.

Velazquez-Campoy A, Leavitt SA and Freire E (2015) Characterization of protein-protein interactions by isothermal titration calorimetry. *Methods in Molecular Biology* **1278**: 183–204.

Vidalain PO, Boxem M, Ge H, Li S and Vidal M (2004) Increasing specificity in high-throughput yeast two-hybrid experiments. *Methods* **32**: 363–370.

Walhout AJ and Vidal M (1999) A genetic strategy to eliminate self-activator baits prior to high-throughput yeast two-hybrid screens. *Genome Research* **9**: 1128–1134.

Wass MN, Fuentes G, Pons C, Pazos F and Valencia A (2011) Towards the prediction of protein interaction partners using physical docking. *Molecular Systems Biology* **7**: 469.

Wells JA and McClendon CL (2007) Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature* **450**: 1001–1009.

Xiong P, Zhang CX, Zheng W and Zhang Y (2017) BindProfX: assessing mutation-induced binding affinity change by protein interface profiles with pseudo-counts. *Journal of Molecular Biology* **429**: 426–434.

Yu HY, Luscombe NM, Lu HX, *et al.* (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Research* **14**: 1107–1118.

Zerbe BS, Hall DR, Vajda S, Whitty A and Kozakov D (2012) Relationship between hot spot residues and ligand binding hot spots in protein-protein interfaces. *Journal of Chemical Information and Modeling* **52**: 2236–2244.

Zhang QC, Petrey D, Deng L, *et al.* (2012) Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**: 556–560.

Zhao H and Beckett D (2008) Kinetic partitioning between alternative protein-protein interactions controls a transcriptional switch. *Journal of Molecular Biology* **380**: 223–236.

Further Reading

Bader GD and Hogue CWV (2002) Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotechnology* **20**: 991–997.

Franzosa E, Linghu B and Xia Y (2009) Computational reconstruction of protein-protein interaction networks: algorithms and issues. *Methods in Molecular Biology* **541**: 89–100.

Goh KI, Cusick ME, Valle D, *et al.* (2007) The human disease network. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 8685–8690.

Jeong H, Mason SP, Barabasi AL and Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* **411**: 41–42.

de Juan D, Pazos F and Valencia A (2013) Emerging methods in protein co-evolution. *Nature Reviews. Genetics* **14**: 249–261.

Klingstrom T and Plewczynski D (2011) Protein-protein interaction and pathway databases, a graphical review. *Briefings in Bioinformatics* **12**: 702–713.

Raman K (2010) Construction and analysis of protein-protein interaction networks. *Automated Experimentation* **2**: 2.

Sanderson CM (2009) The Cartographers toolbox: building bigger and better human protein interaction networks. *Briefings in Functional Genomics & Proteomics* **8**: 1–11.

Snider J, Kotlyar M, Saraon P, *et al.* (2015) Fundamentals of protein interaction network mapping. *Molecular Systems Biology* **11** (12): 848.

Stelling J, Sauer U, Szallasi Z, Doyle FJ and Doyle J (2004) Robustness of cellular functions. *Cell* **118**: 675–685.