View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

FEBS Letters 586 (2012) 2751–2763

FEBS Letters

journal homepage: www.FEBSLetters.org

Review

Domain-mediated protein interaction prediction: From genome to network

Jüri Reimand [a,b,*], Shirley Hui [a,b], Shobhit Jain [a,c], Brian Law [a,c], Gary D. Bader [a,b,c,*]

[a] The Donnelly Centre, University of Toronto, 160 College Street, Toronto, Ontario, Canada M5S 3E1
[b] Department of Molecular Genetics, University of Toronto, 1 King's College Circle, Toronto, Ontario, Canada M5S 1A8
[c] Department of Computer Science, University of Toronto, 10 King's College Circle, Toronto, Ontario, Canada M5S 3G4

ARTICLE INFO

Article history:
Received 25 March 2012
Accepted 17 April 2012
Available online 3 May 2012

Edited by Marius Sudol, Gianni Cesareni, Giulio Superti-Furga and Wilhelm Just

Keywords:
Protein interaction prediction
Network evolution
Peptide recognition module
Functional mutations

ABSTRACT

Protein–protein interactions (PPIs), involved in many biological processes such as cellular signaling, are ultimately encoded in the genome. Solving the problem of predicting protein interactions from the genome sequence will lead to increased understanding of complex networks, evolution and human disease. We can learn the relationship between genomes and networks by focusing on an easily approachable subset of high-resolution protein interactions that are mediated by peptide recognition modules (PRMs) such as PDZ, WW and SH3 domains. This review focuses on computational prediction and analysis of PRM-mediated networks and discusses sequence- and structure-based interaction predictors, techniques and datasets for identifying physiologically relevant PPIs, and interpreting high-resolution interaction networks in the context of evolution and human disease.

© 2012 Federation of European Biochemical Societies. Published by Elsevier B.V.
Open access under CC BY-NC-ND license.

1. Introduction

Eukaryotic signaling and gene regulatory networks are formed by precisely controlled, specific protein–protein and protein–DNA interactions that are ultimately encoded in the genome. Understanding how these networks are encoded in the genome will enable a number of scientific advances. We will be able to predict biologically relevant protein interactions directly from the genome and understand how genomic changes lead to modifications in interaction networks, both over evolution and within a population or individual organism. Further, we will be able to design new networks at the DNA level for synthesis and expression in an organism. Since cellular and physiological phenotypes are the product of molecular interactions, understanding the link from genome to network will help us predict phenotype from genotype.

A number of increasingly powerful experimental techniques provide large datasets of genomes and molecular interaction networks, but many challenges remain before we can accurately relate the two information types. Computational methods can help address a number of these challenges. Primarily, methods can be developed to accurately predict molecular interactions and their binding sites directly from the genome. This will simplify protein interactome mapping and help elucidate many biological processes. Once the binding sites of interactions are known, we can identify how changes at the DNA level affect those sites and thus affect interactions, removing, adding or rewiring them. This review focuses on protein–protein interactions (PPIs), but the ideas can also be applied to other areas where protein domains bind short linear motifs, such as protein–DNA interactions.

In general, predicting protein–protein interactions from the genome is difficult, perhaps comparable to the protein fold prediction problem. Even when three-dimensional protein structures are available, predicting their interactions is challenging, especially if any conformational changes occur upon binding [1]. However, important subclasses of interactions are sufficiently simple to study using established computational techniques, representing a good starting point to study the genome to network relationship. The simplest types of protein–protein interactions are those mediated by peptide recognition modules (PRMs). PRMs are protein domains that, like Lego blocks, recognize short, linear, specific, and characteristic amino acid motifs (peptides) in other proteins. Such domains are reused in different combinations in many proteins with different functions [2–4]. PRMs are suitable for protein–protein interaction mapping as they are widespread in eukaryotic genomes and are relatively easy to detect through sequence similarity to known family members [5–7]. PRMs and their binding partners can be determined with a wide range of high-throughput experimental methods, such as phage display, peptide chips and yeast two-hybrid [8–13]. Furthermore, PRMs are involved in important biological processes including eukaryotic signaling systems and

* Corresponding authors at: The Donnelly Centre, University of Toronto, 160 College Street, Toronto, Ontario, Canada M5S 3E1.
E-mail addresses: Juri.Reimand@utoronto.ca (J. Reimand), Shirley.Hui@utoronto.ca (S. Hui), Shobhit.Jain@utoronto.ca (S. Jain), BM.Law@utoronto.ca (B. Law), Gary.Bader@utoronto.ca (G.D. Bader).

0014-5793 © 2012 Federation of European Biochemical Societies. Published by Elsevier B.V. Open access under CC BY-NC-ND license.
http://dx.doi.org/10.1016/j.febslet.2012.04.027

human disease. Consequently, several types of PRMs are relatively well-studied and many associated datasets are available.

Prediction of PRM-mediated protein interactions from the genome involves multiple steps. Given the preferred binding motif for a PRM, all proteins with the motif are potential binding partners. If the putative binding site is accessible and the proteins are expressed at the same time and place in the cell, the interaction is likely to be physiologically relevant. Thus, mapping the peptide recognition preferences of all PRMs will enable us to predict protein interactions from the genome for a substantial subset of proteins (for instance, over 850 human proteins have PDZ, SH3, SH2, WW, and protein kinase domains according to the HPRD database [14]). The following experimental and computational procedure enables us to predict protein interaction networks from genome sequence. First, a genome is sequenced, all genes are then identified using gene finding software and computationally translated to proteins, followed by detection of PRMs in protein sequences. Second, the binding specificity of each PRM is mapped experimentally or predicted computationally, and the proteome is scanned for potential interaction partners. Third, the resulting protein–protein interactions are scored for physiological relevance using multiple sources of evidence (e.g. co-expression, co-localization). Finally, top candidates are experimentally tested using orthogonal protein interaction mapping methods. The result of this procedure is a high-confidence and high-resolution protein–protein interaction network mediated by PRMs, with information on all PRM-related binding sites [8,10].

This network mapping procedure also enables us to identify genome changes that cause corresponding network changes, an important source of information about the genome–network relationship. Modifying a genome and observing the resulting network changes, via study of binding site changes, would help us identify network-encoding sections of the genome and decipher their meaning. While large-scale experiments of this type are difficult, abundant data about sequence evolution and population variation are increasingly mapped by genome projects. Even though corresponding protein interaction networks are often incomplete, these can be constructed using the network mapping methodology described above and then interpreted in the context of genomic changes. This research will help us understand evolutionary processes and also to predict the functional consequences of inherited and somatic disease-associated mutations. For example, by comparing the genomes of a normal and a diseased individual, we can pinpoint the mutations that cause permutations in signaling pathways or protein complexes that may be involved in the disease phenotype. This would, in turn, provide insights into the genetic basis of specific diseases and new directions for treatment-seeking research.

2. Peptide recognition module mediated interaction networks

An important part of solving the genome-to-network problem involves the development of computational methods to predict protein–protein interactions from the genome. This task is difficult in general, but we can start by focusing on peptide recognition modules. Dozens of PRMs are known, and a few abundant PRMs with well-studied structures provide good starting points for network prediction.

The PDZ domain is one of the simplest PRMs, since it often recognizes the C-terminal tails of other proteins, although other binding modes are known [15]. PDZ domains are 80–90 amino acids in length and fold into a compact structure containing 5–6 beta strands and 2 alpha helices. The binding pocket is formed by the second beta strand, the second alpha helix, and a GLGF loop that preferentially binds a hydrophobic C-terminus. Proteins with PDZ domains are often multi-domain adaptors that regulate ion channels, localize signaling components to the membrane, participate in cell polarity complexes (e.g. the tight junction), and are involved in neural development. The WW domain is another simple PRM with a short 30–40 amino acid length. Its binding mode is more complex than the C-terminal binding mode of the PDZ domain, since it recognizes proline-rich motifs (e.g. PPXY) in any accessible part of a protein [16,17]. WW domains have a three-stranded anti-parallel beta sheet core and generally contain two signature conserved tryptophan residues spaced approximately 20 residues apart. Proteins with WW domains are involved in many signaling pathways, including those in growth control and ubiquitin-mediated proteolysis. As a final PRM example, SH3 domains are larger (~60 amino acids) and more complex than WW domains, with multiple known binding modes, two of which have the domain binding in opposite orientations at the same binding site [18]. SH3 domains are composed of five beta strands organized into two perpendicular beta sheets interrupted by a 3-10 helix. They bind characteristic proline-rich motifs, but can also bind proline-free motifs containing arginine or lysine [10]. SH3 domains are abundant in eukaryotes and are involved in a wide range of cellular processes, including actin cytoskeleton regulation, cell signaling (e.g. receptor tyrosine kinase pathways) and endocytosis.

PRM binding preferences have been mapped using a variety of experimental methods, including phage display and peptide microarray experiments. In phage display experiments, a large-scale combinatorial peptide library is presented to a given PRM on the surface of phage particles, and bound peptides are identified by sequencing the corresponding phage DNA [9,10,19]. To represent an exhaustive set, extremely large libraries (up to approximately 10 billion peptides) can be created containing every possible binding target of up to a particular length; 1.3 billion peptides are needed to cover all seven residues using 20 amino acids. Biased phage libraries can be created by keeping select positions constant to explore longer peptides. In protein microarray experiments, purified domains are immobilised on a solid surface and probed using fluorescently labeled peptides, allowing several hundred domains to be tested for binding against thousands of peptides simultaneously [13,20,21]. For peptide chip experiments, synthesised peptides are displayed on a protein cellulose membrane chip to domains. These experiments are limited to libraries with sizes in the thousands, so are often designed to use only peptides matching a known binding motif for a given domain type [11,12,22,23]. Either the domains or peptides are displayed on the chip, followed by binding of the interaction partners [11–13]. In both peptide chip and protein microarray experiments, binding is generally measured semi-quantitatively, for example using a colorimetric assay. These techniques have been applied to detect interactions involving PRMs such as PDZ, SH2 and SH3 [9,10,18,21].

Domain–peptide interactions from published experiments for PRMs, such as the SH3, WW and PDZ domains, can be found in the Domino and PDZBase databases [24,25]. These interactions can be combined with domain-mediated protein interactions derived using techniques such as yeast two-hybrid to achieve higher confidence protein interaction networks [8,10]. Protein–protein interactions involving various domains from different organisms can be obtained from interaction databases such as iRefIndex, which consolidates PPIs from multiple interaction databases [26]. Combining all of this information with other evidence for physiologically relevant protein interactions, such as co-expression and co-localization, further improves the confidence of a predicted PRM-based protein interaction network.

Overview of machine learning.

Machine learning refers to a family of computational methods that can recognize complex patterns in a given dataset and make decisions on previously unseen data. Binary classification methods can discriminate between objects from two classes using previously available information about those objects. For instance, a predictor may decide if a given domain and peptide pair will physically interact by analyzing the properties of known interacting and non-interacting domain–peptide interactions (e.g. primary, secondary or tertiary protein structural features of interaction participants). Many types of machine learning methods exist, some of which can also perform quantitative, probabilistic, or multi-class predictions.

A machine learning-based predictor is given extensive training data about objects with known classes such as known domain–peptide interactions (positive examples) and non-interacting domain–peptide pairs (negative examples). In a pre-processing step, information about the objects is systematically represented as features. A predictor, such as a support vector machine or Bayesian classifier, is then trained to learn how and which features can best be used to discriminate the objects into their correct classes. The goal of machine learning is then to apply this discrimination process to classify previously unseen objects. Therefore, predictors need to avoid over-fitting, that is capturing features unique to the training data rather than underlying general patterns.

Accurate and generalizable machine learning relies on the use of relevant, diverse, and reliable training data. While positive examples of domain–peptide interactions are often described in the literature, reliable evidence of non-occurring interactions is more difficult to compile. Therefore, various methods to generate reliable artificial negative interactions are employed.

Evaluation of machine learning methods involves estimating a predictor's ability to extrapolate to new data. A basic technique for this is cross-validation, in which the training data is randomly partitioned into several (e.g. ten) subsets of equal size. For each subset, an independent predictor is trained with the remainder of the data, and tested by classifying the data in the subset. The performance of all predictors is then averaged across all runs to achieve a general estimate of performance on 'unseen' data. In the case of domain–peptide interaction prediction, variations of cross-validation may be used. For example, entire sets of interactions involving specific domains or peptides may be held out for testing to reduce the similarity between training and testing data, which would otherwise produce inflated estimates of performance. In addition to cross-validation, blind testing on previously unseen data should be carried out to obtain an unbiased measure of predictor performance.

Using the results from these validation strategies, statistics such as the number of correctly predicted interactions (true positives) and number of predicted positives that are actually negatives (false positives) can be calculated. By varying the predictor's discrimination score threshold (the minimum predicted score that qualifies as a positive classification), a plot of the true positive vs. false positive rates can be made. The area under the receiver operating curve (AUROC) is a single balanced measure that accounts for true and false positives and is commonly used to compare and evaluate machine learning predictors, though other measures exist [27,28]. Multiple other measures also exist to assess different aspects of predictor performance.

2.1. Computational prediction of domain–peptide interactions

The biological importance of PRMs, their recognition of short simple linear motifs, and the availability of experimentally determined interaction data have prompted the development of domain–peptide interaction prediction methods by multiple groups. Such methods are based on established bioinformatics, physical, statistical and machine learning techniques, which have been successfully used to accurately predict interactions for proteins containing PRMs such as the PDZ, SH2, SH3 and protein serine–threonine kinase domains [29–34]. Although the details of these methods vary greatly, such predictors are built by following several general steps that are summarized in Fig. 1. The following discussion focuses on the computational prediction of PDZ domain–peptide interactions; however the methods presented are applicable to other PRMs.

2.2. Sequence-based domain–peptide interaction prediction

A fast and simple method for predicting domain–peptide interactions involves a position weight matrix (PWM) that captures a domain's binding preferences which is used to score a list of potential peptide binders. A PWM is constructed based on a set of verified ligands and is a matrix of the probabilities of observing a particular residue at a given ligand position. PWMs are commonly used to compute a score indicating the binding preference of a domain for a given peptide. Tonikian et al. used PWMs to predict human PDZ interactions and to identify viral proteins that mimicked domain specificities [9]. Stiffler et al. developed a variant of the PWM that contained weights describing the relative preference of a PDZ domain for amino acids at positions in the ligand compared to other domains [21]. The inherent limitation of PWMs is their inability to model dependencies between ligand residue positions. PWMs may also perform poorly when there are too few experimentally determined peptide ligands available for a given protein. Furthermore, the PWM model cannot easily consider additional biological information to help reduce the number of false positives.

More sophisticated prediction methods use machine learning to address the limitations of simple methods such as the PWM. For example, Eo et al. used a support vector machine (SVM) to predict PDZ domain interactions, limited to those involving G-coupled proteins [35]. Chen et al. used a Bayesian model to predict interactions for the entire PDZ domain family using data from a protein microarray experiment [34]. The authors demonstrated their model's ability to predict mouse PDZ domain–peptide interactions and, to a lesser extent, interactions in other organisms. Our group developed a regression framework using positive (quantitative) and negative (qualitative) mouse PDZ domain interaction data to predict PDZ domain–peptide binding affinity [36]. While these methods can predict PDZ domain interactions, their common limitation is that they were trained and validated using limited interaction data for only a subset of PDZ domains. Thus, it is unclear if these can be used to predict interactions for all PDZ domains on a proteome scale. To address this, our group trained an SVM using all the high-throughput PDZ domain–peptide interaction data available at the time for human and mouse. We used the predictor to scan the proteomes of multiple organisms for PDZ-mediated interactions and showed that it outperformed existing state-of-the-art sequence-based predictors for proteome scanning [33].

2.3. Structure-based domain–peptide interaction prediction

Structural features within the domain-binding pocket of a PRM play an important role in determining binding specificity. Therefore the use of domain structure features in training should result in a predictor with improved performance and the ability to

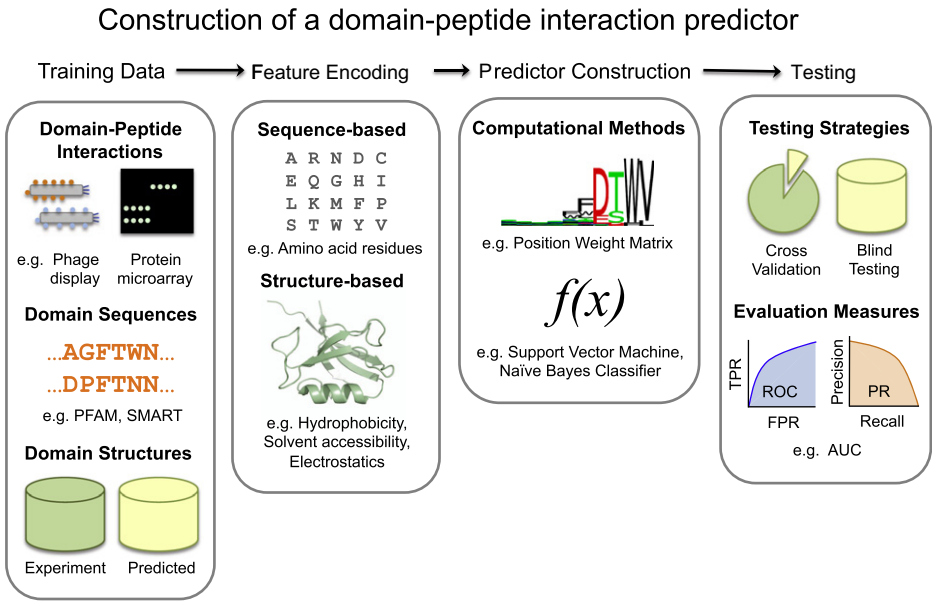

Fig. 1. Construction of a domain–peptide interaction predictor. Training data involve known domain–peptide interaction pairs from experiments, as well as sequence and structure information about domains from databases. Feature encoding is performed in order to transform the data into numeric values that best describe considered information. These features may be sequence-based, describing the amino acids in a given sequence, or structure-based, describing the features involved in protein folding and stability. The predictor is constructed with computational methods, such as a position weight matrix (PWM) or one of many machine learning algorithms. Predictor performance is then evaluated with commonly used metrics (i.e. AUROC scores), different cross-validation strategies and blind testing.

predict new interactions. For instance, Hue et al. used a SVM to predict PPIs using a kernel derived from protein structure information [37]. Other methods using structure information to predict domain–peptide interactions have also been developed. Sanchez et al. used an empirical force field to calculate structure-based energy functions for human SH2 domain interactions [38]. Fernandez-Ballester et al. constructed PWMs of all possible SH3-ligand complexes in yeast using homology modeling [39]. Smith et al. used protein backbone sampling to predict binding specificity for 85 human PDZ domains [40]. Kaufmann et al. developed an optimised energy function to predict the binding specificity of PDZ domain–peptide interactions for 12 PDZ domains [41]. Finally, our group trained an SVM using PDZ domain structure and peptide sequence information. We used the predictor to perform proteome scanning on multiple organisms for hundreds of PDZ domains [42].

2.4. Limitations of domain–peptide interaction prediction methods

Since sequence-based methods are trained using domain and peptide sequence information only, their performance is known to depend on the sequence similarity of a given domain to the domains in the training set. We showed that the ability of our sequence-based predictor to correctly predict interactions for blind test domains decreased as the domain's similarity to the training domains decreased. Furthermore, when the test domain was less than 60% similar, the performance was comparable to a naïve nearest neighbor predictor, whose prediction criteria are based solely on sequence similarity between the domains and the peptides [33]. In other related work, Shao et al. built a sequence-based predictor of PDZ domain–peptide binding affinity. They also observed that the average performance of their predictor depended on how similar a test PDZ domain was to its closest training domain [36]. Thus, sequence-based predictors are in general more likely to correctly predict interactions for domains that are well-represented in the training set in terms of sequence similarity. For structure-based prediction methods, the main challenge is that three-dimensional structures are not available for most domains. However, structures are often available for one or several domain family members. Since PRMs are evolutionarily conserved, they show good sequence similarity. It is therefore possible to use homology modeling and generate reliable structures for many PRMs that lack experimentally determined structures. This will increase the number of structures available for training and testing. For example, on average across human, mouse, worm, and fly, PDZ domains are ~60% pairwise similar in sequence. We used homology modeling to generate 65 PDZ domain structures to train our structure-based SVM domain–peptide predictor. Since homology models may contain inaccuracies, we limited the structural features to the binding site, which is more conserved and therefore more reliably modeled in comparison to regions such as loops. All models had greater than 50% sequence similarity to their template structure (average 90%) [42]. At this threshold, models are expected to have the correct fold with most inaccuracies arising from structural variation in the templates and incorrect reconstruction of loops [43,44].

Most machine learning methods require both positive and negative data for training. As the limited availability of negative interaction data is a problem, collections of random peptides or permuted peptide sequences have been used to represent true negative interactions in training data. However, such approaches resulted in lower predictor accuracy in comparison to training with real negatives [45,46]. Randomly shuffling the interacting partners or pairing partners that are known to be in different cellular compartments are also useful methods for creating negative samples. Unfortunately, these methods create a constraint on the distribution of negatives and artificially influence the predictor to distinguish between positive and negative interactions. This leads to biased estimates of predictor performance when cross-validation is used [45]. Therefore the generation of biologically meaningful artificial negative training data is not fully addressed currently. A recent database archives negative protein–protein interaction data, which may be generally useful for PPI prediction methods in the future [47].

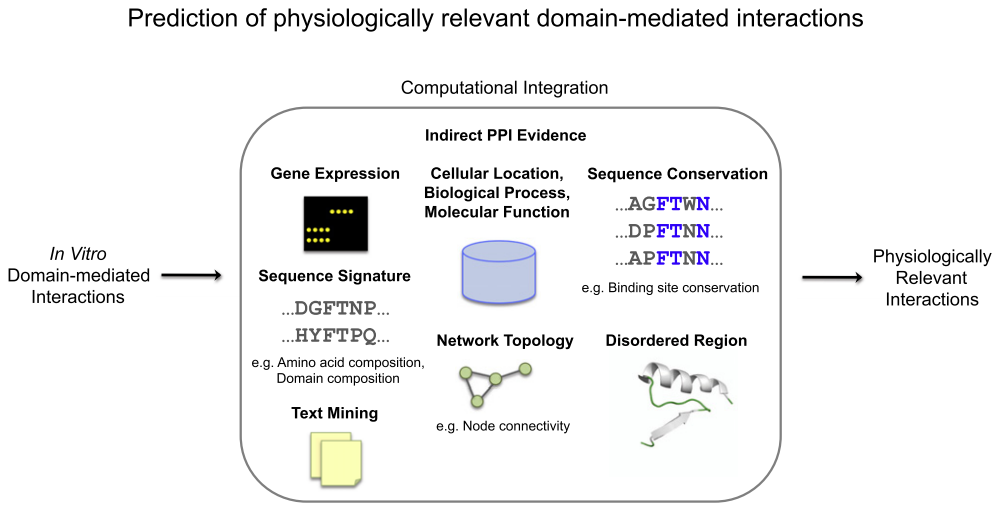

Fig. 2. Prediction of physiologically relevant domain-mediated interactions. Besides domain–peptide interactions, additional data is required to predict high-confidence, physiologically relevant domain–peptide interactions. Indirect evidence such as gene expression, protein function and sequence similarity can be used to identify in vivo protein interactions. Numerous methods allow PPI prediction using single sources of evidence, while machine learning techniques such as naïve Bayes can be used to combine multiple datasets to predict the biologically relevant domain–peptide mediated interactions.

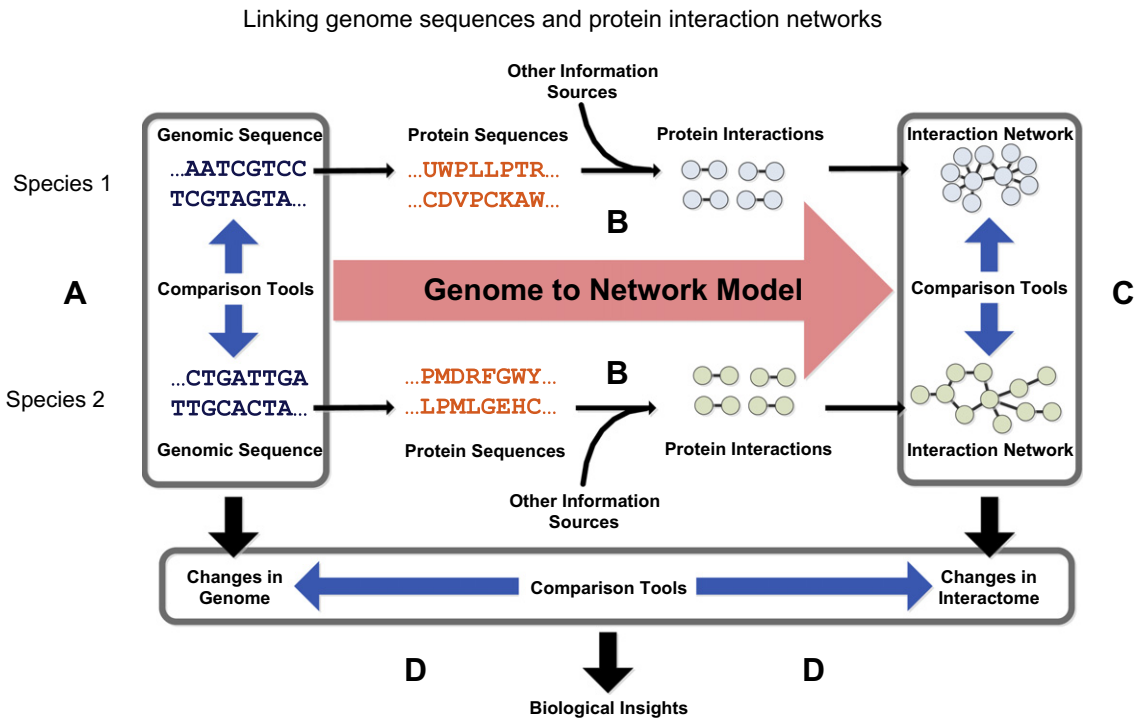

Fig. 3. Linking genome sequences and protein interaction networks. The genome-to-network model involves four classes of analyses. First, comparison of two or more genomes will reveal DNA and protein sequence differences between species (A), and sequence alignment algorithms are readily available. Second, analysis of PRMs and protein sequences will enable construction of protein–protein interaction networks from genomes (B). Third, comparison of two or more protein–protein interaction networks using network alignment algorithms (C) will enable investigation of the differences of interactomes across species. Finally, genome-to-network and network-to-genome analyses (D) will allow us to interpret genome sequence changes in the context of interactomes, and vice versa.

2.5. Structure-based and sequence-based methods predict different interactions

Predictors trained using different features produce different predictions. We quantified this phenomenon for PDZ domains by comparing the domain–peptide predictions from sequence-based and structure-based predictors to known PDZ mediated PPIs [42]. By considering predictions from both methods, 11% of known PDZ-mediated PPIs were recovered. However, 72% of the results were obtained by either the sequence-based or the structure-based predictor. Our analysis showed that different sets of interactions were found by the predictors, which can be attributed to the use of different features for training. Therefore, to obtain the greatest coverage of domain mediated PPIs, it is important to combine different feature encodings and presumably different prediction methods.

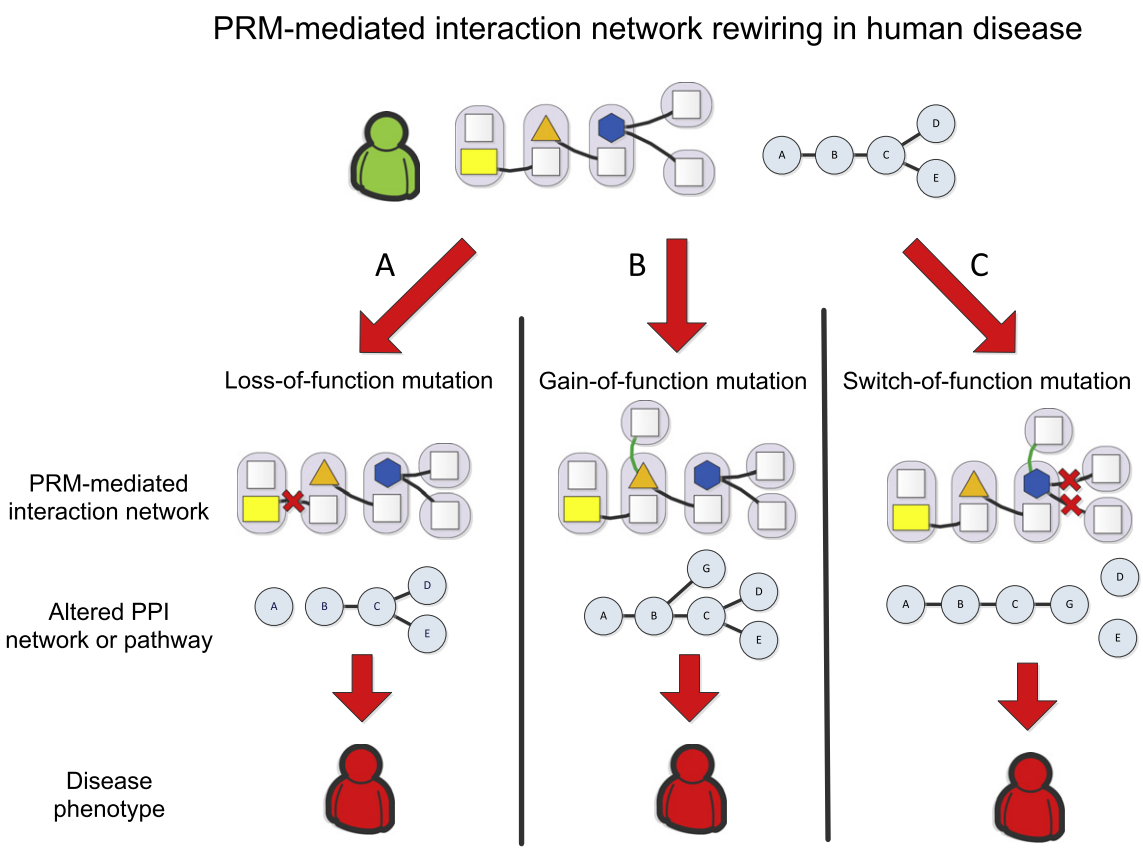

Fig. 4. PRM-mediated interaction network rewiring in human disease. Disease-related DNA mutations in critical protein regions related to PRM-mediated interactions can cause multiple types of network rewiring resulting in altered pathways and phenotypes, broadly classified into three groups. Most mutations are expected to be loss-of-function (A), e.g. in which important residues in peptide recognition modules or binding sites are mutated such that an interaction no longer occurs (red cross). In contrast, gain-of-function mutations (B) e.g. are caused by mutations that introduce new binding sites, creating novel interactions between PRMs and target proteins (green edge). The third class of 'switch-of-function' mutations (C) [143] e.g. covers modifications in PRMs or binding sites such that some interactions are introduced and others deleted, resulting in network rewiring of multiple interactions.

2.6. Improving domain–peptide interaction prediction methods

While current PRM interaction predictors perform reasonably well, other domain features should be considered in the future. It is known, for example, that the structural flexibility of the domain-binding pocket can contribute to a PDZ domain's ability to bind specific ligands [48–50]. Recently, a model of PDZ domain backbone flexibility was used to successfully predict domain binding specificity for a subset of human PDZ domains [40]. Therefore, domain backbone flexibility features may be used to train a predictor with improved performance. Structure and sequence features can also be combined with other features to build a single predictor that utilizes all available types of information. Alternatively, ensemble predictors may be developed which combine the results of structure-based and sequence-based predictors. Molecular dynamics algorithms for protein-peptide docking should also be used. While these methods can produce highly quality docking models [51,52], they may be computationally intensive or difficult to apply to domains with no available structures. Since domains may prefer to bind to more than one binding motif, a multiple specificity model may prove useful. For instance, Gfeller et al. modeled the multiple specificities of PDZ domains using several PWMs, following the observation that ligand residue positions are often significantly correlated [53]. Finally, while the discussed methods help predict accurate *in vitro* domain–peptide interactions, additional incorporation of cellular context information is required for predicting interactions that occur *in vivo*.

3. Physiologically relevant PRM-mediated protein–protein interactions

Proteins will only interact if they recognize each other, and are temporally and spatially co-located in the cell. Domain–peptide interaction predictors described above allow us to discover protein pairs that recognize each other. Additional sources of evidence must be considered to accurately score domain–peptide interactions by their physiological relevance, such as the correlation of the expression profiles of the corresponding genes, their involvement in related biological processes, and their presence in the same cellular compartment (Fig. 2). Gene expression profiles, cellular location of proteins, functional annotation (molecular function and biological process), sequence signatures, literature references, and known experimental interactions can be obtained from diverse biological data sources and combined to predict physiologically relevant protein interactions. A number of computational methods have been developed for evaluating protein interactions using single sources of evidence, while others combine multiple types of knowledge using ensemble approaches. As domain-mediated networks are only now emerging, few methods have been developed specifically for these data. However, the collection of methods developed for analysing traditional protein–protein interaction networks can be combined with sequence- and structure-based domain–peptide interaction prediction methods discussed above to define physiologically relevant high-resolution PRM-mediated interaction networks.

3.1. Cellular location, biological process, molecular function

Proteins are more likely to interact with each other when they co-localize in the same cellular compartment or are part of the same biological processes. Gene Ontology (GO) is a useful and popular taxonomy that contains a hierarchy of controlled terms regarding cellular location, biological process and molecular function [54]. GO terms are used to annotate genes and proteins based on experimental and computational evidence and literature curation. This resource can be used to quantify the functional relationship between different proteins using a straightforward comparison of associated annotations, that is, two proteins are related if they have many annotations in common. More elaborate semantic similarity measures consider the entire GO hierarchy in comparing two interacting proteins, that is, two proteins are related if they have many similar annotations in common. For instance, two GO terms that are close in the hierarchy are similar.

Semantic similarity provides a quantitative measurement of the likeness of concepts belonging to an ontology. In the context of PPIs, higher semantic similarity scores between GO terms annotated to a protein pair indicate a higher likelihood of these proteins interacting *in vivo*. Guo et al. compared a number of network-based and information content-based semantic similarity methods in distinguishing true and false human PPIs, and concluded that the average (AVG) method by Resnik performed best in AUROC analysis [55–58]. Xu et al. compared the AVG and maximum (MAX) methods by Resnik to a number of semantic similarity methods specifically developed for GO, and concluded that the MAX method by Resnik outperforms others when considering the three ontologies of GO either individually or together [55,59–62]. More recently, our group developed the Topological Clustering Semantic Similarity (TCSS) method, which uses a novel normalization technique before computing similarity [63]. TCSS improves the performance of PPI predictions in all three branches of GO compared to other semantic similarity measures.

Prediction of protein–protein interactions using GO has several limitations. Notably, GO annotations are often noisy, as more than one third of all annotations and ~75% of human gene annotations are assigned using automated methods [64]. Such low-confidence annotations, labeled as 'Inferred from Electronic Annotation' (IEA), should be excluded from predictions when higher quality annotations are available. Additionally, the structure of GO is often unbalanced since some biological processes are studied more extensively than others, leading to ascertainment biases in predicting protein interactions. As semantic similarity measures use knowledge structured in the form of ontologies, other ontologies could be substituted for GO. Some describe highly structured biological pathway mechanisms, such as the BioPAX pathway representation standard [65]. Large amounts of curated pathway data are available in this format, such as from the Reactome pathway database [66]. Further development of semantic similarity methods that consider such ontologies could improve PPI prediction.

3.2. Gene expression

Gene expression is a frequently used measure for assessing the confidence and biological relevance of predictions from high-throughput PPI experiments. As proteins must be expressed in order to interact, interacting proteins should be co-expressed at the same time and are likely to have similar gene expression profiles. The association between protein interactions and correlated gene expression profiles has been demonstrated in several studies. Co-expressed genes in yeast and bacteriophage T7 were shown to be enriched in protein interactions, and clusters of gene expression profiles frequently contained interacting proteins in yeast [67]. Jansen et al. demonstrated a strong correlation between the gene expression profiles of yeast proteins involved in the same complex [68]. Bhardwaj et al. compared the gene expression profiles of interacting and random gene pairs in *Escherichia coli*, and concluded that genes encoding for interacting proteins have a stronger expression pattern correlation that is also more conserved than for random protein pairs [69]. Consequently, PPI prediction methods frequently use strong co-expression of genes as an evidence source for protein interactions [70,71].

While gene expression data is a useful source of evidence, it has a number of limitations. Adler et al. studied curated Reactome pathways in the context of the human gene expression atlas, and concluded that co-expression is sufficient for reconstructing pathways such as metabolism and translation, while dynamic signaling processes are captured to a lesser extent [72]. Liu et al. noted that six large protein complexes, including the ribosome, provided the majority of the signal between expression correlation and protein interactions in several gene expression datasets in yeast, while many other protein complexes did not show an association [73]. Further, tissue-specific and developmental programs regulate gene expression in multi-cellular organisms, meaning that the global co-expression of potentially interacting proteins is not necessarily informative of their co-expression in a given cellular state. Future work to improve the confidence of co-expression data for high-resolution PRM-mediated interaction networks will involve novel methods for determining global co-expression of genes, such as MEM [74]. Emerging experimental and computational technologies (e.g. RNA-seq and the discovery of the alternative splicing code [75]) will also help distinguish between multiple alternative transcripts of a single gene, some of which include a PRM or ligand while others do not.

3.3. Sequence signature

Protein interactions can also be predicted based on correlated sequence motifs. These motifs are learned from existing PPIs using only sequence data and characterize direct binding, but also could be related to protein function, which is in turn predictive of PPIs [76]. Methods based on information content analyze co-occurring subsequences of proteins with experimentally verified interactions, and use these patterns for predicting new interactions. Pitre et al. developed the Protein–protein Interaction Prediction Engine (PIPE), which finds co-occurrences of subsequences in pairs of proteins with known interactions [77]. Sprinzak and Margalit identified over-represented sequence signatures in known PPIs and then used this information for predicting novel interactions [78]. Najafabadi and Salavati introduced a method based on codon usage as a predictor for PPIs [79].

Machine learning methods use sequence information regarding a gold standard set of positive and negative PPIs to classify new pairs of potentially interacting proteins. Various approaches mainly differ in their encoding of sequence features and choice of learning functions. For instance, Martin et al. encoded the sequence information for a protein pair by a product of signatures [80], while Shen et al. proposed the use of conjoint triads, *i.e.*, frequencies of continuous subsequences of three residues [76]. Guo et al. used the auto-correlation values of seven different physicochemical scales for protein sequences as protein interaction predictors [81]. Roy *et al.* explored the contribution of pure amino acid composition for predicting protein interactions and concluded that this simple feature outperforms domains and other sequence features such as tuples and signature products [82].

A major limitation of sequence signatures for predicting protein interactions is the generally weak correlation between sequence and functional similarity. Limitations of these machine learning methods are similar to the ones described above, notably including the lack of well-defined true negative examples. For instance, Yu

et al. evaluated the effect of positive-to-negative ratio in training and test sets for SVM based methods and found that it had considerable effect on classifier accuracy [83]. Lastly, use of sequence signatures to refine high-resolution PRM-mediated interaction networks must avoid duplicate counting of the domain-motif interaction knowledge already used to generate the original network.

3.4. Network topology

Much work has been done in defining the relationship between PPI network topology and biological function, with the conclusion that two proteins that have many shared neighbors in a PPI network are more likely to interact [84]. The property of highly connected components, *i.e.*, network cohesiveness, in small-world networks is often used to assess the confidence of predicted protein–protein interactions. Goldberg and Roth showed that true interactions have higher neighborhood cohesiveness as compared to false interactions [85]. Conversely, a predicted PPI is more likely to be true if it shows a higher degree of neighborhood cohesiveness. Bader et al. proposed that interacting proteins with shared interactors are more likely to be biologically relevant [86]. Yu et al. predicted interactions in protein networks by completing their partially connected components, applying the assumption that proteins within the same protein cluster are likely to interact with each other [87].

An important challenge for network-based protein interaction predictors is the identification of appropriate topological clusters in networks. Larger cluster sizes lead to an increased rate of false positives, while overly small clusters have few positive predictions. Clustering and cohesiveness analysis of PRM-mediated protein interaction networks may require additional research, as their topological properties may differ from traditional PPI networks. Finally, prediction of PRM-mediated protein interactions based on known interactions will require careful filtering of data to avoid duplicate counting of evidence.

3.5. Text mining

Text mining, the technique of automated information extraction from literature, can also be used to predict PPIs. While text mining is error-prone and unlikely to improve substantially in the near future due to challenges in the computational analysis of natural language, it has been shown to be useful in predicting protein–protein interactions. For instance, scientific publications contain more references to interacting proteins than expected, even if no previous experimental evidence of their interaction exists [88].

The simplest way to extract PPIs from literature is to detect co-occurring protein names and apply statistical methods to find significantly frequent co-references [89]. More complex text mining methods rely on natural language processing techniques that attempt to parse the meaning of sentences in which the proteins of interest are mentioned. For instance, Ramani et al. developed a text mining pipeline of machine learning and natural language processing methods to predict human protein–protein interactions from MedLine abstracts, and recovered a network with comparable accuracy to existing PPI networks [90]. The MedScan information extraction system is a similar predictor that involves a natural language parser for full sentences [91].

Automatic extraction of information from millions of scientific publications is a computational challenge, since the complexity of a natural language processing process is directly proportional to predictor accuracy but inversely proportional to computational efficiency. Purely statistical methods of interaction prediction are sensitive to noise and indirect interactions, and are likely to miss the meaning of negated interactions. Much work is currently being done to improve text mining accuracy. For instance, the BioCreative workshop promotes the development of text mining tools, including protein name and PPI recognition [92].

3.6. Additional types of PPI evidence

Many sources of PPI evidence exist and not all can be covered here in detail. We present a few additional useful ones below.

3.6.1. Evolutionary conservation

Protein interactions conserved across species are likely to be biologically relevant and this information is predictive of true positive interactions. For example, the I2D database contains protein interactions translated across species through orthology relations [93]. Also, binding sites detected by PRM-focused methods are more likely to be physiologically relevant if they are conserved across species [94].

3.6.2. Correlated mutations

Evolutionary mutations within the binding sites of a PPI are known to correlate, as significant mutations in one binding surface must be compensated for by mutations in the other binding surface to maintain binding [95,96]. Several methods allow detection of correlated mutations in multiple sequence alignments of protein orthologs with domains and motifs [97–99].

3.6.3. Protein binding specificity

Binding competition affects protein interactions in vivo [100]. Quantitative protein concentration and affinity data allows us to assess competition, but these data are not readily available. However, an interaction between two proteins is less likely to compete and thus more likely to be correctly predicted if it involves proteins interacting with only each other and not with any other partners.

3.6.4. Surface accessibility

Predicted binding sites are more likely to be relevant if they are accessible on the protein surface and thus available for binding. Surface accessibility can be predicted computationally from protein structures using tools such as the Eukaryotic Linear Motif (ELM) structure filter [101]. Amino acid sequence-based predictors such as PHDacc are useful when no known protein structure is available [102]. Linear motifs are known to cluster in disordered regions, and disorder predictors, such as and GlobPlot and DISOPRED, have been shown to be highly informative for identifying relevant PRM binding sites in proteins [94].

3.7. Combining PPI evidence sources using ensemble methods

We have discussed most PPI evidence sources as singular predictors, meaning that a single type of data is used to classify protein pairs as either interacting or non-interacting. In contrast, ensemble approaches, generally using machine learning, combine various lines of evidence into a single predictor, which generally improves predictor performance substantially. As another advantage, ensemble approaches can consider multiple weak evidence sources that are individually insufficient for PPI prediction, but are informative when combined with stronger evidence. Several research groups have independently suggested the use of ensemble methods for predicting protein interactions, though data sources, techniques, and implementations vary widely [70,71,86,103,104].

Bayesian integration is the most widely used ensemble technique for PPI prediction. Although other machine learning approaches have been used for this task, such as logistic regression, random forests, decision trees, and support vector machines, Bayesian integration remains the method of choice due to its simple

probabilistic framework and ability to handle missing data. The objective of a Bayesian PPI prediction model is to estimate the probability that a given protein pair interacts, given the biological evidence supporting the interaction. For simplicity, a naïve Bayesian model assumes complete independence between different evidence sources. Jansen et al. proposed the use of Bayesian networks on a fully dependent feature set of experimental PPI data and naive integration of indirect evidence such as mRNA co-expression, biological function, and gene essentiality in *S. cerevisiae* [105]. Rhodes et al. employed a similar but semi-naïve Bayesian strategy to combine homologous PPI, gene co-expression, GO process, and domain-based sequence evidence in human interaction networks, assuming a certain level of dependence between different evidence sources [71]. Scott and Barton extended the Bayesian probabilistic framework for the prediction of human PPIs with more features, including local network topology, co-expression, orthology to known interacting proteins, sub-cellular localization, co-occurrence of domains, and post-translational modifications [106]. Bayesian predictors have been also applied in the context of domain-specific interaction networks. Tonikian et al. used phage display, yeast two-hybrid, and peptide array screening to independently identify SH3 domain binding partners [8]. The authors then combined the results of these complementary techniques using a Bayesian algorithm to generate a high-confidence SH3-mediated interaction network for yeast.

4. Mapping how sequence changes affect the network

Even if the possibility of mapping an entire interactome based on genomic sequence is distant, progressive steps towards such a goal would still provide scientific benefit. The combination of domain–peptide interaction prediction with additional biologically relevant evidence sources will produce a high-confidence and high-resolution PPI network (Fig. 3). This can be combined with additional high-resolution data, such as binding sites from three-dimensional protein structures accumulated in the Protein Databank [107] to increase proteome coverage of the network [108]. The network and the genome to network mapping method can then be used to predict the functional effect of evolutionary mutations and sequence changes on the protein interactome. This would enable us to better understand the relationship between protein sequence and function, of which protein–protein interactions are one component, and predict the functional consequences of evolutionary or disease-associated mutations. We have excellent computational tools to identify sequence changes, *e.g.*, sequence alignment algorithms, and similar tools such as network alignment algorithms are needed for interactome analysis. The combination of such tools will allow identification of modifications in sequences and corresponding networks, and enable us to relate genome-level changes to interaction networks.

4.1. Protein–protein interaction network alignment

As with gene and protein sequences, two or more interaction networks can be aligned to one another for comparison to reveal all conserved and altered interactions. PPI network alignments, like sequence alignments, can also reveal functional similarity and orthology between individual proteins. For instance, if the interaction partners of two proteins in distinct species are orthologous, the proteins are also likely to be orthologous. PPI network alignments can also reveal similarity between larger functional units, such as protein complexes and pathways [109]. PPI network alignment algorithms must account for imperfect matches, just as sequence alignment algorithms allow mismatches and gaps, as PPI networks are known to rewire over time [110,111].

As with sequence alignment, there are two distinct approaches for PPI network alignment, local and global alignment. PathBLAST, one of the first published PPI network alignment algorithms, identifies putative protein–protein interaction paths conserved between two species using a local approach [109,112]. Alternatively, global PPI network alignment methods attempt to align all or most of the proteins in two or more networks. Unlike local methods, global methods make no assumptions about the size or shape of conserved protein interaction patterns, acting on all proteins rather than small sub-networks. However, global methods are likely to create more false positive matches than local alignments, as many more interactions are aligned, even those with weak evidence. Still, the IsoRank, GRAAL, and Graemlin families of global methods produce alignments with significant levels of functional similarity between aligned proteins [113–115].

The local and global models of network alignment are similar but perform different tasks. Local methods can be used to identify conserved building blocks of biological systems such as protein complexes and pathways, whereas global network alignment methods can be used to analyze the variations between the entire interactomes of two or more species. This latter analysis is likely to reveal large-scale, topological trends that emerge over evolutionary time, such as increasing network complexity and redundancy [116]. Much work is still required to improve network alignment methods and a number of concepts from sequence alignment have not yet been extended to network alignment. For instance, no network alignment algorithm considers how gene or genome duplication events affect the network, similar to sequence analysis identifying large-scale genome duplication events [117]. Also, there are currently no published computational methods that can align PRM-mediated protein interactions, though existing algorithms are likely to be adaptable.

4.2. Evolutionary analysis of network alignments

The study of PPI network alignments will improve understanding of how both genomes and interactomes evolve. Network changes, mediated by sequence changes, are selected for by their phenotypic result. Therefore, investigating network-level properties of a protein can help explain protein evolution. For example, central proteins with numerous interaction partners, known as hubs, are generally more conserved at the sequence level [118,119]. It has been speculated that mutations in hub proteins are more likely to cause loss of interactions, thus disturbing the proteins' ability to function. However, there are many unsolved problems in this area, such as the mechanisms behind the emergence of hubs in PPI networks [120–122].

PPI network alignments between species will highlight the extent of conservation in different regions of the interactome over evolutionary time. As genes encoding core cellular machinery are often conserved across species, the corresponding interaction networks are also probably conserved. Other more dynamically evolving processes are likely to show divergence between the network evolution rate and the mutation rate of the underlying protein sequences. Comparison of the SH3-mediated network of *C. elegans* (unpublished data) with the yeast SH3 network [8] revealed an expanded set of worm SH3 domains with conserved fundamental roles already established in yeast, although interactions within the network are heavily rewired. This observation indicates that network alignment may identify functionally related proteins based on their network location that cannot be easily identified via sequence-based orthology methods.

4.3. Evolution of domain specificity and domain–peptide interactions

Investigation of protein binding sites in a high-resolution PRM-mediated interaction network will provide new insights into protein evolution, interaction and function. Kim et al. analyzed one of the first high-resolution interaction networks, derived from protein 3D structures, and recognized that hub proteins with many different binding interfaces evolve slower and are more likely to be essential than hubs with few interfaces [123]. Such distinctions can be observed only in high-resolution interaction networks. Even removed from a network context, ligand information is highly useful for evolutionary PPI studies. By identifying protein residues targeted by PRMs, one can highlight functional regions that are more likely to be sensitive to mutations. The fact that PRMs often target disordered regions indicates that these regions are under evolutionary constraints, even though this may not be obvious from sequence analysis.

Further research will lead to improved understanding of the function and evolution of different parts of the proteome [124,125]. The study of high-resolution PRM-mediated interaction network alignments will advance our understanding of the role of sequence changes mediating network modifications. For instance, a number of models have been proposed for network evolution that explain the power law node degree distribution observed in PPI networks, including the "preferential attachment" hypothesis which states that new proteins will more likely interact with proteins that already have many interactions [126–130]. The tracing of the evolution of a domain's specificity and interactions would identify which of these models is more likely correct. Additional unresolved problems include whether highly connected domains slowly accumulate interactions over time or suddenly gain multiple interactions with a single mutation. Knowledge of how interactions are formed and destroyed between proteins will lead to increased understanding of how genes, complexes and pathways evolve. We now need high-resolution PPI networks, such as those mediated by PRMs, across multiple species to enable research into these network evolution problems.

5. PRM-mediated protein–protein interaction networks in human disease

PRMs are central in cell signaling systems [3,4,15,18,131–134] and have been implicated in numerous diseases [135–137]. For instance, cell polarity disruption, due to perturbed PDZ and SH3 domains, is involved in tumor metastasis and various immune deficiencies, while protein kinase and phosphosite mutations play important roles as cancer drivers. Systematic analysis of high-confidence, high-resolution protein interaction networks will therefore help explain disease phenotypes and identify new diagnostic and predictive biomarkers and new therapeutic targets (Fig. 4).

5.1. Disease mutations and interaction network rewiring

DNA mutations underlie the majority of inherited human disease. Monogenic diseases involve single gene variants and typically follow Mendelian rules of heritability, while complex diseases such as diabetes and some types of mental illness, like schizophrenia, involve many disease predisposition genes. In contrast to disease caused by genetically inherited mutations, cancer is mostly driven by sporadic, malignant mutations in somatic tissues, though many predisposing genetic factors of cancer, such as BRCA1/2 mutations, are known to exist [138]. Disease genes are tracked in the OMIM cataloge of genetic disorders [139], disease mutations are stored in the Human Gene Mutation Database [140], and somatic cancer mutations are collected in the COSMIC database [141].

Broadly, disease mutations are either loss-of-function or gain-of-function depending on how the mutation affects the biochemical mechanism leading to the disease. In the context of high-resolution protein interaction networks, the main focus of analysis involves disease mutations that affect binding domains or ligands and cause interaction rewiring. Point mutations (SNPs or SNVs) are easiest to interpret when they affect single PRM binding sites and ligands, as opposed to stop codon mutations and multi-amino acid alterations, which are likely to alter whole protein regions potentially containing multiple binding sites. There are three effects of mutations on network topology: they can disable protein–protein interactions due to reduced binding affinity of PRMs or ligands, they can introduce new interactions between proteins by strengthening binding affinity between binding interfaces, or they can change existing interaction partners. Analysis of mutations associated with human diseases in high-resolution protein interaction networks will provide insight into the mechanism of action of the mutation and its role in causing disease.

5.2. Interpretation of disease-association DNA mutations in protein interaction networks

Functional interpretation of disease mutations attempts to link disease phenotypes with their underlying genetic causes and biochemical mechanisms. A number of criteria are used to identify potentially disease-causing mutations. Protein coding mutations can have diverse impacts, such as on the stability, expression, subcellular location, and interactions of the protein. Disease-associated mutations are often considered to be more serious than sequence variants frequently seen in the general population. Wang and Moult compared disease mutations and common SNPs and showed that 90% of disease mutations have a functional impact on the protein and its stability, while most common SNPs are functionally neutral [142].

Computational analysis of disease mutations and their functional impact, based on many criteria, is an active field of research (e.g. [143], reviewed in [144,145]). Several statistical and machine learning methods have been developed that consider the physicochemical properties of amino acids, protein structure, sequence, and conservation in evaluating the impact of mutations. Less work has been done in evaluating disease mutations that specifically affect protein binding. Schuster-Böckler and Bateman used homology modeling of protein structures to define high-confidence interaction interfaces and found numerous mutations that alter these sites in human disease [146]. Teng et al. studied SNP-induced electrochemical changes in protein interaction interfaces from 3D structures and concluded that biochemical properties of involved amino acids are not sufficient for functional predictions [147]. However, they also showed that SNPs associated with disease and SNPs in highly conserved positions tend to create greater changes in binding energy. These studies are limited due to the small number of high-confidence protein structures, but high-resolution protein interaction networks with greater coverage will allow precise interpretation of many more human disease mutations [108]. Mapping a large number of disease-associated mutations and their network effects will also enable a better understanding of the relationship between genomes and networks.

5.3. Domain-mediated protein interaction networks and cancer

Consideration of domain-mediated protein interaction networks in mutation analysis will help researchers better explain mutations in complex diseases, such as cancer. Recent cancer genomic projects map a variety of molecular profiles, including

somatic DNA mutations, gene expression, DNA copy number and methylation [148,149]. Such multidimensional datasets are useful for integrated analysis of high-resolution protein interaction networks. For instance, the sequence- and structure-based predictors described above could be used to assess the functional impact of somatic mutations, or the gene expression and copy number aberration data can be used to construct high-confidence networks relevant to the particular cancer under study.

The collection of known somatic mutations in cancer is characterized by a 'long tail': a few genes are mutated in numerous cancer samples and types, while hundreds of genes harbor rare mutations observed only in a few samples. Consequently, researchers focus on frequently mutated genes, and less effort is directed to genes with rare mutations that are more difficult to interpret. High-resolution protein interaction network information will be useful to identify rare mutations that affect binding sites in specific signaling pathways. This analysis will provide a hypothesis about the mechanism of action of a mutation, highlighting it as potentially important for tumor development. Many known links exist between important regulatory networks and domain-mediated protein interactions affected by cancer mutations. For instance, TP53 is the most frequently mutated gene in human cancers, encoding the tumour-suppressing transcription factor that regulates DNA damage response genes and apoptosis. TP53 is regulated by various post-translational modifications, in particular phosphorylation by protein kinase domains. Several phosphorylation sites are involved in the inhibition of TP53 and these are frequently mutated in cancer [150]. This example demonstrates the link between transcriptional regulation and domain-mediated interaction networks of post-translational modifications. The epidermal growth factor receptor EGFR is another prominent example of a cancer gene with well-described domain-specific mutations. The active site of kinase domain EGFR is frequently mutated in lung cancer, and these mutations, comprising 40–45% of patients with EGFR mutations, are used as predictive clinical biomarkers of treatment response [151]. As active site mutations directly determine the kinase activity of EGFR, these will have an impact on downstream signaling and PRM-mediated interaction networks.

6. Conclusion

Understanding how networks are encoded in the genome will help address numerous scientific problems. The approach of building high-confidence, high-resolution protein interaction networks based on peptide recognition modules using computational and experimental methods will provide a useful set of data covering mainly eukaryotic cell signaling systems. We look forward to combining this information with high-resolution networks derived from other sources to eventually develop a complete high-resolution molecular interaction map of the cell. This map will be useful for understanding how DNA mutations alter phenotype at the cellular and organism levels.

References

[1] Lensink, M.F. and Wodak, S.J. (2010) Docking and scoring protein interactions: CAPRI2009. Proteins 78 (15), 3073–3084.
[2] Bhattacharyya, R.P. et al. (2006) Domains, motifs, and scaffolds: the role of modular interactions in the evolution and wiring of cell signaling circuits. Annu. Rev. Biochem. 75, 655–680.
[3] Pawson, T., Gish, G.D. and Nash (2001) SH2 domains, interaction modules and cellular wiring. Trends Cell Biol. 11 (12), 504–511.
[4] Pawson, T. and Nash (2003) Assembly of cell regulatory systems through protein interaction domains. Science 300 (5618), 445–452.
[5] Letunic, I. et al. (2006) SMART 5: domains in the context of genomes and networks. Nucleic Acids Res. 34 (Database issue), D257–D260.
[6] Mulder, N.J. et al. (2007) New developments in the InterPro database. NucleicAcids Res. 35 (Database issue), D224–D228.
[7] Bateman, A. et al. (2004) The Pfam protein families database. Nucleic Acids Res. 32 (Database issue), D138–D141.
[8] Tonikian, R. et al. (2009) Bayesian modeling of the yeast SH3 domain interactome predicts spatiotemporal dynamics of endocytosis proteins. PLoS Biol. 7, el000218.
[9] Tonikian, R. et al. (2008) A specificity map for the PDZ domain family. PLoS Biol. 6 (9), e239.
[10] Tong, A.H. et al. (2002) A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. Science 295 (5553), 321–324.
[11] Landgraf, C. et al. (2004) Protein interaction networks by proteome peptides canning. PLoS Biol. 2 (1), E14.
[12] Wiedemann, U. et al. (2004) Quantification of PDZ domain specificity, prediction of ligand affinity and rational design of super-binding peptides. J. Mol. Biol. 343 (3), 703–718.
[13] Hu, H. et al. (2004) A map of WW domain family interactions. Proteomics 4 (3), 643–655.
[14] Goel, R. et al. (2012) Human protein reference database and human proteinpedia as resources for phosphoproteome analysis. Mol. Biosyst. 8 (2), 453–463.
[15] Nourry, C., Grant, S.G. and Borg, J. (2003) PDZ domain proteins: plug and play! Sci. STKE 2003, RE7.
[16] Macias, M.J., Wiesner, S. and Sudol, M. (2002) WW and SH3 domains, two different scaffolds to recognize proline-rich ligands. FEBS Lett. 513, 30–37.
[17] Zarrinpar, A., Bhattacharyya, R. and Lim, W.A. (2003) The structure and function of proline recognition domains. Sci. STKE 2003, RE8.
[18] Pawson, T. and Gish, G.D. (1992) SH2 and SH3 domains: from structure to function. Cell 71, 359–362.
[19] Tonikian, R. et al. (2007) Identifying specificity profiles for peptide recognition modules from phage-displayed peptide libraries. Nat. Protoc. 2 (6), 1368–1386.
[20] MacBeath, G. and Schreiber, S.L. (2000) Printing proteins as microarrays for high-throughput function determination. Science 289, 1760–1763.
[21] Stiffler, M.A. et al. (2007) PDZ domain binding selectivity is optimized across the mouse proteome. Science 317, 364–369.
[22] Huang, H. et al. (2008) Defining the specificity space of the human SRC homology 2 domain. Mol. Cell. Proteomics: MCP 7 (4), 768–784.
[23] Wu, C. et al. (2007) Systematic identification of SH3 domain-mediated human protein–protein interactions by peptide array target screening. Proteomics 7 (11), 1775–1785.
[24] Ceol, A. et al. (2007) DOMINO: a database of domain–peptide interactions. Nucleic Acids Res. 35, D557–D560.
[25] Beuming, T. et al. (2005) PDZBase: a protein–protein interaction database for PDZ-domains. Bioinformatics 21 (6), 827–828.
[26] Razick, S., Magklaras, G. and Donaldson, I.M. (2008) iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinformatics 9, 405.
[27] Davis, J. and Goadrich, M. (2006)The Relationship Between Precision-Recall and ROC Curves, Vol. 3, ACM, Pittsburgh.
[28] Baldi, P. et al. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16, 412–424.
[29] Lehrach, W.P., Husmeier, D. and Williams, C.K. (2006) A regularized discriminative model for the prediction of protein–peptide interactions. Bioinformatics 22, 532–540.
[30] Yaffe, M.B. et al. (2001) A motif-based profile scanning approach for genome-wide prediction of signaling pathways. Nat. Biotechnol. 19, 348–353.
[31] Wunderlich, Z. and Mirny, L.A. (2009) Using genome-wide measurements for computational prediction of SH2–peptide interactions. Nucleic Acids Res. 37, 4629–4641.
[32] Brinkworth, R.I., Breinl, R.A. and Kobe, B. (2003) Structural basis and prediction of substrate specificity in protein serine/threonine kinases. Proc. Natl. Acad. Sci. USA 100, 74–79.
[33] Hui, S. and Bader, G.D. (2010) Proteome scanning to predict PDZ domain interactions using support vector machines. BMC Bioinformatics 11, 507.
[34] Chen, J.R. et al. (2008) Predicting PDZ domain–peptide interactions from primary sequences. Nat. Biotechnol. 26 (9), 1041–1045.
[35] Eo, H.S. et al. (2009) A machine learning based method for the prediction of Gprotein-coupled receptor-binding PDZ domain proteins. Mol. Cells 27, 629–634.
[36] Shao, X. et al. (2011) A regression framework incorporating quantitative and negative interaction data improves quantitative prediction of PDZ domain–peptide interaction from primary sequence. Bioinformatics 27 (3), 383–390.
[37] Hue, M. et al. (2010) Large-scale prediction of protein–protein interactions from structures. BMC Bioinformatics 11, 144.
[38] Sanchez, I.E. et al. (2008) Genome-wide prediction of SH2 domain targets using structural information and the FoldX algorithm. PLoS Comput. Biol. 4, el000052.
[39] Fernandez-Ballester, G. et al. (2009) Structure-based prediction of the *Saccharomyces cerevisiae* SH3–ligand interactions. J. Mol. Biol. 388, 902–916.
[40] Smith, C.A. and Kortemme, T. (2010) Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. J. Mol. Biol. 402, 460–474.
[41] Kaufmann, K. et al. (2011) A physical model for PDZ-domain/peptide interactions. J. Mol. Model 17, 315–324.
[42] Hui, S., Xing, X. and Bader, G.D. (submitted for publication) Predicting PDZ domain mediated protein interactions from structure.

[43] Zhang, Y. (2009) Protein structure prediction: when is it useful? Curr. Opin. Struct. Biol. 19, 145–155.
[44] Fischer, D. (2006) Servers for protein structure prediction. Curr. Opin. Struct. Biol. 16, 178–182.
[45] Ben-Hur, A. and Noble, W.S. (2006) Choosing negative examples for the prediction of protein–protein interactions. BMC Bioinformatics 7 (Suppl 1), S2.
[46] Lo, S.L. et al. (2005) Effect of training datasets on support vector machine prediction of protein–protein interactions. Proteomics 5, 876–884.
[47] Smialowski, P. et al. (2010) The Negatome database: a reference set of non-interacting protein pairs. Nucleic Acids Res. 38, D540–D544.
[48] Appleton, B.A. et al. (2006) Comparative structural analysis of the Erbin PDZ domain and the first PDZ domain of ZO-1. Insights into determinants of PDZ domain specificity. J. Biol. Chem. 281 (31), 22312–22320.
[49] Skelton, N.J. et al. (2003) Origins of PDZ domain ligand specificity. Structure determination and mutagenesis of the Erbin PDZ domain. J. Biol. Chem. 278 (9), 7645–7654.
[50] Chen, Q. et al. (2007) Solution structure and backbone dynamics of the AF-6 PDZ domain/Bcr peptide complex. Protein Sci. 16, 1053–1062.
[51] Cavasotto, C.N. and Abagyan, R.A. (2004) Protein flexibility in ligand docking and virtual screening to protein kinases. J. Mol. Biol. 337, 209–225.
[52] Antes, I. (2010) DynaDock: a new molecular dynamics-based algorithm for protein–peptide docking including receptor flexibility. Proteins 78, 1084–1104.
[53] Gfeller, D. et al. (2011) The multiple-specificity landscape of modular peptide recognition domains. Mol. Syst. Biol. 7, 484.
[54] Ashburner, M. et al. (2000) Gene ontology: tool for the unification of biology. Gene Ontology Consortium. Nat. Genet. 25 (1), 25–29.
[55] Resnik, P. (1995) Using information content to evaluate semantic similarity in a taxonomy, in: IJCAI'95: Proceedings of the 14th International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
[56] Lin, D. (1998) An information-theoretic definition of similarity, in: Proceedings of the 15th International Conference on Machine Learning, Morgan Kaufmann. pp. 296–304.
[57] Jiang, J.J. and Conrath, D.W. (1997) Semantic similarity based on corpus statistics and lexical taxonomy, in: International Conference Research on Computational Linguistics (ROCLING X). p. 9008+.
[58] Guo, X. et al. (2006) Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics 22 (8), 967–973.
[59] Tao, Y. et al. (2007) Information theory applied to the sparse gene ontology annotation network to predict novel gene function. Bioinformatics 23, i529–i538.
[60] Schlicker, A. et al. (2006) A new measure for functional similarity of gene products based on gene ontology. BMC Bioinformatics 7, 302.
[61] Wang, J.Z. et al. (2007) A new method to measure the semantic similarity of GO terms. Bioinformatics 23, 1274–1281.
[62] Xu, T., Du, L. and Zhou, Y. (2008) Evaluation of GO-based functional similarity measures using *S. cerevisiae* protein interaction and expression profile data. BMC Bioinformatics 9, 472.
[63] Jain, S. and Bader, G.D. (2010) An improved method for scoring protein–protein interactions using semantic similarity within the gene ontology. BMC Bioinformatics 11, 562.
[64] Reimand, J., Arak, T. and Vilo, J. (2011) g:Profiler – a web server for functional interpretation of gene lists (2011 update). Nucleic Acids Res. 39, W307–W315.
[65] Demir, E. et al. (2010) The BioPAX community standard for pathway data sharing. Nat. Biotechnol. 28, 935–942.
[66] Matthews, L. et al. (2009) Reactome knowledge base of human biological pathways and processes. Nucleic Acids Res. 37, D619–D622.
[67] Ge, H. et al. (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. Nat. Genet. 29, 482–486.
[68] Jansen, R., Greenbaum, D. and Gerstein, M. (2002) Relating whole-genome expression data with protein–protein interactions. Genome Res. 12, 37–46.
[69] Bhardwaj, N. and Lu, H. (2005) Correlation between gene expression profiles and protein–protein interactions within and across genomes. Bioinformatics 21, 2730–2738.
[70] Li, D. et al. (2008) PRINCESS, a protein interaction confidence evaluation system with multiple data sources. Mol. Cell Proteomic. 7, 1043–1052.
[71] Rhodes, D.R. et al. (2005) Probabilistic model of the human protein–protein interaction network. Nat. Biotechnol. 23, 951–959.
[72] Adler, P. et al. (2009) Ranking genes by their co-expression to subsets of pathway members. Ann. NY Acad. Sci. 1158, 1–13.
[73] Liu, C.T., Yuan, S. and Li, K.C. (2009) Patterns of co-expression for protein complexes by size in *Saccharomyces cerevisiae*. Nucl. Acids Res. 37 (2), 526–532.
[74] Adler, P. et al. (2009) Mining for coexpression across hundreds of datasets using novel rank aggregation and visualization methods. Genomebiology 10 (12), R139.
[75] Barash, Y. et al. (2010) Deciphering the splicing code. Nature 465, 53–59.
[76] Shen, J. et al. (2007) Predicting protein–protein interactions based only on sequences information. Proc. Natl. Acad. Sci. USA 104, 4337–4341.
[77] Pitre, S. et al. (2006) PIPE: a protein–protein interaction prediction enginebased on the re-occurring short polypeptide sequences between known interacting protein pairs. BMC Bioinformatics 7, 365.
[78] Sprinzak, E. and Margalit, H. (2001) Correlated sequence-signatures as markers of protein–protein interaction. J. Mol. Biol. 311, 681–692.
[79] Najafabadi, H.S. and Salavati, R. (2008) Sequence-based prediction of protein–protein interactions by means of codon usage. Genome Biol. 9, R87.
[80] Martin, S., Roe, D. and Faulon, J.L. (2005) Predicting protein–protein interactions using signature products. Bioinformatics 21, 218–226.
[81] Guo, Y. et al. (2008) Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Res. 36, 3025–3030.
[82] Roy, S. et al. (2009) Exploiting amino acid composition for predicting protein–protein interactions. PloS One 4, e7813.
[83] Yu, C.Y., Chou, L.C. and Chang, D.T. (2010) Predicting protein–protein interactions in unbalanced data using the primary structure of proteins. BMC Bioinformatics 11, 167.
[84] Sharan, R., Ulitsky, I. and Shamir, R. (2007) Network-based prediction of protein function. Mol. Syst. Biol. 3, 88.
[85] Goldberg, D.S. and Roth, F.P. (2003) Assessing experimentally derived interactions in a small world. Proc. Natl. Acad. Sci. USA 100, 4372–4376.
[86] Bader, J.S. et al. (2004) Gaining confidence in high-throughput protein interaction networks. Nat. Biotechnol. 22, 78–85.
[87] Yu, H. et al. (2006) Predicting interactions in protein networks by completing defective cliques. Bioinformatics 22, 823–829.
[88] Snel, B. et al. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res. 28, 3442–3444.
[89] Szklarczyk, D. et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. 39, D561–D568.
[90] Ramani, A.K. et al. (2005) Consolidating the set of known human protein–protein interactions in preparation for large-scale mapping of the human interactome. Genome Biol. 6, R40.
[91] Daraselia, N. et al. (2004) Extracting human protein interactions from MEDLINE using a full-sentence parser. Bioinformatics 20, 604–611.
[92] Arighi, C.N. et al. (2011) Overview of the BioCreative III workshop. BMC Bioinformatics 12 (Suppl 8), S1.
[93] Brown, K.R. and Jurisica, I. (2005) Online predicted human interaction database. Bioinformatics 21, 2076–2082.
[94] Beltrao, P. and Serrano, L. (2005) Comparative genomics and disorder prediction identify biologically relevant SH3 protein interactions. PLoS Comput. Biol. 1 (3), e26.
[95] Goh, C.S. and Cohen, F.E. (2002) Co-evolutionary analysis reveals insights into protein–protein interactions. J. Mol. Biol. 324, 177–192.
[96] Jothi, R. et al. (2006) Co-evolutionary analysis of domains in interacting proteins reveals insights into domain–domain interactions mediating protein–protein interactions. J. Mol. Biol. 362, 861–875.
[97] Pazos, F. and Valencia, A. (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Eng. 14, 609–614.
[98] Pazos, F. and Valencia, A. (2002) In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins 47, 219–227.
[99] Clark, N.L. and Aquadro, C.F. (2010) A novel method to detect proteins evolvingat correlated rates: identifying new functional relationships between coevolving proteins. Mol. Biol. Evol. 27, 1152–1161.
[100] Zarrinpar, A., Park, S.H. and Lim, W.A. (2003) Optimization of specificity in a cellular protein interaction network by negative selection. Nature 426, 676–680.
[101] Via, A. et al. (2009) A structure filter for the eukaryotic linear motif resource. BMC Bioinformatics 10, 351.
[102] Rost, B., Sander, C. and Schneider, R. (1994) PHD – an automatic mail server for protein secondary structure prediction. Comput. Appl. Biosci. 10, 53–60.
[103] Patil, A. and Nakamura, H. (2005) Filtering high-throughput protein–protein interaction data using a combination of genomic features. BMC Bioinformatics 6, 100.
[104] Lin, N. et al. (2004) Information assessment on predicting protein–protein interactions. BMC Bioinformatics 5, 154.
[105] Jansen, R. et al. (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. Science 302 (5644), 449–453.
[106] Scott, M.S. and Barton, G.J. (2007) Probabilistic prediction and ranking of human protein–protein interactions. BMC Bioinformatics 8, 239.
[107] Berman, H.M. et al. (2000) The protein data bank. Nucleic Acids Res. 28 (1), 235–242.
[108] Wang, X. et al. (2012) Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nature Biotechnol. 30 (2), 159–164.
[109] Kelley, B.P. et al. (2004) PathBLAST: a tool for alignment of protein interaction networks. Nucleic Acids Res. 32, W83–W88.
[110] Beltrao, P. and Serrano, L. (2007) Specificity and evolvability in eukaryotic protein interaction networks. PLoS Comput. Biol. 3, e25.
[111] Sun, M.G.F. et al. (2012) Signaling network evolution: rewiring and signatures of conservation in signaling. PLoS Comput. Biol. 8, e1.
[112] Kelley, B.P. et al. (2003) Conserved pathways within bacteria and yeast as revealed by global protein network alignment. Proc. Natl. Acad. Sci. USA 100, 11394–11399.
[113] Flannick, J. et al. (2006) Graemlin: general and robust alignment of multiple large interaction networks. Genome Res. 16, 1169–1181.
[114] Singh, R., Xu, J. and Berger, B. (2008) Global alignment of multiple protein interaction networks with application to functional orthology detection. Proc. Natl. Acad. Sci. USA 105, 12763–12768.

[115] Kuchaiev, O. et al. (2010) Topological network alignment uncovers biological function and phylogeny. J. R. Soc. Interface. 7, 1341–1354.
[116] Kitano, H. (2004) Biological robustness. Nat. Rev. Genet. 5, 826–837.
[117] Kellis, M., Birren, B.W. and Lander, E.S. (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428, 617–624.
[118] Krylov, D.M. et al. (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Res. 13, 2229–2235.
[119] Carlson, M.R. et al. (2006) Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics 7, 40.
[120] Dosztanyi, Z. et al. (2006) Disorder and sequence repeats in hub proteins and their implications for network evolution. J. Proteome Res. 5, 2985–2995.
[121] Ekman, D. et al. (2006) What properties characterize the hub proteins of the protein–protein interaction network of *Saccharomyces cerevisiae*? Genome Biol. 7, R45.
[122] Drummond, D.A., Raval, A. and Wilke, C.O. (2006) A single determinant dominates the rate of yeast protein evolution. Mol. Biol. Evol. 23, 327–337.
[123] Kim, P.M. et al. (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314, 1938–1941.
[124] Kim, P.M., Korbel, J.O. and Gerstein, M.B. (2007) Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. Proc. Natl. Acad. Sci. USA 104, 20274–20279.
[125] Bellay, J. et al. (2011) Bringing order to protein disorder through comparative genomics and genetic interactions. Genome Biol. 12 (2), R14.
[126] Barabasi, A.L. and Albert, R. (1999) Emergence of scaling in random networks, Science.
[127] Przulj, N., Corneil, D.G. and Jurisica, I. (2004) Modeling interactome: scale-free or geometric? Bioinformatics 20, 3508–3515.
[128] Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5, 101–113.
[129] Ravasz, E. and Barabasi, A.L. (2003) Hierarchical organization in complex networks. Phys. Rev. E Stat. Nonlin. Soft. Matter. Phys. 67, 026112.
[130] Karev, G.P. et al. (2002) Birth and death of protein domains: a simple model of evolution explains power law behavior. BMC Evol. Biol. 2, 18.
[131] Lee, H.J. and Zheng, J.J. (2010) PDZ domains and their binding partners: structure, specificity, and modification. Cell Commun. Signal 8, 8.
[132] Pawson, T. and Scott, J.D. (1997) Signaling through scaffold, anchoring, and adaptor proteins. Science 278 (5346), 2075–2080.
[133] Pawson, T. (1995) Protein modules and signalling networks. Nature 373 (6515), 573–580.
[134] Pawson, T. and Nash, P. (2000) Protein–protein interactions define specificity in signal transduction. Genes Dev. 14 (9), 1027–1047.
[135] Dev, K.K. (2004) Making protein interactions druggable: targeting PDZ domains. Nat. Rev. Drug Discov. 3, 1047–1056.
[136] Doorbar, J. (2006) Molecular biology of human papillomavirus infection and cervical cancer. Clin. Sci. 110, 525–541.
[137] Moyer, B.D. et al. (1999) A PDZ-interacting domain in CFTR is an apical membrane polarization signal. J. Clin. Invest. 04, 1353–1361.
[138] King, M.C., Marks, J.H. and Man dell, J.B. (2003) Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science 302, 643–646.
[139] Amberger, J. et al. (2009) McKusick's online Mendelian inheritance in man (OMIM). Nucleic Acids Res. 37, D793–D796.
[140] Stenson, P.D. et al. (2009) The human gene mutation database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. Hum. Genomics 4 (2), 69–72.
[141] Forbes, S.A. et al. (2011) COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res. 39, D945–D950.
[142] Wang, Z. and Moult, J. (2001) SNPs, protein structure, and disease. Hum. Mutat. 17, 263–270.
[143] Reva, B., Antipin, Y. and Sander, C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 9, ell8.
[144] Mooney, S. (2005) Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief Bioinform. 6, 44–56.
[145] Ng, P.C. and Henikoff, S. (2006) Predicting the effects of amino acid substitutions on protein function. Annu. Rev. Genomics Hum. Genet. 7, 61–80.
[146] Schuster-Böckler, B. and Bateman, A. (2008) Protein interactions in human genetic diseases. Genome Biol. 9 (1), R9.
[147] Teng, S. et al. (2009) Modeling effects of human single nucleotide polymorphisms on protein–protein interactions. Biophys. J. 96, 2178–2188.
[148] McLendon, R. et al. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455, 1061–1068.
[149] Bell, D. et al. (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474, 609–615.
[150] Bech-Otschir, D. et al. (2001) COP9 signalosome-specific phosphorylation targets p53 to degradation by the ubiquitin system. EMBO J. 20, 1630–1639.
[151] Sharma, S.V. et al. (2007) Epidermal growth factor receptor mutations in lung cancer. Nat. Rev. Cancer 7, 169–181.