Phys. Biol. **2** (2005) S1–S16 doi:10.1088/1478-3975/2/2/S01

Prediction of physical protein–protein interactions

András Szilágyi, Vera Grimm, Adrián K Arakaki and Jeffrey Skolnick

Center of Excellence in Bioinformatics, University at Buffalo, State University of New York,
901 Washington St, Buffalo, NY 14203, USA

E-mail: skolnick@buffalo.edu

Received 14 December 2004
Accepted for publication 2 February 2005
Published 19 April 2005
Online at stacks.iop.org/PhysBio/2/S1

Abstract

Many essential cellular processes such as signal transduction, transport, cellular motion and most regulatory mechanisms are mediated by protein–protein interactions. In recent years, new experimental techniques have been developed to discover the protein–protein interaction networks of several organisms. However, the accuracy and coverage of these techniques have proven to be limited, and computational approaches remain essential both to assist in the design and validation of experimental studies and for the prediction of interaction partners and detailed structures of protein complexes. Here, we provide a critical overview of existing structure-independent and structure-based computational methods. Although these techniques have significantly advanced in the past few years, we find that most of them are still in their infancy. We also provide an overview of experimental techniques for the detection of protein–protein interactions. Although the developments are promising, false positive and false negative results are common, and reliable detection is possible only by taking a consensus of different experimental approaches. The shortcomings of experimental techniques affect both the further development and the fair evaluation of computational prediction methods. For an adequate comparative evaluation of prediction and high-throughput experimental methods, an appropriately large benchmark set of biophysically characterized protein complexes would be needed, but is sorely lacking.

1. Introduction

In the highly crowded environment of a living cell (figure 1), biological macromolecules occur at a concentration of 300–400 g l^{-1} and they physically occupy a significant fraction (typically 20–30%) of the total volume. Most proteins interact, at least transiently, with other protein molecules; indeed, many essential cellular processes such as signal transduction, transport, cellular motion and most regulatory mechanisms are mediated by protein–protein interactions.

Given their biological importance [1], the development of methods to detect and characterize protein–protein interactions and assemblies is a major theme of functional genomics and proteomics efforts [2, 3]. As discussed in further detail below, currently, two main types of experimental methods are used to detect such interactions: the yeast two-hybrid screen (*Y2H*) [4], which is mainly limited to the detection of binary interactions, and the combination of large-scale affinity purification with mass spectrometry (MS) to detect and characterize multiprotein complexes [5–7]. First applied to yeast [8–11], these methods revealed the dense network of interactions linking proteins in the cell, but their error rate is high [12]. The coverage of *Y2H* screens seems incomplete, with many false negatives and false positives as evidenced by the limited overlap between sets of interacting proteins identified by different groups [10] and between those identified by *Y2H* and other approaches [13]. For yeast, there are several efforts to assemble a consistent network of reliable interactions from protein–protein interaction data sets produced by different methods [14–16]. There is clearly the need to develop large-scale benchmark sets of interacting proteins that have been experimentally validated by biophysical methods such as ultracentrifugation or light scattering.

This discrepancy among experimental methods has prompted keen interest in the development of computational methods for inferring protein–protein interactions [17–19].

1478-3975/05/020001+16$30.00 © 2005 IOP Publishing Ltd Printed in the UK

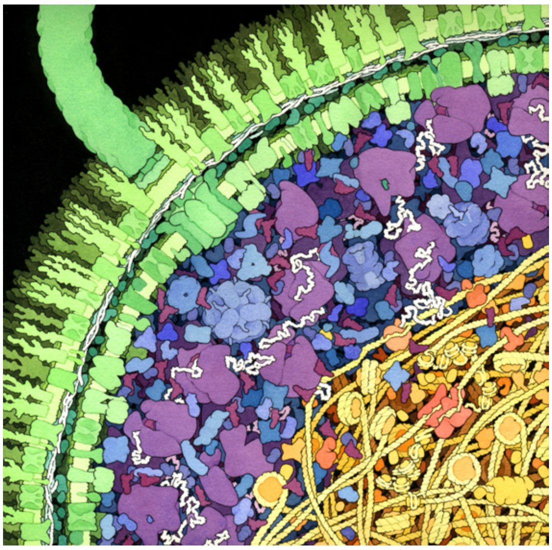

Figure 1. Representation of the approximate numbers, shapes and density of packing of macromolecules inside a cell of *Escherichia coli*. (Illustration by David S Goodsell; reprinted with permission.)

Many consider protein–protein interactions in the most general context and often refer to 'functionally interacting proteins' [19], implying that the proteins cooperate to carry out a given task without actually (or necessarily) engaging in physical contact. Other methods attempt to predict direct physical interactions between proteins. Such approaches range from the prediction of the binding interface without the prediction of the full three-dimensional quaternary structure to techniques that provide such quaternary structure predictions. In what follows, we flesh out these ideas as well as describe in additional detail the state-of-the-art of various high-throughput experimental approaches. The prediction of direct, physical interactions is the main focus of the present review.

1.1. Types of protein–protein interactions

Protein–protein interactions can be classified in various ways such as homo- versus hetero-oligomeric, obligate versus non-obligate and transient versus permanent. However, the boundaries between these classes are blurred and protein interactions can be regarded to span a continuous 'interaction space' rather than a set of discrete classes.

Many proteins form strong, stable interactions, giving rise to permanent protein complexes. Because these complexes are much easier to study, most of the available experimental data (such as x-ray structures) have been obtained from stable complexes. However, transient protein–protein interactions are equally important: they play a major role in signal transduction, electron cascades and other essential physiological processes. Nooren *et al* distinguish between 'weak' transient complexes that exist *in vivo* in an equilibrium of different oligomeric states and 'strong' transient complexes with binding affinities in the nanomolar range that dissociate only upon triggering [20]. Since transient interactions often neither form stable crystals nor give good NMR structures, transient complexes are notoriously hard to study experimentally. This is also reflected in the small number of validated complexes found by Nooren and Thornton [20] (weak: 16, strong: 23).

1.2. Inference of interacting sites and interfaces

One type of prediction approach addresses the following question: given the sequence or the structure of a protein, which regions or residues are likely to be parts of its interface with another protein?

Knowing where the binding region of a protein is located can help in guiding both experiments and other types of predictions. For example, mutagenesis experiments can be designed to pinpoint functionally important residues of receptors and other binding proteins. Information on likely binding sites can even be a starting point for drug design when the given interaction needs to be inhibited or mimicked [21]. On the other hand, when the prediction of the structure of a complex based on the structures of the component proteins (i.e. protein–protein docking) is desired, knowledge of the binding regions can be used to reduce the size of the configuration space to search. As evidenced by assessments of blind prediction experiments such as critical assessment of predicted interactions (CAPRI) [22], this reduction is extremely helpful for docking, and the success or failure of the procedure often depends on having some knowledge (either from biochemical experiments or prediction) of the interacting regions.

The basis of methods for predicting the interfacial residues from protein sequence alone is the somewhat controversial concept that residues at protein–protein interfaces are more conserved across different protein families than other surface residues. Earlier studies, based on only a small number of complexes, supported this hypothesis. Recently, Caffrey *et al* [23] have tested this approach on an expanded, non-redundant set of 64 protein–protein interfaces. They found that even though individual residues at the protein interface are usually more conserved than other surface residues, if the analysis is performed by examining candidate surface patches, then the difference in conservation scores between actual interface patches and other patches becomes too small to allow prediction of the interface by conservation alone. The most conserved surface patch has an average overlap of only 36–39% with the actual interface. Another result of this study is that obligate interfaces differ from transient ones in two aspects: they have significantly fewer alignment gaps at the interface than the rest of the protein surface, and their buried interface residues are more conserved than the partially buried ones. Even though residue conservation is insufficient for predicting interfaces, there is the hope that it can be useful for prediction if applied together with other information such as phylogenetic relationships [24] as well as residue propensities [25] and physical properties [26].

When the structure of the individual molecules is known or can be reliably predicted, then one can utilize knowledge from numerous observations regarding the nature of protein–protein interfaces to predict the interacting regions. Simple principles of protein–protein recognition such as complementarity of shape, electrostatic interactions and hydrogen bonding have long been recognized [27, 28]. In about one-third of the interfaces, a recognizable hydrophobic core is found, surrounded by inter-subunit polar interactions; the rest of the interfaces show a varied mixture of small hydrophobic patches, polar interactions and sometimes water molecules scattered over the interfacial area [29, 30]. The amino acid composition of interfaces is characteristic, and different types of interfaces (such as domain–domain, homo- or hetero-oligomeric and permanent or transient) can often be distinguished from each other using the observed residue frequencies alone [26, 31–33]. It is interesting to note that different studies often report slightly different or even contradictory results, in part, depending on whether they investigate interfaces as contiguous surface patches or just define the interface as the set of individual residues in contact with another subunit (see [33] for some critical notes on the 'surface patch' approach).

It has long been recognized that some residues within the binding interface make a dominant contribution in the stabilization of protein complexes. These residues, defined by having a significant drop in the binding affinity when mutated to alanine are called 'hot spots'. It has been shown that hot spots correlate well with residue conservation [34, 35]. Recently, Halperin *et al* [36] demonstrated that both experimental hot spots and conserved residues tend to couple across the protein–protein interface, and the local packing density tends to be higher (about as high as within protein cores) around them than at other spots within the interface. Favorable conserved pairs include glycine coupled with aromatic, charged and polar residues, as well as aromatic residue coupling; on the other hand, charged pairs were under-represented. These results deepen our understanding of the nature of protein–protein interfaces and can lead to improved prediction methodologies.

1.3. Inference of interacting partners

Another type of question is the following: 'Given a set of protein sequences (or structures), which pairs of proteins are likely to have interactions?' Our goal in asking a question like this is to reconstruct the protein–protein interaction network for a set of proteins; ideally, we would like to extend the analysis to the whole proteome of an organism. The network of all interactions within an organism (not necessarily limited to protein molecules) is sometimes called the *interactome*. While functional linkages between proteins (as inferred by various methods for genome analysis) can often suggest direct, physical interactions between them [19, 37, 38], functional linkage is clearly a broader concept and does not necessarily involve direct physical interaction. Evidence of direct binding, however, is a good indication of functional relatedness, and therefore, knowledge of the interactome is a significant step toward understanding the functional organization of the cell.

In recent years, high-throughput experimental methods such as the yeast two-hybrid method and mass spectrometry have been used to elucidate the protein–protein interaction network of several organisms [8, 10, 11, 39, 40], even though the accuracy of these methods is often lower than expected and often the conclusions are inconsistent [41]. Nevertheless, the resulting data sets have been subject to intensive analysis. In particular, the topologies of the networks have been studied in great detail, and they were found to be small-world, scale-free and modular [42].

Because the experimental data on interaction networks are known for only a few organisms, it is an important question whether interaction annotation can be transferred from one organism to another. It turns out that protein–protein interactions can readily be transferred when a pair of proteins has a joint sequence identity of $>80\%$ or a joint E-value $<10^{-70}$ [43]. Based on this finding, an online database of interologs (orthologous pairs of interacting proteins) has been created [43]. Other observations that can form the basis of interaction predictions include the following: (1) proteins that can functionally substitute for one another tend to have anti-correlated distribution patterns across organisms [44]; (2) interacting proteins tend to exhibit similar phylogenetic trees [45]; (3) the interaction network has certain conserved motifs [46]; (4) interacting proteins tend to have similar phylogenetic profiles and a similar gene neighborhood; (5) they tend to be involved in gene-fusion events and (6) their co-evolution leads to some identifiable correlated mutations in their sequences [38].

Prediction approaches based on sequence and genome analysis do not always provide fully reliable answers regarding the presence or absence of a putative interaction. In these cases, looking at the structural details of the putative interaction using an experimentally determined or even a predicted structure can be of help. This leads to another class of interaction prediction methods: those that use a structure-based approach [40].

2. Structure-independent methods of protein–protein interaction prediction

In this section, we focus on structure-independent methods for the prediction of protein–protein interactions that are based on *a priori* biological knowledge. Thus, methods that rely heavily on experimental data (e.g. learning features from known protein interaction partners) are not discussed here.

2.1. Methods based on gene context conservation

The conservation of different types of genomic context information that can be extracted from the comparative analysis of genomes can be used to predict functional interactions between gene products [37]. The application of this type of method to the prediction of protein–protein interactions suffers from the problem that functional interaction does not necessarily imply direct physical interaction [19]. Recently, new ways of exploiting genomic context for the prediction of functional associations between proteins have been developed, but their correlation with direct physical interactions has not been investigated [47].

2.1.1. Co-occurrence of genes in related species (phylogenetic profiles). The underlying hypothesis of the phylogenetic profile method is that functionally linked proteins co-evolve. If this hypothesis is true, then these proteins should have homologs in the same subset of organisms. In the phylogenetic profile method, each analyzed protein is represented as a string of bits (a phylogenetic profile), indicating the presence or absence of homologs of the given protein in a set of genomes [17, 48]. Proteins exhibiting similar phylogenetic profiles are predicted to be functionally linked, i.e. they participate in a common structural complex or metabolic pathway. Recently, Wu *et al* extended the method by taking into account the probability that a given arbitrary degree of similarity between two profiles would occur by chance. This extension allows inference to be done at any desired level of confidence [49].

In their evaluation of methods of protein function prediction by genomic context, Huynen *et al* found direct physical interactions for only 34% of the *Mycoplasma genitalium* proteins analyzed by phylogenetic profiles [50]. The main disadvantages of the phylogenetic profile method are that only complete genomes can be used as input (because that is the only way to be certain that homologs of a given gene are absent in a given organism) and that somewhat arbitrary thresholds must be set to dictate whether the homolog is present. Moreover, given the observation that proteins can be homologous but have different function, the presence of a homolog does not guarantee that the specified function is conserved across the set of organisms. Perhaps the method could be improved by restricting the bit strings to sets of conserved orthologs.

2.1.2. Conservation of local genomic context. This method is based on the fact that neighboring genes in bacterial and archaeal genomes tend to encode proteins that show physical or functional interactions with each other. Conservation of the genomic context across different genomes can be detected based on (1) analysis of gene order and operon architecture [51] or (2) analysis of gene clusters, defined as sets of genes that occur on the same DNA strand and have gaps between the adjacent genes of 300 base pairs or less [52].

Huynen *et al* found that ∼63% of proteins encoded by gene pairs conserved as neighbors in phylogenetically distant genomes physically interact, either directly (30%) or indirectly (33%) [50]. The main limitation of methods based on the conservation of the local genomic context is that they cannot be applied to eukaryotes, where, with only a few exceptions, genes appear to be randomly distributed [53]. Also, genome annotation errors such as incorrect assignment of translation starts, frame shifts and missed or incorrectly predicted genes complicate the comparative analysis of gene orders [54].

2.1.3. Gene fusion analysis. Functional interaction can be inferred from the presence of proteins in an organism that have homologs in another organism fused into a single protein. The existence of a fusion protein in one genome (called a 'Rosetta Stone sequence' [55] or a 'composite protein' [56]) allows the prediction of the interaction between the single-domain proteins in other genomes, even when they are not encoded by neighboring genes. The essential assumption of this method, i.e. genes linked by fusion are at least functionally related, has been validated by the analysis of several complete genomes [50, 57]. Enright *et al* employed BLAST [58] comparisons to establish orthology between proteins in the query and the reference genomes, and Marcotte *et al* used the Pfam [59] and ProDom [60] databases with the same purpose. Recently, Truong and Ikura have proposed a method for domain fusion analysis that does not directly rely on sequence comparison and can be applied to large non-redundant databases [61]. They start with Pfam domain assignments of each protein in Swiss-Prot + TrEMBL [62] and then apply successive relational algebra operations to identify putative functional linkages.

Huynen *et al* found evidence of direct physical interaction for 55% of the *M. genitalium* proteins analyzed by gene fusion [50]. The main disadvantage of gene-fusion analysis is that this approach is limited by the number of fusion events, which varies for different types of genes. For instance, certain structural and functional groups, e.g. proteins with an alpha/beta fold [63] have a higher propensity to be involved in gene-fusion events. Metabolic enzymes exhibit a tendency to participate in multiple gene-fusion events that is three times higher than a protein selected at random [64].

2.2. Methods based on co-evolution of interacting proteins

Physically interacting proteins generally evolve in a coordinated fashion that preserves relevant contacts between them [65]. Thus, methods based on this principle are more likely to predict relationships between proteins, which are not only functional but also reflect direct physical interactions.

2.2.1. Phylogenetic tree similarity. Phylogenetic trees of interacting proteins have a higher degree of similarity than those constructed based on non-interacting proteins. On the basis of this concept, Goh *et al* evaluated the similarity of phylogenetic trees of the two domains of phosphoglycerate kinase as the linear correlation between the pairwise distance matrices used to build the trees [66]. They used the co-evolution of the two domains of phosphoglycerate kinase to quantify the co-evolution of chemokines and their receptors. This approach was extended to different test sets by Pazos and Valencia, who proposed a more general application of the approach for predicting protein interactions, including a rigorous statistical evaluation [67]. However, while their approach was able to identify protein families that interact, it could not identify specific interacting partners between the two protein families because the authors only incorporated one homologous protein per organism. In recent modifications of this method, this problem has been solved in a few specific cases by analyzing the correlation between sequence similarity distance matrices constructed for protein families [45, 68]. A limitation of methods based on phylogenetic tree similarity is that a good quality multiple sequence alignment including homologs of the two proteins from the same set of organisms is required.

2.2.2. Differential accumulation of correlated mutations (in silico two-hybrid method). In an early series of papers, Pazos *et al* employed correlated mutation analysis of multiple sequence alignments to detect sets of residues that interact across protein interfaces [65]. They also predicted interdomain contact regions of heat-shock protein Hsc70. This work was subsequently followed up by an approach, named the '*in silico* two-hybrid method', that can identify putative partners as well as the regions of the sequence that might interact [18]. The procedure consists of a search algorithm to find pairs of multiple sequence alignments with a distinctive co-variation signal. The method is based on the hypothesis that co-adaptation of interacting proteins can be detected by the presence of a particular number of compensatory mutations in the corresponding homologs of different species. Significant results using this method have been reported in a number of cases including the use of this approach to construct a 'complete' interaction network in *E. coli*. The main advantage of the *in silico* two-hybrid method is that it indicates not only a possible interaction, but also the possible protein region involved; this information can be used to guide quaternary structure prediction algorithms, e.g. protein–protein docking simulations.

2.2.3. Co-evolution of gene expression. Based on the observation that interacting proteins are frequently co-expressed to maintain the correct stoichiometry among interacting partners, Fraser *et al* have shown that the expression co-evolution can be used for the computational prediction of protein–protein interactions [69]. They find that the co-evolution of expression in yeast is a more powerful predictor of physical interaction than is the co-evolution of amino acid sequence. A limitation of this approach is that it may not be easily applicable to organisms with insignificant codon bias due to gene expression levels.

3. Structure-based methods of protein–protein interaction prediction

3.1. Modeling of protein–protein interactions by homology

Protein–protein interactions can be modeled by similarity, using known structures of protein complexes whose components are homologous or similar to the proteins whose interactions are to be modeled.

In a straightforward extension of the MODELLER homology modeling technique [70], known structures of protein complexes were used to evaluate the inter-subunit interactions in putative complexes comprising homologs of each subunit, using a scoring function based on the propensities of residue pairs to span protein–protein interfaces [71]. Using this technique, ~30 000 links between pairs of ~10 000 modeled sequences were predicted, with an estimated false positive rate of 25%. These predicted links have been included in the MODBASE database, which contains homology models for ~660 000 sequences in the Swiss-Prot/TrEMBL database [71].

In a similar effort, Aloy and Russell [72] described a method to test putative interactions on complexes of known structure. Given a 3D complex and alignments of homologues of the interacting proteins, they used an empirical potential to assess the fit of any possible interacting pair on the complex. In an evaluation of the method, all known complexes gave significant scores except for peptidase/inhibitor complexes, which are known to interact via many main-chain to main-chain contacts. The method, named 'interaction prediction through tertiary structure' was later made available as an online server [73]. Using this approach, combined with screening by electron microscopy (EM), models (partial or complete) were built for 54 protein complexes in yeast and a structure-based network of molecular machines was constructed [74].

The assumption behind these homology-based approaches is that interaction information can be extrapolated from one complex structure to homologs of the interacting proteins. Indeed, it has been demonstrated that close homologs (30–40% or higher sequence identity) almost always interact in the same way (i.e. the RMSD between the corresponding interacting regions is low) [75]. Similarity only in fold (without additional evidence for a common ancestor) was found to be only rarely associated with a similarity in interaction. This suggests that there is a twilight zone of sequence similarity where the interaction may or may not be similar. Threading-based techniques can be used to handle sequences in this 'twilight zone'.

3.2. Threading-based method: MULTIPROSPECTOR

To capture more distantly related or even analogous proteins, the idea of multimeric threading was introduced [76], extending fold recognition approaches [77] to multiple chains. For dimeric threading, two target sequences are threaded onto each structure in a representative, non-redundant library of folds. Templates that match one of the query sequences with a *Z*-score higher than a threshold are collected. The template structures in the two resulting sets are then examined. If two templates (one from each set) form a dimer, then each of the two target sequences is threaded onto its template in the presence of the other subunit. The structure–sequence alignments are optimized during this double chain threading using a knowledge-based interfacial potential [78], thereby allowing the predicted structure of the complex to influence the individual sequence–structure alignments.

The template pairs with the highest *Z*-score (energy in standard deviation units relative to the mean) and the lowest interfacial energies are subjected to further filtering. If no alternative monomeric structure with a higher *Z*-score can be found for any of the sequences and the interaction energy is below a certain threshold and the *Z*-scores for the complex are above 5.0, then this pair of sequences is predicted to form a dimer.

The approach was tested on a benchmark set of true monomers, homo- and heterodimers with excellent results. When MULTIPROSPECTOR was tested on 2457 known interactions of yeast proteins, 144 were correctly identified; this small number clearly shows the limitations due to the fact that an existing template structure is needed for correct predictions. MULTIPROSPECTOR has been applied on a genomic scale to the *Saccharomyces cerevisiae* proteome [79]. This study yielded 7321 predicted interactions, with only 374 found in experimental yeast interaction data. However, the overlap between different large-scale experimental sets is similarly small [79].

3.3. Computational protein–protein docking

3.3.1. Overview. Docking aims to predict the native three-dimensional structure of a multimeric protein complex given the atomic coordinates of its constituent proteins. The docking procedure is, in general, facilitated by prior knowledge, e.g. knowing the binding site will help in restricting the search space significantly (for an overview see [80]). Often, proteins had been studied for a long time before their structures were solved; therefore, the residues involved in their binding are known or their identification is easy, as e.g. for serine proteases and antibodies. As mentioned above, certain types of residues, called 'hot spots', have a major contribution to the binding energy [81, 82], and prior knowledge of them can facilitate the search. Correlated hot spots are more conserved than expected from a random distribution and might be identified by theoretical methods [36]. Known structures of homologous complexes are also helpful: in the CAPRI competition, excellent results were obtained in docking a superantigen toxin to the beta-chain of the T-cell receptor (TCR) because the complex had a close homolog which was used for comparative modeling [83]. NMR restraints were also combined with docking methods to restrict the search space [84]. However, due to the advent of structural genomics, often the sequence and three-dimensional structure of a protein may be the first information obtained, without any experimental data on function or binding. As the sampling of the 'interface space' by the known complex structures is still sparse [85], there is a large number of 'zero-knowledge' cases for which docking could generate putative structures for complexes of proteins assumed to be in contact. All docking methods rely on the assumption that interacting proteins have a certain degree of shape complementarity, a notion first formulated by Emil Fischer in 1895 to explain the substrate binding of enzymes. While observations for many protein complexes for which atomic structures could be obtained show a high level of similarity between the bound (i.e. in the complex) and unbound structures, most proteins undergo small to large conformational changes upon binding, commonly known as 'induced fit', the prediction of which has proved to be one of the greatest challenges in docking.

3.3.2. Representation of the protein surface. One crucial component of most docking algorithms is the choice of the computational representation of the protein's surface, depending on the sampling strategy used and the features to be correlated. Few methods use the atomic structure directly [86]. Many methods represent the structure either by mapping it to a grid [87] or by spherical harmonic expansions, while others take only a set of points ('sparse critical points') [88] based on the Connolly surface [89] and represent the surface as triangles with their normal vectors attached. Any of these surface representations can additionally be 'softened' to allow for flexibility of the side-chains. Long and flexible or incorrectly positioned side-chains within the interface can prevent successful docking. Several cases, such as kallikrein A and bovine pancreatic trypsin inhibitor (BPTI), are known to be particularly difficult docking candidates because of protruding amino acids in the interface. Some of these residues appear to be 'key' or 'anchor' side-chains, interacting with structurally constrained pockets, while others, mostly on the periphery of the binding pocket, show 'induced fit' behavior [90, 91]. Another approach to handle protruding side-chains is their truncation. Gabb *et al* and Chen *et al* reported that this solution unfavorably affected their results [92, 93], while altering the geometric weight of grid cells for the most variable side-chains, e.g. lysine improved the outcome in nearly all test cases [94]. Other methods of surface softening include a low-resolution docking method [95, 96], the use of a simplified model for selected side-chains [97], and the thickening of the surface layer [98]. All modern docking techniques use some approximate solution to handle side-chain flexibility.

3.3.3. Search algorithm. Even using a rigid body approximation, the remaining six-dimensional conformational search space is large. Due to new algorithms and increasing computer power, it is sometimes possible to perform an exhaustive search. The search scheme chosen is directly dependent on the type of surface representation. For grid representations, the most popular strategy is based on Fourier correlation. Introduced in 1992 by the groundbreaking work of Katchalski and coworkers, this technique allows the calculation of correlations between the points of two grids simultaneously for all possible translations, leading to a considerable speed-up of the search [87]. It has been implemented in a variety of docking programs including FTDock/3D-Dock [92, 99], GRAMM [95], DOT [100] and ZDock [101]. A similar Fourier-based approach for the fast calculation of correlations using spherical harmonics has been implemented in the program HEX [102]. Some other algorithms are also capable of searching the entire rotational and translational space, notably the matching of surface cubes [103], genetic algorithms (GAPDOCK [104], DARWIN [105]) as well as methods based on Boolean operations (BiGGER [106]) or a pseudo-Brownian Monte Carlo approach (ICM [86]). Sampling the conformational space evenly should yield (at least in theory) several near-native protein orientations besides millions of incorrectly docked conformations. To eliminate the incorrectly docked complexes, filtering is applied, exploiting the expected complementarities between the two (or more) molecules. Geometric fit alone is not capable of distinguishing near-native from non-native complexes. Protein–protein interfaces vary widely in shape, size, amino acid content, hydrophobicity, electrostatics and other features [28, 31, 107]. The 'complementarity idea' has therefore been extended to coupled electrostatic fields [92, 108] and hydrophobic complementarities [109, 110]. For several complexes, such as the trypsin-BPTI [111] and the barnase–barstar system [112], electrostatics is the major contributor to the binding process and stability. Other complexes as elastase–OMTKY or chymoptrypsin–OMTKY show desolvation-driven complex formation [113]. Consequently, desolvation effects have been taken into account in many algorithms. One quite successful example is the atomic contact energy (ACE) model [114]. This approach involves replacing atom–atom contacts by atom–water contacts and has been implemented in ZDOCK [93] and other algorithms [115]. Apart from complementarities, knowledge-based potentials are commonly used to discriminate native conformations from non-native ones.

3.3.4. Refinement. After filtering out the incorrectly docked structures, a small number of candidate models remain. At this stage of the docking procedure, neglected side-chain flexibility can be re-introduced, and a subsequent refinement step might improve the model. Methods for refinement include a biased probability side-chain optimization method (ICM [86]) or side-chain minimization (Multidock [99]). The simultaneous correction of main-chain displacements seems to be quite successful for small main-chain movements (ROSETTA [116]). Algorithms taking side-chain flexibilities into account performed slightly better in CAPRI rounds 1 and 2 [83], compared to methods without any flexibility. However, they failed to provide correct predictions for weakly binding complexes and complexes with large backbone displacement between bound and unbound states.

3.3.5. Main-chain flexibility. The correct prediction of protein–protein orientations with substantial backbone displacement between bound and unbound forms, as seen in several transient complexes in signal transduction, seems to be impossible using a ‘rigid body’ type approach. All docking methods failed to predict a homodimer with a backbone displacement of $\sim$12 Å in the third CAPRI round [117], and only a few acceptable results were obtained for the complex between the protein kinase from *Lactobacillus casei* (Hpr-K) and its substrate (P-Hp), with a difference of 2 Å between the bound and unbound forms [83]. For many enzyme–inhibitor complexes, the ‘rigid body assumption’ might yield reasonable docking results even if the complex obeys an induced-fit recognition mechanism [118]. Only a few methods try to include main-chain flexibility directly in the calculations. Sandak *et al* studied the docking of Calmodulin with its M13 peptide, allowing for domain or substructural movement in the receptor or the ligand structure [119, 120]. Several experimentally confirmed binding modes could be reproduced remarkably well. For complexes where the hinge region is known from structural comparisons or experimental data, this method has the potential to provide good predictions.

3.3.6. Type of complex. The success of a docking experiment is also dependent on the type of complex to be predicted. Vajda *et al* introduced a classification scheme for protein–protein complexes, based on the level of difficulty to find the native conformation by means of docking [117]. According to this scheme, complexes with major backbone displacement are as nearly unpredictable as transient interactions. Stable enzyme–inhibitor systems with evolutionarily optimized interfaces [30, 121] are, in general, ‘easy cases’ whose structures are usually found independently of the docking method with high accuracy [122]. On the other hand, antibody–antigen systems could be called ‘hard cases’: first of all, the solved crystal structures do not resemble real-world scenarios since most are high-affinity antibodies designed for a specific purpose. In addition, the interface of those complexes often has grooves and deep pockets, in good agreement with the idea that the backbone geometry in these complexes is not as optimized as, e.g., in serine proteases. The notoriously hard to predict transient interactions are in the same difficulty class [117].

3.3.7. Improvements. Recent interesting developments include a structure-based method to identify interfacial residues by means of docking followed by an analysis of enriched low-energy conformations [123]. For some complexes, e.g. the CAPRI target T07 (TCR beta chain–SpeA complex), a binding site different than the experimentally determined one was found. In this example, the predicted site was located in the interface between the TCR alpha and the TCR beta chain (PDB-entry 1tcr). Another interesting approach is the incorporation of external, e.g. biochemical, data directly into the conformational search by up- or down-weighting certain intermolecular residue contacts [124].

3.3.8. Genome-scale docking. Although interesting lessons can be learned from docking experiments with single protein–protein systems, the docking-based prediction of protein interactions on a genomic scale is an even more exciting undertaking. One of the few approaches in this direction is the docking of very approximate molecular protein structures [125] based on Vakser’s low-resolution docking. Another interesting study described the successful reconstruction of homo-tetramers from comparative models of a single subunit using docking and comparative modeling techniques [126]. Obviously, the performance of these methods, when applied on a genomic scale, is still far behind that of sequence-based, large-scale interaction prediction.

3.4. Other structure-based methods

Several methods have been developed for the structure-based prediction of protein binding sites. Many of these approaches just predict functionally important sites; here, we only mention those that specifically focus on the residues involved in protein–protein binding. Relying on the finding that there are structurally conserved residues in binding sites [35], neural networks trained with a reduced representation of the interacting patch and sequence profile were able to detect 73% of the residues at protein–protein interfaces in a test set [127]. In an extension of the evolutionary trace approach, Aloy *et al* [128] developed an automated method that maps invariant polar residues in a multiple sequence alignment onto a protein structure and identifies spatial clusters of these residues as being putative functional sites. The procedure proved useful for filtering putative complex structures obtained by protein–protein docking. Using a related approach, Landgraf *et al* [129] defined two types of scores based on spatial clusters of residues and an associated multiple alignment and found that (1) a 'regional conservation score' is useful for identifying functional residue clusters as well as for the prediction of poorly conserved, transient protein–protein interfaces; (2) a 'similarity deviation score' is useful for finding specificity-conferring regions.

Given the structure of a complex, hot spots have been predicted using a simple physical model [130]. In a test of this 'computational alanine-scanning' procedure, 79% of hot spots were correctly predicted [131] and this procedure formed the basis of a successful redesign of a number of protein–protein interfaces [132].

3.5. Interfacial potentials

Interfacial potentials are used in most structure-based methods to evaluate prospective protein–protein conformations. They are based on the idea that energy-like parameters such as free energy should discriminate native from non-native conformations. The native complex structure is thought to be at the global thermodynamic minimum [133, 134] of the free-energy function. However, calculating the free energy is complicated. Knowledge-based potentials [135–137], also called statistical effective energy functions [138], have become increasingly popular; easy and fast to calculate, they have been incredibly successful in protein fold recognition [139–142], structure prediction [78, 143, 144] and other fields. They can be built at various levels of detail and can be atomic or residue-based. By comparing a given feature (e.g. side-chain contacts) to a reference state where the putative interaction is assumed to be absent, it can be turned into an energy-like quantity. The physical basis of this approach is somewhat controversial [145–148], mainly because of the construction of a so-called reference state which is essential for the quality of the derived potential [149]; however, a good correlation between such knowledge-based and physics-based potentials has been observed [150].

4. Structure-based versus structure-independent methods

4.1. Advantages and disadvantages of structure-based methods

Even though protein–protein interactions can often be inferred from sequence information and genome analysis alone, it is ultimately the fine atomic details of an interaction that determine the binding affinity and the specificity of binding one biomolecule to another. Structure-based methods analyze protein–protein interactions at this level, and therefore have the potential to be more accurate and decisive than methods that do not use structure information. In addition, they have the ability to provide predicted structures of complexes [74] which can be essential for understanding the function of molecular machines [151].

However, structure-based methods also have several disadvantages, the main one being the lack of structural templates for most types of interactions. The number of structurally distinct interactions is estimated to be ∼10 000, of which less than ∼2000 are known [152]. Extrapolating the current growth rate (200–300 new interactions per year), it would take two decades or more before most interaction types are known [152], although proposed initiatives for structural genomics of complexes [153] are expected to speed up this process. Still, for the time being, homology-based complex modeling suffers from the lack of templates for many protein pairs, which means that considerably less reliable, *ab initio* types of methods such as protein–protein docking have to be applied.

Another limitation is that it is difficult to crystallize weakly interacting complexes. It is widely believed that the majority of functional protein–protein interactions are transient and do not form complexes stable enough for crystallization or even NMR studies [154]. Therefore, by using these techniques, we may never be able to obtain experimental structures for some of the most important protein–protein interactions.

Structure-based methods critically depend on the energy functions used to evaluate proposed conformations, and, when applicable, on the algorithm used to sample conformational space. In nature, some interactions show overlapping specificities [155] while others are remarkably specific [156]. Recent success in designing protein–protein interaction specificity [132] suggests that, despite some shortcomings, energy functions capable of reproducing the specificity of protein–protein binding already exist, and structure-based computational methods can now be used to modulate and reengineer protein–protein interaction networks in living cells [157].

4.2. Utility of protein complex models

Protein structures and knowledge of the interactions between specific proteins are essential to understand the molecular mechanisms of biological systems. In general, proteins do not act in an isolated manner; instead, they are organized in multiprotein complexes, whether permanent or transient, that allow them to perform essential roles in all kinds of biological processes. An important practical motivation for the determination of new protein structures and their complexes is the fact that the cause of many genetic diseases can be traced back to deficiencies in single gene products or in their interaction. The design of therapeutic drugs is also facilitated by the availability of experimental protein structures or good quality protein models. Therefore, protein structures can provide insights into how implicated gene products interact among themselves or with other partners [158].

5. Evaluation of protein–protein interaction predictions

5.1. Need for a biophysically characterized, gold standard set of complexes

In order to evaluate the prediction methods described in the preceding sections, a 'gold standard' reference set is needed. The requirements for this reference set depend on what type of prediction method it will be used to evaluate.

To evaluate methods for the prediction of interacting partners, a set of confirmed interacting partners is needed. Although only a fraction of interactions in existing protein–protein interaction databases is correct and confirmed, a reasonably large interaction partner data set can be compiled. For the purposes of a comparative assessment of large-scale data sets of protein–protein interactions, von Mering *et al* [12] assembled a reference set of known interactions from two catalogs of protein complexes in yeast. One (http://mips.gsf.de/proj/yeast/catalogues/complexes/index.html) is maintained at MIPS [159] and the other is a part of the Bioknowledge database (YPD) [160]. The reference set contains 10 907 trusted interactions and could readily be used for the evaluation of interaction partner prediction methods. Also, the protein–protein interaction database DIP [161] includes a subset (named CORE) that contains the interactions believed to be correct.

For the evaluation of predictions of interacting sites, regions and binding modes, a representative set of known structures of protein complexes is needed. Recently, Keskin *et al* [85] compiled a non-redundant set of the structures of protein interfaces from structures found in the Protein Data Bank. This set contains 3799 clusters of interfaces and can be a good starting point for evaluating prediction methods. Recently, we have compiled a non-redundant template library of protein complexes (unpublished), which can be utilized as a benchmark set as well.

Prediction methods that rely on some energy function or scoring function in order to decide whether an interaction is present would greatly benefit from available data on binding affinities and data on the contribution of individual residues to the free energy of binding. Such data are available in various databases, e.g. KDBI [162] and BID [163] (see the databases section for more detail). False positive and false negative predictions could be evaluated in a more sophisticated way if experimental binding data and predicted binding energies could be compared. However, at present, we are not aware of an integrated, representative, non-redundant data set of protein complexes/interfaces with associated binding energy data or other biophysical parameters. Construction of such a data set is sorely needed not only for the validation and benchmarking of theoretical approaches but also to validate the various proposed high-throughput experimental methods.

5.2. Databases of protein–protein interactions

Several databases contain experimental information on protein–protein interactions. Due to the varying reliability of various experimental techniques, the accuracy of this information varies on a wide range. Table 1 shows a list of some of these databases.

The data in the dedicated protein–protein interaction databases such as DIP and the biomolecular interaction database (BIND) come from various sources. Besides direct submissions, most come from high-throughput experiments and from manual or automatic processing of literature reporting data from small-scale experiments. DIP currently contains data ∼45 000 interactions between ∼17 000 proteins; BIND lists ∼63 000 interactions between ∼35 000 proteins.

In the past couple of years, data from high-throughput experiments such as the mapping of the protein interaction network of yeast [8, 11] and *Caenorhabditis elegans* [164] have tremendously increased the coverage of these databases. Currently, about 80% of interactions in DIP come from high-throughput data [165]. However, the large size of such data sets makes it impractical to verify individual interactions by the same methods as those used in small-scale experiments. Using two forms of computational assessment, namely the expression profile reliability (EPR) index and the paralogous verification method (PVM), Deane *et al* [41] estimated that about 50% out of 8000 pairwise yeast protein interactions in DIP are reliable. The interactions believed to be correct have been separated as a subset of DIP denoted as the CORE, which currently includes about 30% of interactions in DIP [165].

Small-scale experiments provide more reliable data about protein–protein interactions. Such data have been extracted from the biomedical literature and manually curated. Automatic data mining procedures have been helpful in this tedious work: Marcotte *et al* [166] applied a Bayesian approach using discriminating words in Medline abstracts to identify papers about protein–protein interactions in yeast, and Donaldson *et al* [167] developed a support vector machine (SVM) to perform a similar task. Human review and curation, however, are still necessary before the data can get incorporated into the databases.

Although databases such as DIP and BIND do a fairly good job cataloging known protein–protein interactions; in most cases, they contain little more than just the type of experiment used to identify the interaction. Only a negligibly small fraction of database entries contain biophysical data such as a dissociation constant. The KDBI (kinetic data of biomolecular interactions) database was created to fill this gap, providing kinetic data, including dissociation constants and various other rate constants, collected from the literature. It is not limited to protein–protein interactions, but includes kinetic data on protein–RNA, protein–DNA, protein–ligand, RNA–ligand and DNA–ligand binding or reaction events as well. KDBI currently contains 8273 entries of 1231 binding or interaction events, which involves 1380 proteins, 143 nucleic acids and 1395 small molecules.

Both DIP and BIND are in the process of including data on protein complexes from the PDB (Protein Data Bank), the database of protein structures. The PDB is a rich source of structures of protein complexes. Often, however, it is difficult or impossible to determine the physiological oligomeric state of a protein in a given PDB entry just by looking at the entry itself. The deposited coordinates in a PDB entry usually

Table 1. Databases of protein–protein interactions.

Database	Type of information	URL	Reference
DIP (Database of Interacting Proteins)	Interactions (direct binding) between proteins	http://dip.doe-mbi.ucla.edu	[161]
IntAct	Interactions (direct binding) between proteins	http://www.ebi.ac.uk/intact	[216]
BIND (Biomolecular Interaction Network Database)	Interactions (binding) between biomolecules	http://www.bind.ca/	[16]
MINT (Molecular INTeraction database)	Interactions (both direct and indirect relationships) between proteins	http://mint.bio.uniroma2.it/mint/	[217]
BRITE (Biomolecular Relations in Information Transmission and Expression)	‘Generalized interactions’ between proteins (including direct binding) [part of KEGG]	http://www.genome.jp/brite/	[218]
InterDom	Integrative database of putative protein domain interactions	http://interdom.lit.org.sg	[219]
PDB (Protein Data Bank)	Atomic structures of proteins, including those of protein complexes	http://www.rcsb.org/pdb/	[176]
PQS (Protein Quaternary Structures)	Quaternary structures of proteins in PDB	http://pqs.ebi.ac.uk/	[168]
Data set of protein–protein interfaces	Structurally non-redundant set of interfaces	http://protein3d.ncifcrf.gov/∼keskino/	[85]
SPIN-PP (Surface Properties of Interfaces—Protein Protein Interfaces)	protein–protein interfaces in PDB	http://honiglab.cpmc.columbia.edu/SPIN/main.html	Unpublished
KDBI (Kinetic Data of Bio-molecular Interactions)	Kinetic parameters of protein–protein and other interactions	http://xin.cz3.nus.edu.sg/group/kdbi/kdbi.asp	[162]
ASEdb (Alanine Scanning Energetics database)	Energetics of side-chain interactions at heterodimeric interfaces, from alanine scanning mutagenesis	http://www.asedb.org	[220]
BID (Binding Interface Database)	Detailed data on protein interfaces	http://tsailab.org/BID/	[163]
Organism-specific databases	Various interactions, functional links	http://mips.gsf.de/proj/ppi/ and links therein	

consist of the contents of the asymmetric unit (ASU), from which the coordinates of the whole crystal system can be generated. The contents of the ASU can define one or more copies of the macromolecule and crystallographic symmetry operations might be required to generate the complete macromolecule. The Protein Quaternary Structures (PQS) database has been created in an attempt to reconstruct the biologically relevant macromolecular structures using the PDB data. To generate PQS entries, multiple copies of the same molecule are separated, and all relevant symmetry operations are applied, followed by calculating the surfaces buried in interfaces in order to discriminate crystal packing artifacts from functional protein–protein contacts. Since this is an inference procedure itself, some erroneous classifications are expected. According to early tests [168], 19% of complexes classified as probable dimers mismatched some other online annotation. Currently, PQS contains about 30 000 entries, with about 9500 entries being monomers and about 10 000 entries being dimers, with the rest of the entries being divided among various higher-order complexes.

The PDB itself has a section (denoted as ‘biounit’) containing structures of biological complexes reconstructed from the original PDB entries. These reconstructed structures come in part from PQS and are subject to the same potential problems as the PQS entries themselves. On the other hand, the well-annotated protein sequence database Swiss-Prot [62] contains reliable information on the biological oligomerization status (as determined by experiments) of many proteins.

Another structure-derived data set is named ‘data set of protein–protein interfaces’ [85] and contains a non-redundant set of interfaces obtained by clustering the interface structures found in PDB into 3799 clusters. The authors identified three different types of clusters according to whether a similar interface is associated with similar global folds of the component proteins. A related structure-derived database is SPIN-PP (Surface Properties of INterfaces—Protein–Protein Interfaces), which contains images of protein–protein interfaces with various physico-chemical properties mapped onto them. It includes 6460 interfaces, with a non-redundant subset (labeled the ‘unique’ set) of 855 entries.

Other databases of interest include BID (Binding Interface Database) and ASEdb (Alanine Scanning Energetics database). BID entries contain detailed protein descriptions, interaction descriptions and data on the contribution of each amino acid to binding, obtained by mining the primary literature describing alanine scanning and other site-directed mutagenesis experiments. The database currently includes 455 interacting protein pairs with over 6417 hot spots documented.

Created in a similar spirit, ASEdb is a searchable database of single alanine mutations in protein–protein, protein–nucleic acid and protein–small molecule interactions in which binding has been experimentally determined.

Protein–protein interaction databases are constantly growing and becoming enriched with new features and data types. Although they provide invaluable data about various types of analyses and functional annotation methods, the number of biophysically fully characterized protein–protein interactions is still small and only covers a tiny fraction of known interactions. As opposed to a proliferation of various databases collecting the same or different types of data, an effort to integrate or cross-link different types of data in different databases would be very desirable.

5.3. CAPRI (Critical Assessment of Predicted Interactions)

CAPRI ([22]; http://capri.ebi.ac.uk/) is a community-wide experiment, analogous to Critical Assessment of Structure Prediction (CASP [169]), but aimed at assessing the performance of protein–protein docking procedures. Like CASP, the predictions are performed blindly and assessed by an independent team by comparison to x-ray structures available prior to publication. Beginning in 2001, five rounds, including 19 targets, have been completed and the evaluation of the first three rounds and a preliminary assessment of the fourth round have been published [83, 170].

Although the set of 19 target complexes is small and not representative (e.g. five of the seven targets in round 1 were antigen–antibody complexes), CAPRI has revealed several limitations of current docking algorithms and taught us a few important lessons. One is their poor ability to handle conformational changes. Most of the docking protocols used in CAPRI treat the molecular components as rigid bodies or only perform limited exploration of conformational space. Approaches that have the ability to explore larger regions of conformational space (e.g. using essential dynamics [171]) are being tested. Another limitation is that, as a number of recent studies have demonstrated [172, 173], with the exception of the size of the interface, most other parameters (e.g. hydrogen bonds, contact propensities) are poor discriminators between specific and non-specific protein association modes. Although the scoring functions used by CAPRI predictors are sophisticated, they are still not sufficiently reliable.

The most important lesson from CAPRI is that prior knowledge about the regions where the component proteins are likely to interact is tremendously helpful for the docking calculations. Such knowledge is sometimes available from biochemical studies, as was the case with a few CAPRI targets. In other cases, predictors can use computational methods to infer interaction sites from patterns of sequence conservation and sequence signatures [174], correlated mutations [65], homology modeling [72], or threading [76]. Although these methods do have their own limitations, combining interaction site prediction with docking procedures appears to be a very effective way to obtain accurate predictions even for difficult targets.

6. Experimental determination of protein–protein interactions

A comprehensive characterization of protein–protein interactions involves qualitative information such as interaction partners, quantitative information on their kinetic and thermodynamic properties as well as structural descriptions at different levels of resolution. Computational methods and experimental techniques complement each other. While theoretical methods fail to give a complete and accurate picture, experimental methods are costly and time consuming and can only be applied to a small subset of all proteins. Thus, computational methods can help prioritize interesting targets that can then be studied by experiment. Furthermore, most prediction methods have been designed to benefit from coarse experimental data.

Recent experimental approaches have, for the first time, enabled researchers to obtain a qualitative picture of interactions on a genomic scale. However, the number of experimental methods for studying protein interactions is too large for a comprehensive overview to be presented here. Thus, we will focus on recent developments regarding larger molecular assemblies at various levels of resolution, their capabilities and their limitations.

6.1. Atomic level

While the library of solved protein structures appears to be essentially complete for single domain proteins [175], this is far from being true for protein–protein complex structures. True macromolecular protein complexes represent only a small fraction of the currently 27 570 entries (October 2004) in the Protein Data Bank (PDB). The PDB [176] has a bias toward proteins that are easy to express, purify and crystallize, and a negative bias toward membrane proteins—and protein–protein complexes. As of October 2004, 18 930 macromolecular assemblies from 29 277 different PDB files (including obsolete ones) are present in the PQS server [168], and this number includes redundant and biologically non-relevant complexes.

Most of the currently available structures have been solved by x-ray crystallography (23 561) or nuclear magnetic resonance spectrometry (4009). To obtain an x-ray structure, several milligrams of the highly purified protein–protein complex are needed and conditions favorable for crystallization have to be found. Certain types of complexes, e.g. membrane proteins, virus envelopes, weak and transient complexes are particularly difficult to crystallize. Larger assemblies tend to give small crystals and large unit cells; these crystals are often weakly diffracting and more sensitive to radiation [177]. Moreover, not all complexes crystallize, and if they do, it may not be in a biologically relevant conformation. Due to their high-resolution, x-ray structures are still assumed to be the ‘gold standard’ (for a review see [177]) and even large complexes as RNA polymerase, ribosomal subunits, the complete ribosome, a proteasome and the GroEL chaperonin [40] have been solved by x-ray. Nuclear magnetic resonance (NMR), unlike x-ray crystallography, was limited to relatively small proteins (300 amino acids, 30–40 kDa) for a long time

[178]. The development of TROSY (transverse relaxation-optimized spectroscopy) [179] has opened up the possibility of using NMR for larger protein assemblies ($M_r > 100$ kDa). The 100 kDa structure of GroES with GroEL, a 14-mer resulted in a well-resolved ^{1}H–^{15}N spectrum [180] and is only one example out of many [178].

A variety of other well-described methods can provide at least some information on the identity of the interacting residues, including site-directed mutagenesis [181] and fluorescence resonance energy transfer (FRET) [182]. FRET can be used to determine the distance between labeled groups of interacting proteins [183]. Hybrid techniques combining chemical crosslinking with subsequent mass spectrometric identification of the crosslinked peptides after proteolytic digestion appear especially well-suited to capture information on residues involved in transient complexes [184, 185]. An interesting technique for the quick detection of interfacial residues is based on cross-saturation effects coupled with TROSY experiments [186]. It was applied to determine the interface region of the FB–Fc fragment complex ($M_r =$ 64 kDa) and in several other recent studies [187, 188].

6.2. *True interfaces versus crystal contacts*

Crystal contacts are artifacts that only appear upon crystallization of proteins. The forces acting at these interfaces are considered too weak to form at cellular concentrations [189]. The number and location of the artificial contacts can vary according to the crystal symmetry.

The discrimination between biological interfaces and crystal contacts in x-ray structures of protein complexes is a difficult task. Because biological interfaces tend to be larger than interfaces arising from crystal contacts, the size of the interface is the best discriminator, providing an error rate of ∼15%. This result can be further improved by the use of a statistical potential [172]. When biological and crystal dimers having large interfaces (and therefore not distinguishable by interface size) were investigated, it was found that a combination of the non-polar interface area and the fraction of buried interface atoms correctly assigns 88% of the biological dimers and 77% of the crystal dimers. These success rates increased to 93–95% when the residue propensity score of the interfaces was taken into consideration [190]. Interfaces from transient complexes often show a high similarity to crystal contacts, making the identification of these interfaces particularly difficult [107].

6.3. *Shape characterization*

At a lower level of resolution, methods such as electron microscopy (EM) and its subclasses single-particle EM and electron tomography can provide information on the overall shape and symmetry of a protein–protein complex which is often sufficient to assemble high-resolution structures of the individual components into larger complexes.

Electron tomography is used to study very large assemblies like organelles in a cellular context at resolutions of 50 Å [191], but could soon reach 20 Å that was formerly the domain of EM. As EM can only produce two-dimensional images, images at many different orientations are needed to reconstruct the three-dimensional structure of the molecule. Furthermore, the sample is damaged by radiation during the procedure, and therefore requires averaging over images from different molecules. While the resolution of non-crystalline probes is generally too low (∼20 Å), two-dimensional crystals reach resolutions high enough to rebuild the backbone structure (bacteriorhodopsin [192], a/b tubulin [193]). For particles larger than ∼300 kDa, single-particle cryo-EM techniques can achieve resolutions up to approximately 5 Å. This is still not sufficient for an atomic structure but computational methods are used quite successfully to dock protein structures or models into the electron density maps [194]. Examples of difficult cases are the membrane proteins of the dengue virus [195] and bacteriorhodopsin [196]. Although single-particle EM is still very time-consuming compared to x-ray crystallography and NMR techniques, it is a very powerful technique, and due to automation efforts could soon match the high-throughput speed of the other methods [197].

6.4. *Interaction partner level*

An even lower level of resolution models that is applicable on a genomic scale is provided by methods that obtain qualitative information about the identity of the interaction partners. Combinations of MS with affinity purification techniques (for a recent review see [198]) have improved rapidly. Tandem affinity purification (TAP) uses a bait protein that is linked to a tag consisting of two parts, with each part being recognized in a separate affinity purification step. This bait protein is recovered from a whole cell lysate, thereby allowing complexes to be analyzed in their normal cellular milieu. Purification is performed under mild conditions so that interacting proteins stay associated and can subsequently be characterized by mass spectrometry. The tagging system is particularly important for the quality of the data. Non-physiological levels of the bait protein can lead to artifacts. Tagging systems specific for protein–protein complexes are under investigation [182]. Although binary interactions of larger complexes cannot be studied separately, the method can capture large assemblies, e.g. the complete human spliceosome with its ∼100 proteins [199, 200]. Other examples include the characterization of the highly symmetrical yeast nuclear pore which consists of various copies of only ∼30 components [201]. For transient and weak interactions, chemical cross-linking coupled with mass spectrometry appears to be the method of choice.

For studying binary protein interactions at the genomic level, the yeast two-hybrid technique [4] was the first, and is still the most widespread method. It is based on the modular nature of yeast transcriptional activators, consisting of a DNA-binding domain and an activation domain. A protein of interest is fused to the DNA-binding domain and another protein to the activation domain. If the two proteins bind to each other, the two activation factor domains are brought into close proximity and the activity of the transcriptional activator is restored, resulting in the transcription of a reporter gene. Whole cDNA libraries with proteins fused to the activation domain can be screened using yeast cells that express the protein of interest fused to a DNA-binding protein. Using strong promoters, even weakly interacting proteins can be detected. As the interaction takes place in the nucleus of the yeast cell and not in its biological context, there are limitations to the types of proteins that can be investigated, and a number of circumstances can also lead to false positive results.

6.5. Applications to genomes

The first large-scale interaction map of the *S. cerevisiae* proteome [8, 10] was obtained by using the yeast two-hybrid method. Two recent MS-based large-scale efforts analyzed the yeast proteome as well. In one, TAP-tagged proteins were used to identify interacting proteins [9]. The other approach used single-step immunopurification and LC-MS/MS (integrated liquid chromatography with mass spectrometry) [11]. The main difference between the two methods is the way the tagged 'baits' are expressed. In the former work, an endogenous promoter is used, while the latter employs inducible over-expression that can lead to an over-representation of interactions that are not seen in the biological system. The overlap of the results obtained by both the methods was quite small. A comprehensive comparison of results from the yeast two-hybrid investigation with the mass spectrometric investigations and others revealed only marginal overlap between the techniques [12]. The percentage of interactions predicted by more than two methods is low, and only 4.5% of the interactions detected by small-scale experiments and high-throughput methods could be found [165]. Systematic investigation of the four large-scale yeast related screenings in comparison with the MIPS database revealed that the accuracy could be significantly improved by combining two or even three data sets from different methods.

6.6. Kinetic and thermodynamic properties

Although some of the methods mentioned above, such as the yeast two-hybrid approach, give semi-quantitative results, the kinetic and thermodynamic description of protein–protein complexes is a field in its own right. Isothermal titration calorimetry (ITC) measures the heat created upon complex formation and allows for the determination of both the binding constant and the enthalpy of binding [202]. Binding constants in the order of 10^9 M, common for enzyme–inhibitor complexes and high-affinity antibodies, cannot be determined by this method. Surface plasmon resonance (SPR) measures the binding affinity of a molecule to a surface-immobilized receptor in real time and also allows the study of the dynamics of protein interactions [203]. Finally, an emerging and very promising technique based on single molecule force microscopy (FM) [204] should be mentioned: it allows for the direct determination of binding forces. Using FM, mechanical properties of single molecules can be investigated (for a review see [205]). When the force needed to break intermolecular bonds is compared to a known reference bond, e.g. a short DNA duplex, it is possible to measure the unbinding force of the complex [206]. This method was used to study specific versus non-specific binding, and with modern chip technology, such experiments can be carried out in a parallel fashion and are therefore capable of high-throughput [207].

7. What can we learn from interaction networks?

The network representation of the pairwise protein–protein interactions existent in an organism provides a powerful framework to study various biological concepts [208]. Some methods take advantage of topological features of interaction networks to predict the function of uncharacterized proteins [209, 210] or to determine novel protein complexes [211]. Other methods transfer interaction networks from one species to another [212–214]. But, most importantly, a network of physical interactions between proteins is a necessary (although obviously not sufficient) step toward whole cell modeling [215].

8. Summary and outlook

Of late, due their biological importance, protein–protein interactions have been the object of increasing attention, especially as they relate to interactions and associations in the entire proteome. Both large-scale experimental and theoretical approaches have progressed in recent years but still much further development is required. A key condition for success is the development of large-scale experimental benchmarks by which the accuracy of high-throughput approaches can be assessed. With regard to computational methods, combined approaches that can reasonably accurately identify putative interacting regions, followed by either homology modeling or multimeric threading, are likely to be the most successful in the short term. Such methods are, however, limited (especially those that attempt to predict quaternary structure) by the library of already solved folds. Docking of proteins on a genome scale is a far more difficult problem. An accurate solution will require the development of better scoring functions as well as techniques that can remodel the side-chains and/or backbone as the protein complex adjusts from the unbound to the bound state. (Even for single proteins, there are a few algorithms that do a good job when significant backbone rearrangement occurs.) Thus, while some progress has been made, the field is clearly in its infancy and much work will be required to bring the prediction of protein–protein interactions to a robust and reliable state.

Glossary

Conserved residues. Residues of proteins that are evolutionarily conserved across members of a protein family (often including proteins with the same function from different species).

Experimental hot spots. Residues at protein–protein interfaces that contribute significantly to the binding affinity of the complex, measured by the drop in the binding affinity when the residue is mutated to alanine.

High-throughput. A class of experimental techniques, distinguished by the ability to characterize a very large number of proteins or genes (such as an entire genome) in a short time.

Interactome. The network of all interactions between molecules (including proteins, nucleic acids and small organic compounds) in an organism.

Interolog. An interaction between two proteins that have similarly interacting counterparts with similar functions in an evolutionarily related species.

Motif. A recurring pattern that usually correlates with a particular function.

Obligate interface. Interface between two proteins that form a permanent, stable complex, as opposed to transient interactions.

Oligomeric (homo- or hetero-). Consisting of a small number of components, which can either be identical (in homo-oligomers) or different (in hetero-oligomers).

Proteomics. The study of the proteome, i.e. the full set of proteins encoded by a genome.

Residue propensity. The tendency of a particular residue to exhibit a certain property, e.g. to appear in specific structural elements or at specific sites of a protein.

References

[1] Alberts B, Bray D, Lewis J, Raff M, Roberts K and Watson J D 1994 *Molecular Biology of the Cell* 3rd edn (New York: Garland)

[2] Frieden C 1971 *Annu. Rev. Biochem.* **40** 653–96

[3] Legrain P, Wojcik J and Gauthier J M 2001 *Trends Genet.* **17** 346–52

[4] Fields S and Song O 1989 *Nature* **340** 245–6

[5] Yates J R III 2000 *Trends Genet.* **16** 5–8

[6] Sobott F and Robinson C V 2002 *Curr. Opin. Struct. Biol.* **12** 729–34

[7] Rigaut G, Shevchenko A, Rutz B, Wilm M, Mann M and Seraphin B 1999 *Nat. Biotechnol.* **17** 1030–2

[8] Uetz P *et al* 2000 *Nature* **403** 623–7

[9] Gavin A C *et al* 2002 *Nature* **415** 141–7

[10] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y 2001 *Proc. Natl Acad. Sci. USA* **98** 4569–74

[11] Ho Y *et al* 2002 *Nature* **415** 180–3

[12] von Mering C, Krause R, Snel B, Cornell M, Oliver S G, Fields S and Bork P 2002 *Nature* **417** 399–403

[13] Janin J and Seraphin B 2003 *Curr. Opin. Struct. Biol.* **13** 383–8

[14] Krause R, von Mering C and Bork P 2003 *Bioinformatics* **19** 1901–8

[15] Spirin V and Mirny L A 2003 *Proc. Natl Acad. Sci. USA* **100** 12123–8

[16] Bader G D, Betel D and Hogue C W 2003 *Nucleic Acids Res.* **31** 248–50

[17] Pellegrini M, Marcotte E M, Thompson M J, Eisenberg D and Yeates T O 1999 *Proc. Natl Acad. Sci. USA* **96** 4285–8

[18] Pazos F and Valencia A 2002 *Proteins* **47** 219–27

[19] Huynen M A, Snel B, von Mering C and Bork P 2003 *Curr. Opin. Cell. Biol.* **15** 191–8

[20] Nooren I M and Thornton J M 2003 *J. Mol. Biol.* **325** 991–1018

[21] Cochran A G 2001 *Curr. Opin. Chem. Biol.* **5** 654–9

[22] Janin J, Henrick K, Moult J, Eyck L T, Sternberg M J, Vajda S, Vakser I and Wodak S J 2003 *Proteins* **52** 2–9

[23] Caffrey D R, Somaroo S, Hughes J D, Mintseris J and Huang E S 2004 *Protein Sci.* **13** 190–202

[24] Lichtarge O, Bourne H R and Cohen F E 1996 *J. Mol. Biol.* **257** 342–58

[25] Ofran Y and Rost B 2003 *FEBS Lett.* **544** 236–9

[26] Jones S and Thornton J M 1997 *J. Mol. Biol.* **272** 121–32

[27] Janin J 1995 *Prog. Biophys. Mol. Biol.* **64** 145–66

[28] Jones S and Thornton J M 1996 *Proc. Natl Acad. Sci. USA* **93** 13–20

[29] Larsen T A, Olson A J and Goodsell D S 1998 *Structure* **6** 421–7

[30] Chakrabarti P and Janin J 2002 *Proteins* **47** 334–43

[31] Lo Conte L, Chothia C and Janin J 1999 *J. Mol. Biol.* **285** 2177–98

[32] Glaser F, Steinberg D M, Vakser I A and Ben-Tal N 2001 *Proteins* **43** 89–102

[33] Ofran Y and Rost B 2003 *J. Mol. Biol.* **325** 377–87

[34] Hu Z, Ma B, Wolfson H and Nussinov R 2000 *Proteins* **39** 331–42

[35] Ma B, Elkayam T, Wolfson H and Nussinov R 2003 *Proc. Natl Acad. Sci. USA* **100** 5772–7

[36] Halperin I, Wolfson H and Nussinov R 2004 *Structure (Camb)* **12** 1027–38

[37] Galperin M Y and Koonin E V 2000 *Nat. Biotechnol.* **18** 609–13

[38] Valencia A and Pazos F 2002 *Curr. Opin. Struct. Biol.* **12** 368–73

[39] Giot L *et al* 2003 *Science* **302** 1727–36

[40] Russell R B, Alber F, Aloy P, Davis F P, Korkin D, Pichaud M, Topf M and Sali A 2004 *Curr. Opin. Struct. Biol.* **14** 313–24

[41] Deane C M, Salwinski L, Xenarios I and Eisenberg D 2002 *Mol. Cell Proteomics* **1** 349–56

[42] Barabasi A L and Oltvai Z N 2004 *Nat. Rev. Genet.* **5** 101–13

[43] Yu H, Luscombe N M, Lu H X, Zhu X, Xia Y, Han J D, Bertin N, Chung S, Vidal M and Gerstein M 2004 *Genome Res.* **14** 1107–18

[44] Morett E, Korbel J O, Rajan E, Saab-Rincon G, Olvera L, Olvera M, Schmidt S, Snel B and Bork P 2003 *Nat. Biotechnol.* **21** 790–5

[45] Goh C S and Cohen F E 2002 *J. Mol. Biol.* **324** 177–92

[46] Albert I and Albert R 2004 *Bioinformatics* **20** 3346–52

[47] Korbel J O, Jensen L J, von Mering C and Bork P 2004 *Nat. Biotechnol.* **22** 911–7

[48] Gaasterland T and Ragan M A 1998 *Microb. Comp. Genomics* **3** 199–217

[49] Wu J, Kasif S and DeLisi C 2003 *Bioinformatics* **19** 1524–30

[50] Huynen M, Snel B, Lathe W 3rd and Bork P 2000 *Genome Res.* **10** 1204–10

[51] Dandekar T, Snel B, Huynen M and Bork P 1998 *Trends Biochem. Sci.* **23** 324–8

[52] Overbeek R, Fonstein M, D'Souza M, Pusch G D and Maltsev N 1999 *Proc. Natl Acad. Sci. USA* **96** 2896–901

[53] von Mering C and Bork P 2002 *Nature* **417** 797–8

[54] Rogozin I B, Makarova K S, Wolf Y I and Koonin E V 2004 *Brief Bioinform.* **5** 131–49

[55] Marcotte E M, Pellegrini M, Ng H L, Rice D W, Yeates T O and Eisenberg D 1999 *Science* **285** 751–3

[56] Enright A J, Iliopoulos I, Kyrpides N C and Ouzounis C A 1999 *Nature* **402** 86–90

[57] Yanai I, Derti A and DeLisi C 2001 *Proc. Natl. Acad. Sci. USA* **98** 7940–5

[58] Altschul S F, Gish W, Miller W, Myers E W and Lipman D J 1990 *J. Mol. Biol.* **215** 403–10

[59] Bateman A *et al* 2004 *Nucleic Acids Res.* **32** D138–41

[60] Corpet F, Servant F, Gouzy J and Kahn D 2000 *Nucleic Acids Res.* **28** 267–9

[61] Truong K and Ikura M 2003 *BMC Bioinform.* **4** 16

[62] Boeckmann B *et al* 2003 *Nucleic Acids Res* **31** 365–70

[63] Hua S, Guo T, Gough J and Sun Z 2002 *J. Mol. Biol.* **320** 713–9

[64] Tsoka S and Ouzounis C A 2000 *Nat. Genet.* **26** 141–2

[65] Pazos F, Helmer-Citterich M, Ausiello G and Valencia A 1997 *J. Mol. Biol.* **271** 511–23

[66] Goh C S, Bogan A A, Joachimiak M, Walther D and Cohen F E 2000 *J. Mol. Biol.* **299** 283–93

[67] Pazos F and Valencia A 2001 *Protein Eng.* **14** 609–14

[68] Ramani A K and Marcotte E M 2003 *J. Mol. Biol.* **327** 273–84

[69] Fraser H B, Hirsh A E, Wall D P and Eisen M B 2004 *Proc. Natl Acad. Sci. USA* **101** 9033–8

[70] Sali A and Blundell T L 1993 *J. Mol. Biol.* **234** 779–815

[71] Pieper U *et al* 2004 *Nucleic Acids Res.* **32** D217–22

[72] Aloy P and Russell R B 2002 *Proc. Natl Acad. Sci. USA* **99** 5896–901

[73] Aloy P and Russell R B 2003 *Bioinformatics* **19** 161–2

[74] Aloy P, Bottcher B, Ceulemans H, Leutwein C, Mellwig C, Fischer S, Gavin A C, Bork P, Superti-Furga G, Serrano L and Russell R B 2004 *Science* **303** 2026–9

[75] Aloy P, Ceulemans H, Stark A and Russell R B 2003 *J. Mol. Biol.* **332** 989–98

[76] Lu L, Lu H and Skolnick J 2002 *Proteins* **49** 350–64

[77] Skolnick J and Kihara D 2001 *Proteins* **42** 319–31

[78] Lu H and Skolnick J 2001 *Proteins* **44** 223–32

[79] Lu L, Arakaki A K, Lu H and Skolnick J 2003 *Genome Res.* **13** 1146–54

[80] Sternberg M J, Gabb H A and Jackson R M 1998 *Curr. Opin. Struct. Biol.* **8** 250–6

[81] Bogan A A and Thorn K S 1998 *J. Mol. Biol.* **280** 1–9

[82] DeLano W L 2002 *Curr. Opin. Struct. Biol.* **12** 14–20

[83] Mendez R, Leplae R, De Maria L and Wodak S J 2003 *Proteins* **52** 51–67

[84] Dominguez C, Boelens R and Bonvin A M 2003 *J. Am. Chem. Soc.* **125** 1731–7

[85] Keskin O, Tsai C J, Wolfson H and Nussinov R 2004 *Protein Sci.* **13** 1043–55

[86] Fernandez-Recio J, Totrov M and Abagyan R 2002 *Protein Sci.* **11** 280–91

[87] Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem A A, Aflalo C and Vakser I A 1992 *Proc. Natl Acad. Sci. USA* **89** 2195–9

[88] Lin S L, Nussinov R, Fischer D and Wolfson H J 1994 *Proteins* **18** 94–101

[89] Connolly M L 1983 *Science* **221** 709–13

[90] Kimura S R, Brower R C, Vajda S and Camacho C J 2001 *Biophys. J.* **80** 635–42

[91] Rajamani D, Thiel S, Vajda S and Camacho C J 2004 *Proc. Natl Acad. Sci. USA* **101** 11287–92

[92] Gabb H A, Jackson R M and Sternberg M J 1997 *J. Mol. Biol.* **272** 106–20

[93] Chen R and Weng Z 2002 *Proteins* **47** 281–94

[94] Heifetz A and Eisenstein M 2003 *Protein Eng.* **16** 179–85

[95] Vakser I A 1995 *Protein Eng.* **8** 371–7

[96] Vakser I A 1996 *Biopolymers* **39** 455–64

[97] Li C H, Ma X H, Chen W Z and Wang C X 2003 *Protein Eng.* **16** 265–9

[98] Eisenstein M and Katchalski-Katzir E 1998 *Lett. Pept. Sci.* **5** 365–9

[99] Jackson R M, Gabb H A and Sternberg M J 1998 *J. Mol. Biol.* **276** 265–85

[100] Mandell J G, Roberts V A, Pique M E, Kotlovyi V, Mitchell J C, Nelson E, Tsigelny I and Ten Eyck L F 2001 *Protein Eng.* **14** 105–13

[101] Chen R, Li L and Weng Z 2003 *Proteins* **52** 80–7

[102] Ritchie D W and Kemp G J 2000 *Proteins* **39** 178–94

[103] Jiang F and Kim S H 1991 *J. Mol. Biol.* **219** 79–102

[104] Gardiner E J, Willett P and Artymiuk P J 2001 *Proteins* **44** 44–56

[105] Taylor J S and Burnett R M 2000 *Proteins* **41** 173–91

[106] Palma P N, Krippahl L, Wampler J E and Moura J J 2000 *Proteins* **39** 372–84

[107] Nooren I M and Thornton J M 2003 *EMBO J.* **22** 3486–92

[108] Sheinerman F B and Honig B 2002 *J. Mol. Biol.* **318** 161–77

[109] Young L, Jernigan R L and Covell D G 1994 *Protein Sci.* **3** 717–29

[110] Berchanski A, Shapira B and Eisenstein M 2004 *Proteins* **56** 130–42

[111] Honig B and Nicholls A 1995 *Science* **268** 1144–9

[112] Schreiber G and Fersht A R 1996 *Nat. Struct. Biol.* **3** 427–31

[113] Camacho C J, Weng Z, Vajda S and DeLisi C 1999 *Biophys. J.* **76** 1166–78

[114] Zhang C, Vasmatzis G, Cornette J L and DeLisi C 1997 *J. Mol. Biol.* **267** 707–26

[115] Camacho C J, Gatchell D W, Kimura S R and Vajda S 2000 *Proteins* **40** 525–37

[116] Gray J J, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl C A and Baker D 2003 *J. Mol. Biol.* **331** 281–99

[117] Vajda S and Camacho C J 2004 *Trends Biotechnol.* **22** 110–6

[118] Betts M J and Sternberg M J 1999 *Protein Eng.* **12** 271–83

[119] Sandak B, Wolfson H J and Nussinov R 1998 *Proteins* **32** 159–74

[120] Schneidman-Duhovny D, Inbar Y, Polak V, Shatsky M, Halperin I, Benyamini H, Barzilai A, Dror O, Haspel N, Nussinov R and Wolfson H J 2003 *Proteins* **52** 107–12

[121] Lawrence M C and Colman P M 1993 *J. Mol. Biol.* **234** 946–50

[122] Jackson R M 1999 *Protein Sci.* **8** 603–13

[123] Fernandez-Recio J, Totrov M and Abagyan R 2004 *J. Mol. Biol.* **335** 843–65

[124] Ben-Zeev E and Eisenstein M 2003 *Proteins* **52** 24–7

[125] Tovchigrechko A, Wells C A and Vakser I A 2002 *Protein Sci.* **11** 1888–96

[126] Berchanski A and Eisenstein M 2003 *Proteins* **53** 817–29

[127] Fariselli P, Pazos F, Valencia A and Casadio R 2002 *Eur. J. Biochem.* **269** 1356–61

[128] Aloy P, Querol E, Aviles F X and Sternberg M J 2001 *J. Mol. Biol.* **311** 395–408

[129] Landgraf R, Xenarios I and Eisenberg D 2001 *J. Mol. Biol.* **307** 1487–502

[130] Kortemme T and Baker D 2002 *Proc. Natl Acad. Sci. USA* **99** 14116–21

[131] Kortemme T, Kim D E and Baker D 2004 *Sci. STKE* **2004** pl2

[132] Kortemme T, Joachimiak L A, Bullock A N, Schuler A D, Stoddard B L and Baker D 2004 *Nat. Struct. Mol. Biol.* **11** 371–9

[133] Ben-Naim A 1990 *Biopolymers* **29** 567–96

[134] Smith G R and Sternberg M J 2002 *Curr. Opin. Struct. Biol.* **12** 28–35

[135] Sippl M J 1990 *J. Mol. Biol.* **213** 859–83

[136] Miyazawa S and Jernigan R L 1996 *J. Mol. Biol.* **256** 623–44

[137] Reva B A, Finkelstein A V, Sanner M F and Olson A J 1997 *Protein Eng.* **10** 865–76

[138] Lazaridis T and Karplus M 1999 *Proteins* **35** 133–52

[139] Melo F and Feytmans E 1997 *J. Mol. Biol.* **267** 207–22

[140] Moont G, Gabb H A and Sternberg M J 1999 *Proteins* **35** 364–73

[141] Lu H, Lu L and Skolnick J 2003 *Biophys. J.* **84** 1895–901

[142] Zhang C, Liu S, Zhou H and Zhou Y 2004 *Protein Sci.* **13** 400–11

[143] Tobi D, Shafran G, Linial N and Elber R 2000 *Proteins* **40** 71–85

[144] Zhou H and Zhou Y 2002 *Protein Sci.* **11** 2714–26

[145] Ben-Naim A 1997 *J. Chem. Phys.* **107** 3698–706

[146] Finkelstein A V, Badretdinov A and Gutin A M 1995 *Proteins* **23** 142–50

[147] Thomas P D and Dill K A 1996 *J. Mol. Biol.* **257** 457–69

[148] Zhang L and Skolnick J 1998 *Protein Sci.* **7** 112–22

[149] Betancourt M R and Thirumalai D 1999 *Protein Sci.* **8** 361–9

[150] Mohanty D, Dominy B N, Kolinski A, Brooks C L III and Skolnick J 1999 *Proteins* **35** 447–52

[151] Aloy P, Ciccarelli F D, Leutwein C, Gavin A C, Superti-Furga G, Bork P, Bottcher B and Russell R B 2002 *EMBO Rep.* **3** 628–35

[152] Aloy P and Russell R B 2004 *Nat. Biotechnol.* **22** 1317–21

[153] Aloy P and Russell R B 2002 *Trends Biochem. Sci.* **27** 633–8

[154] Vakser I A 2004 *Structure (Camb.)* **12** 910–2

[155] Landgraf C, Panni S, Montecchi-Palazzi L, Castagnoli L, Schneider-Mergener J, Volkmer-Engert R and Cesareni G 2004 *PLoS Biol.* **2** E14

[156] Zarrinpar A, Park S H and Lim W A 2003 *Nature* **426** 676–80

[157] Kortemme T and Baker D 2004 *Curr. Opin. Chem. Biol.* **8** 91–7

[158] Zhou H X 2004 *Curr. Med. Chem.* **11** 539–49

[159] Mewes H W *et al* 2004 *Nucleic Acids Res.* **32** D41–4

[160] Costanzo M C *et al* 2001 *Nucleic Acids Res.* **29** 75–9

[161] Salwinski L, Miller C S, Smith A J, Pettit F K, Bowie J U and Eisenberg D 2004 *Nucleic Acids Res.* **32** D 449–51

[162] Ji Z L, Chen X, Zhen C J, Yao L X, Han L Y, Yeo W K, Chung P C, Puy H S, Tay Y T, Muhammad A and Chen Y Z 2003 *Nucleic Acids Res.* **31** 255–7

[163] Fischer T B *et al* 2003 *Bioinformatics* **19** 1453–4

[164] Li S *et al* 2004 *Science* **303** 540–3

[165] Salwinski L and Eisenberg D 2003 *Curr. Opin. Struct. Biol.* **13** 377–82

[166] Marcotte E M, Xenarios I and Eisenberg D 2001 *Bioinformatics* **17** 359–63

[167] Donaldson I *et al* 2003 *BMC Bioinform.* **4** 11

[168] Henrick K and Thornton J M 1998 *Trends Biochem. Sci.* **23** 358–61

[169] Venclovas C, Zemla A, Fidelis K and Moult J 2003 *Proteins* **53** (Suppl. 6) 585–95

[170] Wodak S J and Mendez R 2004 *Curr. Opin. Struct. Biol.* **14** 242–9

[171] Amadei A, Linssen A B and Berendsen H J 1993 *Proteins* **17** 412–25

[172] Ponstingl H, Henrick K and Thornton J M 2000 *Proteins* **41** 47–57

[173] Janin J and Rodier F 1995 *Proteins* **23** 580–7

[174] Lichtarge O and Sowa M E 2002 *Curr. Opin. Struct. Biol.* **12** 21–7

[175] Kihara D and Skolnick J 2003 *J. Mol. Biol.* **334** 793–802

[176] Berman H M *et al* 2002 *Acta Crystallogr. D Biol. Crystallogr.* **58** 899–907

[177] Hendrickson W A 2000 *Trends Biochem. Sci.* **25** 637–43

[178] Riek R, Pervushin K and Wuthrich K 2000 *Trends Biochem. Sci.* **25** 462–8

[179] Pervushin K, Riek R, Wider G and Wuthrich K 1997 *Proc. Natl Acad. Sci. USA* **94** 12366–71

[180] Fiaux J, Bertelsen E B, Horwich A L and Wuthrich K 2002 *Nature* **418** 207–11

[181] Wells J A 1991 *Methods Enzymol.* **202** 390–411

[182] Phizicky E, Bastiaens P I, Zhu H, Snyder M and Fields S 2003 *Nature* **422** 208–15

[183] Sali A, Glaeser R, Earnest T and Baumeister W 2003 *Nature* **422** 216–25

[184] Rappsilber J, Siniossoglou S, Hurt E C and Mann M 2000 *Anal. Chem.* **72** 267–75

[185] Melcher K 2004 *Curr. Protein Pept. Sci.* **5** 287–96

[186] Takahashi H, Nakanishi T, Kami K, Arata Y and Shimada I 2000 *Nat. Struct. Biol.* **7** 220–3

[187] Morrison J, Yang J C, Stewart M and Neuhaus D 2003 *J. Mol. Biol.* **333** 587–603

[188] Takeuchi K, Takahashi H, Sugai M, Iwai H, Kohno T, Sekimizu K, Natori S and Shimada I 2004 *J. Biol. Chem.* **279** 4981–7

[189] Carugo O and Argos P 1997 *Protein Sci.* **6** 2261–3

[190] Bahadur R P, Chakrabarti P, Rodier F and Janin J 2004 *J. Mol. Biol.* **336** 943–55

[191] Baumeister W 2002 *Curr. Opin. Struct. Biol.* **12** 679–84

[192] Henderson R and Schertler G F 1990 *Philos. Trans. R. Soc. Lond.* **B326** 379–89

[193] Nogales E, Wolf S G and Downing K H 1998 *Nature* **391** 199–203

[194] Gao H *et al* 2003 *Cell* **113** 789–801

[195] Zhang W, Chipman P R, Corver J, Johnson P R, Zhang Y, Mukhopadhyay S, Baker T S, Strauss J H, Rossmann M G and Kuhn R J 2003 *Nat. Struct. Biol.* **10** 907–12

[196] Grigorieff N, Ceska T A, Downing K H, Baldwin J M and Henderson R 1996 *J. Mol. Biol.* **259** 393–421

[197] Zhu Y *et al* 2004 *J. Struct. Biol.* **145** 3–14

[198] Aebersold R and Mann M 2003 *Nature* **422** 198–207

[199] Rappsilber J, Ryder U, Lamond A I and Mann M 2002 *Genome Res.* **12** 1231–45

[200] Zhou Z, Licklider L J, Gygi S P and Reed R 2002 *Nature* **419** 182–5

[201] Rout M P, Aitchison J D, Suprapto A, Hjertaas K, Zhao Y and Chait B T 2000 *J. Cell. Biol.* **148** 635–51

[202] Pierce M M, Raman C S and Nall B T 1999 *Methods* **19** 213–21

[203] Leatherbarrow R J and Edwards P R 1999 *Curr. Opin. Chem. Biol.* **3** 544–7

[204] Binnig G, Quate C F and Gerber C 1986 *Phys. Rev. Lett.* **56** 930–3

[205] Clausen-Schaumann H, Seitz M, Krautbauer R and Gaub H E 2000 *Curr. Opin. Chem. Biol.* **4** 524–30

[206] Albrecht C, Blank K, Lalic-Multhaler M, Hirler S, Mai T, Gilbert I, Schiffmann S, Bayer T, Clausen-Schaumann H and Gaub H E 2003 *Science* **301** 367–70

[207] Blank K *et al* 2003 *Proc. Natl Acad. Sci. USA* **100** 11356–60

[208] Jansen R and Gerstein M 2004 *Curr. Opin. Microbiol.* **7** 535–45

[209] Vazquez A, Flammini A, Maritan A and Vespignani A 2003 *Nat. Biotechnol.* **21** 697–700

[210] Bu D *et al* 2003 *Nucleic Acids Res.* **31** 2443–50

[211] Bader G D and Hogue C W 2002 *Nat. Biotechnol.* **20** 991–7

[212] Matthews L R, Vaglio P, Reboul J, Ge H, Davis B P, Garrels J, Vincent S and Vidal M 2001 *Genome Res.* **11** 2120–6

[213] Wojcik J, Boneca I G and Legrain P 2002 *J. Mol. Biol.* **323** 763–70

[214] Wojcik J and Schachter V 2001 *Bioinformatics* **17** (Suppl. 1) S296–305

[215] Slepchenko B M, Schaff J C, Macara I and Loew L M 2003 *Trends Cell. Biol.* **13** 570–6

[216] Hermjakob H *et al* 2004 *Nucleic Acids Res.* **32** D 452–5

[217] Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M and Cesareni G 2002 *FEBS Lett.* **513** 135–40

[218] Kanehisa M, Goto S, Kawashima S, Okuno Y and Hattori M 2004 *Nucleic Acids Res.* **32** D 277–80

[219] Ng S K, Zhang Z, Tan S H and Lin K 2003 *Nucleic Acids Res.* **31** 251–4

[220] Thorn K S and Bogan A A 2001 *Bioinformatics* **17** 284–5