INSTITUTE OF PHYSICS PUBLISHING PHYSICAL BIOLOGY

Phys. Biol. **2** (2005) S36–S43 doi:10.1088/1478-3975/2/2/S04

Character and evolution of protein–protein interfaces

Ivica Reš and Olivier Lichtarge

Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

E-mail: lichtarge@bcm.tmc.edu

Received 7 March 2005
Accepted for publication 9 May 2005
Published 27 May 2005
Online at stacks.iop.org/PhysBio/2/S36

Abstract

Protein–protein interactions create the macromolecular assemblies and sequential signaling pathways essential for cell function. Their number far exceeds the number of proteins themselves and their experimental characterization, while improving, remains relatively slow. For these reasons, novel computational methods have important roles to play in understanding the physical basis of protein interactions, and in constraining the molecular basis of their specificity. This paper discusses methods based on multiple sequence alignments of protein homologues and phylogenetic trees.

Introduction

Protein–protein interactions play a central role in health and disease. When they occur as a series of sequential individual events, they create signaling pathways and cellular networks tightly regulated by transient protein–protein interactions. When they occur in parallel, they mediate the assembly of large macromolecular machines, such as the proteins bound together in RNA polymerase II. Finally, abnormal interactions among misfolded proteins can fuel the growth of toxic cellular aggregates linked to neurodegenerative diseases, such as Alzheimer's, Parkinson's, prion encephalopathies, as well as type II diabetes and cystic fibrosis [1].

For these reasons, the control of protein–protein interactions is increasingly seen as the next frontier in pharmaceuticals [2]. Some drugs could be designed to promote protein–protein complex formation, such as Cyclosporine, which attenuates the immune response otherwise leading to organ transplant rejection [3]. Other drugs might inhibit protein interactions, such as a peptide designed to block poly-glutamine aggregation and shown to block Huntington's disease in a fly model [4].

This paper will focus on the current state of computational analysis of molecular interfaces and their evolution. After general comments on the scope of the protein–protein interaction problem, it will summarize the properties of protein interfaces and the energetics of surface residues (hot spots). Then it will consider the evolutionary information such as residue conservation, correlations and phylogenetic trees, and it will end with a review of machine learning applications. Reviews of important topics such as docking, molecular dynamics and Monte Carlo simulations can be found elsewhere [5, 6].

Scope of the protein–protein interaction problem

Our understanding of protein–protein interactions is far from complete. In order to estimate their diversity, an analysis recently compared the number of genes in genomes from bacteria, yeast, fly and worm with the number of protein–protein interactions detected by yeast two hybrid [7–10], and by affinity purification followed by mass spectrometry [11, 12]. The result is a remarkably consistent linear correlation ($R^2 = 0.96$) [13]. This leads the authors to estimate that, discounting proteins with more than 30% sequence identity, one can expect of the order of 10 000 distinct structural types of protein–protein interaction. However, the protein structure data bank [14] contains thus far representatives of only 2000–3000 distinct interactions [13, 15], or at most only 30% of the estimated total. To put this number in perspective, the proteome is thought to contain about 1000 distinct protein folds [16], about 80% of which may already have a structural representative [17].

Many techniques aim to bridge this large gap in the structural description of protein–protein interactions. A number of biochemical and biophysical methods provide

isolated structural clues, as recently reviewed in [18]. Electron microscopy and electron tomography provide insights into large molecular assemblies, typically at 15–30 Å resolution, but the best data come from x-ray crystallography and NMR spectroscopy of smaller complexes, which provide detailed atomic resolution views of protein–protein interfaces (resolutions of 2–5 Å). New structures accumulate very slowly though: at an average rate of 200–300 per year [13]. Among these a few very large molecular machines are solved at a rate of 1 or 2 per year: chaperonin [19]; the proteasome [20]; RNA polymerase [21]; the ribosome [22]; the GroEL/GroES complex by NMR [23]; photosystem I [24]; the light harvesting complex [25]; the signal recognition particle [26] and viral structures [27, 28].

To complement these painstaking experimental studies, computational techniques aim to analyze and to predict protein–protein interactions. One basic approach is to compute the physical interactions among atoms and molecules to simulate proteins on computer, for example by using molecular dynamics or force fields to score docking models [5, 6]. This would be equivalent to *de novo* approaches of protein structure prediction. Another approach, more akin to homology modeling, uses evolutionary information. It is based on the fact that sequences with more than 30% sequence identity can be expected not only to share a common ancestry and fold [29], but also common interaction propensities [30] since their joint evolutionary analysis is sufficient to identify their common functional sites [31, 32]. Specifically, a study of domain pairs in different structures showed that they nearly always interact similarly if they have greater than 30–40% sequence identity [30]. This is not true if their similarity is limited to having the same fold.

The methods we review below extract information from the multiple sequence alignment (MSA), either based on conservation of columns in MSA, or on various types of correlations: between different columns of the MSA, between columns and evolutionary trees or between the trees themselves. In the former case, the idea is that the column conservation reflects the evolutionary importance of corresponding residues, while in the latter case the importance is reflected in the fact that different residues (columns) or tree branches change in ways that are coordinated with functional changes. The basic information obtained by using these techniques is the evolutionary importance of residues in protein, without underlying physical or chemical reasons. Information from MSA can be combined with structural, physical and chemical properties, and integrated using different machine learning methods.

The composition of protein–protein interfaces depends on the type of interaction

Multiple studies have compared diverse types of protein–protein interactions. One should first note that interface residues have been defined in two ways. Either as those that change by at least 1 $Å^2$ in solvent accessible surface area (ASA) upon forming a complex [33]. Alternately, a residue is part of an interface if its distance from the interacting partner is smaller than some cut-off of 4 or 5 Å [34, 35]. The different interactions considered have typically included homo-oligomers and hetero-oligomers (interaction between identical or non-identical chains), obligate interactions (without which the individual proteins are not found as stable structures *in vivo*) and transient (which associate and dissociate *in vivo*) or permanent interactions [36]. Functional differences were also considered, including: enzyme-inhibitor, antibody–antigen, enzyme complexes [37, 38].

The results showed that while general differences do exist, they are seldom sufficiently pronounced to predict either the structural or functional type of interaction [39]. Homomultimer interfaces are more hydrophobic than those between heteromultimers [37, 40]. Similarly, the amino acid composition and residue–residue contact preferences are different in six types of protein complexes: same structural domain versus different structural domains, permanent versus transient complexes and homo-oligomers versus hetero-oligomers [41]. A comparison of 16 weak and 23 strong transient homodimers has shown that the weak homodimers have smaller, more planar and more polar contact areas. The strong transient dimers often undergo large conformational changes upon complexation (or dissociation) [42]. Another study of 122 homodimer interfaces has shown that the interface is often made of a core of buried residues, surrounded by a rim of solvent accessible residues [43].

Hot spots

The difficulty in identifying more specific and predictive interfacial features may well lie in the fact that each interface is itself heterogeneous: only a fraction of its residues contribute to function or to the energetics of binding. As a result the analysis of entire interface is always noisy, regardless of how it is defined. The electrostatic contributions to protein–protein interactions can vary from stabilizing to strongly destabilizing, according to a study done on four protein complexes [44]. This unequal contribution of residues to binding free energy is most directly demonstrated by alanine scanning mutagenesis (whereby individual residues are mutated to alanine and the protein function is then assayed) [45]. The few residues that make up the bulk of binding energy define ‘hot spots’. Among heterodimeric complexes, these hot spots appear to be enriched in Trp, Tyr, Arg, residues that can make multiple types of interactions [46], and they were often surrounded by hydrophobic rings, presumably to occlude bulk solvent [46]. The total electrostatic contribution to binding was found to be inversely correlated with buried total and non-polar surface area [44].

To understand hot spots, a number of studies focused on their energetics and modeled possible binding free energy functions. A simple physical model only takes into account limited changes in the backbone conformation and does not explicitly include water [47]. The free energy function has terms for the van der Waals interactions, Coulomb electrostatics, hydrogen bonding, solvation energy and amino-acid-dependent backbone angle probabilities. This model identified 69% of hot spots on 19 protein complexes, and

this improved to 79% when only the interface residues were considered. The hydrogen bonding term proved to be the most significant: without it hot spot prediction accuracy falls to 47%. This simple model [47] was further shown to be well correlated with more sophisticated molecular mechanics Poisson–Boltzmann calculations [48], and it may thus capture the essential physics of interactions. Nevertheless, the physics of hot spots is still not entirely elucidated since a different Monte Carlo study suggests instead that hydrophobic interactions are the most relevant ones [49].

From a different perspective, other studies have focused on the structural conservation properties of hot spots. Ma *et al* [50] studied the correlations between interface propensities of structurally conserved residues and experimental enrichment of hot spots. Based on ten protein families and the data on amino acid preferences in hot spots from [46], the authors obtained a correlation coefficient of 0.7. This study was then extended to compare the coupling of structurally conserved residues and of hot spots across protein interfaces [51]. Residues from distinct chains were considered coupled if the distance between the centers of any of their atoms was less than 5 Å. These conserved residues were coupled across the interfaces nearly twice as often than expected by chance, and their association with hotspots was 5.2 times greater than expected by chance. Lastly, residue packing was higher around hot spots versus non-hotspots (on average 69.2 $Å^3$ versus 38.1 $Å^3$), and a strong correlation ($r^2 = 0.94$) between the hot spot contribution to free energy change and across the interface local atomic packing has been found [51]. The authors pointed out that such packing excludes the solvent, and thus may stabilize an interaction by lowering the dielectric constant and increasing the electrostatic and hydrogen bond interactions [46, 51]. Overall, however, the interacting pairs of interface residues, the charge-conserved residue pairs seem to be disfavored across the interface [51], in keeping with the heterogeneity of evolutionary pressures within an interface.

A recent study suggests that the insufficiently dehydrated hydrogen bonds play an important role in protein interactions, based on a data set of 1476 high-resolution protein structures [52]. Most backbone hydrogen bonds are 'wrapped' by nonpolar groups, except for a few hot spots which are 'underwrapped'. These hot spots become dramatically stabilized by the removal of water, indicating their important contribution to binding sites [52].

Recently, small-world networks [53] have been used to study protein interactions. The basic idea of this approach is to represent proteins as networks (graphs), where residues are the nodes and interactions are the edges [54]. These networks can be used to identify the residues at or near hot spots [55, 56].

Evolution of interfaces

Besides energetics and structure, a different line of study of interfaces considers evolution. Most simply, residues involved in interactions are less likely to vary. This hypothesis can be quantified by considering the conservation of columns in MSAs of related proteins from multiple organisms, and calculating the information entropy or some related measure of conservation. Thus, based on the similarity scores of MSAs, the interface residues of six homodimer families were more conserved than the rest of the protein surface [57]. Similarly, the interfaces in a larger set of proteins that included homodimers were also more conserved (based on the Von Neumann entropy) than the rest of the surface, but this fact alone was insufficient to predict the location of the interface [58].

Conservation can nevertheless be useful to distinguish true interfaces in protein dimers from false ones due to crystallization artifacts [33]. In 53 homodimers and 65 monomers, when information entropy was combined with the change in residue solvent accessible surface area upon complexation, they could discriminate between biological and non-biological protein–protein interface with an accuracy of 86% [59]. Size and conservation together can thus discriminate biological from non-biological contacts [33].

A problem with this simple model of evolutionary conservation is that for it to work best, interfaces would have to be immutable. But this would require in turn that each distinct interaction in protein networks would depend on entirely different proteins. In fact, we know that the opposite is true: pathways are opportunistically built by cannibalizing and adapting already existing protein parts such as SH2, SH3, PDZ and many other modules or domains. Similarly, enzymes such as kinases have repeatedly mutated to alter ligand specificity while maintaining their key phosphorylation capability. A more natural model of evolution should account therefore not only for the absolute conservation of key functional residues, but also for the systematic variation of specificity determinants as they adapt to different evolutionary constraints.

Evolutionary trace of proteins

In recognition of these issues, a set of techniques called evolutionary tracing (ET) was developed to identify patterns of variation in multiple sequence alignments that match the patterns of functional divergences suggested by phylogenetic trees [31]. Methods based on this idea have been tested repeatedly and shown to predict functional sites [60–65]. The key concept of ET [31] is to build a phylogenetic tree based on the alignment and then to rank every column according to whether its residue variation pattern correlates with the branching pattern of the evolutionary tree. As the tree is used from root to tip to divide the alignment into a hierarchy of groups and subgroups, ET asks each time whether a column's residues are invariant within individual groups, even if they vary among them. In its most straightforward implementation, a residue has rank 1 if it is invariant in the entire family. It has rank 2 if it is different between the first two branches of the tree, but is invariant within each one. It has rank 3 if it is variable between at least two of the first three branches but is invariant within all three and so forth. If the rank of a residue is low, its natural variations are necessarily always associated with major evolutionary branchpoints suggesting that they matter. If it is high, the opposite is true: it varies even between nearly identical species suggesting that

residue matters little. This simple ranking procedure assigns to every residue a relative rank of importance during evolution. Remarkably, top-ranked residues display a number of highly desirable features: they cluster spatially in native structures; they map out the functional sites of a protein and they indicate the determinants of functional specificity.

The universal 3D clustering of evolutionary important residues in protein structures is now established. It arises by mapping the better ranked ET residues onto the protein structure in order to visualize unusual areas of the protein where evolutionary influential residues co-locate. Such co-localization was seen first in protein that binds other proteins, such as SH2 and SH3, and in proteins that binds DNA [31, 66]. More recently, clustering of top-ranked trace residues was observed to be significant, compared to a random draw of residues, in 45 out of 46 proteins [32] and again in 92% of a larger test set of 79 proteins [67]. Other studies in other laboratories yielded similar conclusions [65, 68, 69]. These many observations were finally formalized and generalized to the entire PDB to show that clustering of evolutionarily important residues identified by ET is a universal property of protein [70].

The ability of ET clusters to predict functional sites has also been repeatedly shown. For example, *bona fide* predictions of interfaces in the G protein α-subunit [71, 72], in regulators of G protein signaling [73, 74], in the nuclear transport carrier NTF2 [75], have been validated by mutations and in the case of RGS by crystallography as well. More generally, larger scale retrospective control studies show that ET successfully narrows the docking searches for binding sites, and that its predicted functional sites overlap those that are known structurally [32], and biochemically as well [62]. In [32], the trace cluster contacted the ligand in 37 out of 38 protein–ligand complexes. Considering a collection of 79 proteins, the functional site overlap was shown to be statistically significant in majority of proteins (range 79–100%, depending on the definition of overlap and the statistics used) [62].

Another key property of ET methods is to identify the biologically meaningful residue variations that underlie functional specificity. This was first suggested by the correlation across the protein–DNA interface between the variation of trace residues and of their contact response element base pairs [66]. It was tested experimentally by swapping trace residues between two members of the RGS family and thus exchanging their activity. Strikingly, the same type of ET-guided redesign of transcription factors from the basic helix–loop–helix family enabled the reciprocal swap of developmental proneural pathway programs from a frog oocyte to a fly embryo [76]. Experiments are now underway to further test predictions of specificity determinants in G protein-coupled receptors [77]. Other studies along these lines were aimed at predicting protein functional subtypes [78] and DNA-binding specificity [79].

The ET approach is not without difficulties, however. Accuracy depends in essential ways on the quality of the multiple sequence alignment, like other techniques, and of the corresponding phylogenetic tree. Both can be controversial and limit automated analysis [80]. One may also wish to account for different evolutionary rates [63]. For now, an approach that performs as well as, or even better, than manual tracing combines entropy-related methods with the use of tree phylogeny information: the importance of a given alignment column position is calculated by a weighted sum over all evolutionary groups of the entropy of the position in each one [81]. Another caveat is that, since the reliability of trace predictions is inferred from their unusual 3D clustering pattern, it is impossible to assess trace quality in the absence of a structure.

Evolutionary correlations

Instead of probing correlations between residue variation and tree variation, other ideas consider directly the correlations between trees or between residues. First, as noted above, interacting binding partners seem to coevolve so that variations linked to phylogenetic changes in one partner are reflected by changes in the other. Accordingly, the phylogenetic trees of interacting protein should be mirror trees, meaning that the trees obtained by reducing alignments of the two proteins to the set of organisms common to both should be superimposable, ideally. This was quantified by calculating the correlations between the distance matrices of the trees [82, 83]. These matrices contain distances between all possible protein pairs from the multiple sequence alignment. One study of six families of ligand–receptor pairs was thus able to find 79% of all known binding partners. This approach was extended to a variety of proteins [84]: 13 proteins with two interacting domains, 53 *Escherichia coli* proteins and to a whole genome of 4300 *E. coli* proteins. The authors assert that the similarity between phylogenetic trees can be used as a predictor of protein–protein interaction with more than 66% of true positives detected at correlations >0.8. If so, the correlations between distance matrices seem to be a good statistical indicator of protein interactions. An important limitation is that the correlations are between distance matrices rather than between the actual phylogenetic trees.

Correlated mutations for pairs of multiple sequence alignments have been used to detect interacting protein pairs or their domains. This '*in silico* two-hybrid method' has been used to predict physically interacting proteins in *E. coli* [85]. Its main limitation seems to be the availability of large multiple sequence alignments for each pair of proteins.

Alternatively, it is also possible to consider correlations directly between interacting residue by alignment columns. The idea is that changes in one column will be matched in another column if the residues are coupled functionally [86] or energetically [87]. For example, within a protein, the average pairwise distance between the residues on the structure tends to be smaller for correlated pairs [88], and this can be used to discriminate between physiological and crystallographic and incorrect 3D configurations of protein domains [89].

There are several unresolved issues related to the use of correlated mutations in MSAs. Once a correlation between residue pairs is found, it is not trivial to decide whether it reflects functional correlations or just a phylogenetic

relationship [86]. Also, a comparison of several algorithms based on covariance in MSAs (among them, the methods used in [87, 88]) showed a surprising lack of agreement in prediction of important residues [90–92]. This may represent many complementary views of a large set of cooperative interactions. Background residue conservation has been suggested as a possible reason for the lack of agreement between correlated mutations algorithms [90] (different covariance algorithms deal with conserved alignment columns in different ways: most tend to agree for the columns which are highly covarying or highly conserved). A comparison of predictions based on correlated mutations with the experimentally determined coupling between residues (using double mutant cycles) shows that these algorithms can find residue pairs which are physically close [91]: however, a direct correlation between these predictions and thermodynamically coupled residues has been found in only one of the four data sets.

Predicting interacting residues using machine learning

Given the variety of these approaches, it is important to consider current attempts to combine physical, chemical and evolutionary analysis of residues at protein–protein interfaces. Thus far, neural networks have been trained to classify with an accuracy of 70% whose surface residues do or do not interact with another protein based on their side chain and their neighbors [34, 93, 94]. While this is promising, it should be mentioned that the data set used in [93] contained some homodimers, while about half of the data set used in [34] were the interfaces between heavy and light chains of antibodies (both of these cases should make the classification easier: homodimer interfaces are on average hydrophobic, and the interface between heavy and light antibody chains is much more conserved than the antibody–antigen interface). Moreover, these results may not improve on the ET approach, which has already been validated experimentally.

There have been related attempts to predict interacting residues but without any structure information. A support vector machine and a Bayesian classifier were used to classify surface residues into interacting and noninteracting sets by exploiting the fact that interface residues tend to form clusters in the primary amino acid sequences in [95]. While none of the features were based on the structural parameters, this work still required to obtain the starting set of surface residues since only those were used in classification. On the set of 77 complexes they achieve accuracy of above 70%. Only the information extracted from primary sequence and a neural network were used on a set of 333 transient proteins [35]. Without giving the overall accuracy, the authors state that for the best 34 cases (out of 333) 94% of predictions were confirmed experimentally. Recently, the same problem was tackled by considering evolutionary importance of sequence residues. A support vector machine-based prediction of interface residues reached the accuracy of 64% [96]. The full range of achieved specificity–sensitivity results is shown in figure 1.

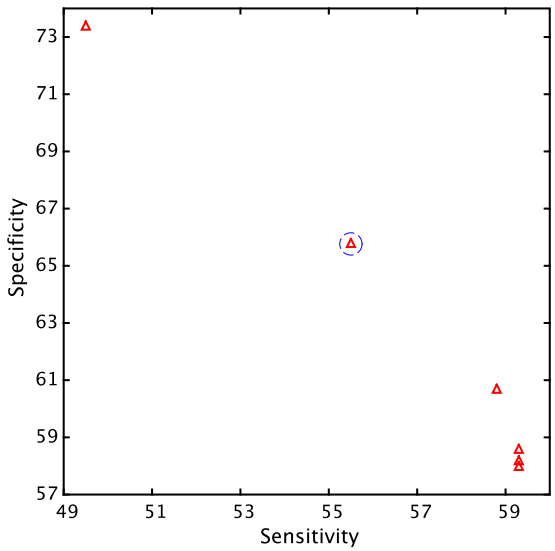

Figure 1. Specificity versus sensitivity plot for prediction of protein interaction sites without using the structure, as obtained in [96]. The circled point corresponds to accuracy of 64%.

Another support vector machine-based study used various combinations of feature vectors to predict interacting residues, with or without the use of structures [97]. The results were slightly improved compared to [93, 34], while the sequence-only-based prediction resulted in the average prediction performance of ∼10% higher than random.

Summary and outlook

The characterization of the critical determinants of protein–protein interactions holds the key to understanding the molecular basis for the assembly of macromolecular machines and of cellular networks. Not only would it allow the design of drugs or of engineered proteins that selectively block or trigger entire signaling pathways, but it would allow us to predict more reliably quaternary structures, which are difficult to derive experimentally. Bulk statistical analysis of interfaces have not led to accurate predictions, perhaps largely due to the heterogeneity of interfaces, where only a few residues forming the hot spots while the others contribute little or not all. But aside from the study of amino acid composition, physical models of binding free energy, and evolutionary models of correlation among residues, trees and structures are much more promising. In particular evolutionary tracing has been extensively validated, is now fully automated, and was shown to accurately predict both the location of interfaces and the determinants of their specificity. Together with the correlation between hot spots and structurally conserved residues, it seems only a matter of time before an accurate PDB-wide view of protein interfaces emerges. The ultimate goal of these studies is the prediction of protein–protein interactions. Several of the methods presented here are difficult to benchmark against small and subjective data sets. This also can only improve in future as a result of structural genomics and physical modeling.

Acknowledgments

The authors gratefully acknowledge support from the March of Dimes (MOD FY03-93), the National Science Foundation (DBI-0318415), and the National Institute of Health (RO1 GM066099).

References

[1] Temussi P A, Masino L and Pastore A 2003 From Alzheimer to Huntington: why is a structural understanding so difficult? *EMBO J.* **22** 355–61

[2] Arkin M R and Wells J A 2004 Small-molecule inhibitors of protein–protein interactions: progressing towards the dream *Nat. Rev. Drug Discov.* **3** 301–17

[3] Schreiber S L and Crabtree G R 1992 The mechanism of action of Cyclosporin A and FK506 *Immunol. Today.* **13** 136–42

[4] Nagai Y, Fujikake N, Ohno K, Higashiyama H, Popiel H A, Rahadian J, Yamaguchi M, Strittmatter W J, Burke J R and Toda T 2003 Prevention of polyglutamine oligomerization and neurodegeneration by the peptide inhibitor QBP1 in *Drosophila Hum. Mol. Genet.* **12** 1253–9

[5] Vajda S and Camacho C J 2004 Protein–protein docking: is the glass half-full or half-empty? *Trends Biotechnol.* **22** 110–6

[6] Wodak S J and Mendez R 2004 Prediction of protein–protein interactions: the capri experiment, its evaluation and implications *Curr. Opin. Struct. Biol.* **14** 242–9

[7] Rain J C *et al* 2001 The protein–protein interaction map of *Helicobacter pylori Nature* **409** 211–5

[8] Uetz P *et al* 2000 A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae Nature* **403** 623–7

[9] Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M and Sakaki Y 2001 A comprehensive two-hybrid analysis to explore the yeast protein interactome *Proc. Natl Acad. Sci.* **98** 4569–74

[10] Giot L *et al* 2003 A protein interaction map of *Drosophila melanogaster Science* **302** 1727–36

[11] Gavin A C *et al* 2002 Functional organization of the yeast proteome by systematic analysis of protein complexes *Nature* **415** 141–7

[12] Ho Y *et al* 2002 Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry *Nature* **415** 39–44

[13] Aloy P and Russell R B 2004 Ten thousand interactions for the molecular biologist *Nat. Biotechnol.* **22** 1371–21

[14] Westbrook J, Feng Z, Chen L, Yang H and Berman H M 2003 The protein data bank and structural genomics *Nucleic Acids Res.* **31** 489–91

[15] Sali A, Glaeser R, Earnest T and Baumeister W 2004 From words to literature in structural proteomics *Nature* **422** 216–25

[16] Chothia C 1992 One thousand families for the molecular biologist *Nature* **357** 543–4

[17] Andreeva A, Howorth D, Brenner S E, Hubbard T J, Chothia C and Murzin A G 2004 SCOP database in 2004: refinements integrate structure and sequence family data *Nucleic Acids Res.* **32** D226–9

[18] Russell R B, Alber F, Aloy P, Davis F P, Korkin D, Pichaud M, Topf M and Sali A 2004 A structural perspective on protein–protein interactions *Curr. Opin. Struct. Biol.* **14** 311–24

[19] Braig K, Otwinowski Z, Hegde R, Boisvert D C, Joachimiak A, Horwich A L and Sigler P B 1994 The crystal structure of the bacterial chaperonin GroEL at 2.8 A *Nature* **371** 578–86

[20] Lowe J, Stock D, Jap B, Zwickl P, Baumeister W and Huber R 1995 Crystal structure of the 20S proteasome from the archaeon *T. acidophilum* at 3.4 Å resolution *Science* **268** 533–9

[21] Zhang G, Campbell E A, Minakhin L, Richter C, Severinov K and Darst S A 1999 Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution *Cell* **98** 811–24

[22] Ban N, Nissen P, Hansen J, Moore P B and Steitz T A 2000 The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution *Science* **289** 905–20

[23] Fiaux J, Bertelsen E B, Horwich A L and Wuthrich K 2002 NMR analysis of a 900K GroEL GroES complex *Nature* **418** 207–11

[24] Ben-Shem A, Frolow F and Nelson N 2003 Crystal structure of plant photosystem I *Nature* **426** 630–5

[25] Liu Z, Yan H, Wang K, Kuang T, Zhang J, Gui L, An X and Chang W 2004 Crystal structure of spinach major light-harvesting complex at 2.72 A resolution *Nature* **428** 287–92

[26] Egea P F, Shan S O, Napetschnig J, Savage D F, Walter P and Stroud R M 2004 Substrate twinning activates the signal recognition particle and its receptor *Nature* **427** 215–21

[27] Grimes J, Basak A K, Roy P and Stuart D 1995 The crystal structure of bluetongue virus VP7 *Nature* **373** 167–70

[28] Nakagawa A *et al* 2003 The atomic structure of rice dwarf virus reveals the self-assembly mechanism of component proteins *Structure* **11** 1227–38

[29] Chothia C and Lesk A M 1986 The relation between the divergence of sequence and structure in proteins *EMBO J.* **5** 823–6

[30] Aloy P, Ceulemans H, Stark A and Russell R B 2003 The relationship between sequence and interaction divergence in proteins *J. Mol. Biol.* **332** 989–98

[31] Lichtarge O, Bourne R and Cohen F E 1996 An evolutionary trace method defines binding surfaces common to protein families *J. Mol. Biol.* **257** 342–58

[32] Madabushi S, Yao H, Marsh M, Kristensen D M, Philippi A, Sowa M E and Lichtarge O 2002 Structural clusters of evolutionary trace residues are statistically significant and common in proteins *J. Mol. Biol.* **316** 139–53

[33] Valdar W and Thornton J 2001 Conservation helps identify biologically relevant crystal contacts *J. Mol. Biol.* **313** 399–416

[34] Fariselli P, Pazos F, Valencia A and Casadio R 2002 Prediction of protein–protein interaction sites in heterocomplexes with neural networks *Eur. J. Biochem.* **269** 1356–61

[35] Ofran Y and Rost B 2003 Predicted protein–protein interaction sites from local sequence information *FEBS Lett.* **544** 236–9

[36] Nooren I M and Thornton J M 2003 Diversity of protein–protein interactions *EMBO J.* **22** 3486–92

[37] Jones S and Thornton J M 1996 Principles of protein–protein interactions *Proc. Natl Acad. Sci.* **93** 13–20

[38] Chakrabarti P and Janin J 2002 Disecting protein–protein recognition sites *Proteins: Struct. Funct. Genet.* **47** 334–43

[39] Ponstingl H, Kabir T, Gorse D and Thornton J M 2005 Morphological aspects of oligomeric protein structures *Prog. Biophys. Mol. Biol.* **89** 9–35

[40] Rodier F, Janin J, Bahadur R P and Chakrabarti P 2003 A dissection of specific and non-specific protein–protein interfaces *J. Mol. Biol.* **336** 943–55

[41] Ofran Y and Rost B 2003 Analysing six types of protein–protein interfaces *J. Mol. Biol.* **325** 377–87

[42] Nooren I M and Thornton J M 2003 Structural characterisation and functional significance of transient protein–protein interactions *J. Mol. Biol.* **325** 991–1018

[43] Rodier F, Janin J, Bahadur R P and Chakrabarti P 2003 Dissecting subunit interfaces in homodimeric proteins *Proteins: Struct. Funct. Genet.* **53** 708–19

[44] Sheinerman F B and Honig B 2002 On the role of electrostatic interactions in the design of protein–protein interfaces *J. Mol. Biol.* **318** 161–77

[45] Clackson T and Wells J A 1995 A hot spot of binding energy in a hormone–receptor interface *Science* **267** 383–6

[46] Bogan A A and Thorn K S 1998 Anatomy of hot spots in protein interfaces *J. Mol. Biol.* **280** 1–9

[47] Kortemme T and Baker D 2002 A simple physical model for binding energy hot spots in protein–protein complexes *Proc. Natl Acad. Sci. USA* **99** 14116–21

[48] Massova I and Kollman P A 1999 Computational alanine scanning to probe protein–protein interactions: a novel approach to evaluate binding free energies *J. Am. Chem. Soc.* **121** 8133–43

[49] Verkhivker G H, Bouzida D, Gehlhaar D K, Rejto P A, Freer S T and Rose P W 2002 Monte Carlo simulations of the peptide recognition at the consensus binding site of the constant fragment of human immunoglobulin g: the energy landscape analysis of a hot spot at the intermolecular interface *Proteins: Struct. Funct. Genet.* **48** 539–57

[50] Ma B, Elkayam T, Wolfson H and Nussinov R 2003 Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces *Proc. Natl Acad. Sci.* **100** 5772–7

[51] Halperin I, Wolfson H and Nussinov R 2004 Protein–protein interactions: coupling of structurally conserved residues and of hot spots across interfaces. implication for docking *Structure* **12** 1027–38

[52] Fernández A and Scheraga H A 2003 Insufficiently dehydrated hydrogen bonds as determinants of protein interactions *Proc. Natl Acad. Sci. USA* **100** 113–8

[53] Watts D J and Strogatz S H 1998 Collective dynamics of small-world networks *Nature* **393** 440–2

[54] Amitai G, Shemesh A, Sitbon E, Shklar M, Netanely D, Venger I and Pietrokovski S 2004 Network analysis of protein structures identifies functional residues *J. Mol. Biol.* **344** 1135–46

[55] del Sol A and O'Meara P 2005 Small-world network approach to identify key residues in protein–protein interaction *Proteins* **58** 672–82

[56] del Sol A and O'Meara P 2005 Topology of small-world networks of protein–protein complex structures *Bioinformatics* **21** 1311–5

[57] Valdar W and Thornton J 2001 Protein–protein interfaces: analysis of amino acid conservation in homodimers *Proteins: Struct. Funct. Genet.* **42** 108–24

[58] Caffrey D R, Somaroo S, Hughes J D, Mintseris J and Huang E S 2004 Are protein–protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci.* **13** 190–202

[59] Elcock A H and McCammon J A 1998 Identification of protein oligomerization states by analysis of interface conservation *Proc. Natl Acad. Sci.* **98** 2990–4

[60] Sowa M E and Lichtarge O 2002 Evolutionary predictions of binding surfaces and interactions *Curr. Opin. Struct. Biol.* **12** 21–7

[61] Lichtarge O, Yao H, Kristensen D M, Madabushi S and Mihalek I 2003 Accurate and scalable identification of functional sites by evolutionary tracing *J. Struct. Funct. Genomics* **4** 159–66

[62] Yao H, Kristensen D M, Mihalek I, Sowa M E, Shaw C, Kimmel M, Kavraki L and Lichtarge O 2003 An accurate, sensitive, and scalable method to identify functional sites in protein structures *J. Mol. Biol.* **326** 255–61

[63] Pupko T, Bell R E, Mayrose I, Glaser F and Ben-Tal N 2002 Rate4site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues *Bioinformatics* **18** (Suppl. 1) s71–7

[64] Glaser F, Pupko T, Paz I, Bell R E, Bechor-Shental D, Martz E and Ben-Tal N 2003 Consurf: identification of functional regions in proteins by surface-mapping of phylogenetic information *Bioinformatics* **19** 163–4

[65] Landgraf R, Xenarios I and Eisenberg D 2001 Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins *J. Mol. Biol.* **307** 1487–502

[66] Lichtarge O, Yamamoto K R and Cohen F E 1997 Identification of functional surfaces of the zinc binding domains of intracellular receptors *J. Mol. Biol.* **274** 325–37

[67] Schueler-Furman O and Baker D 2003 Conserved residue clustering and protein structure prediction *Proteins: Struct. Funct. Genet.* **52** 225–35

[68] Armon A, Graur D and Ben-Tal N 2001 ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information *J. Mol. Biol.* **313** 447–63

[69] Neuvirth H, Raz R and Schreiber G 2004 ProMate: a structure based prediction program to identify the location of protein–protein binding sites *J. Mol. Biol.* **338** 181–99

[70] Mihalek I, Res I, Yao H and Lichtarge O 2003 Combining inference from evolution and geometric probability in protein structure evaluation *J. Mol. Biol.* **331** 263–379

[71] Lichtarge O, Bourne H R and Cohen F E 1996 Evolutionarily conserved G$\alpha\beta\gamma$ binding surfaces support a model of the G protein–receptor complex *Proc. Natl Acad. Sci. USA* **93** 7507–11

[72] Onrust R, Herzmark P, Chi P, Garcia P D, Lichtarge O, Kingsley C and Bourne H R 1997 Receptor and $\beta\gamma$ binding sites in the alpha subunit of the retinal G protein transducin *Science* **275** 381–4

[73] Sowa M E, He W, Wensel T G and Lichtarge O 2000 A regulator of G protein signaling interaction surface linked to effector specificity *Proc. Natl Acad. Sci. USA* **97** 1483–8

[74] Sowa M E, He W, Slep K C, Kercher M A, Lichtarge O and Wensel T G 2001 Prediction and confirmation of a site critical for effector regulation of RGS domain activity *Nat. Struct. Biol.* **8** 234–7

[75] Cushman I, Bowman B R, Sowa M E, Lichtarge O, Quiocho F A and Moore M S 2004 Computational and biochemical identification of a nuclear pore complex binding site on the nuclear transport carrier NTF2 *J. Mol. Biol.* **344** 303–10

[76] Quan X J, Denayer T, Yan J, Jafar-Nejad H, Philippi A, Lichtarge O, Vleminckx K and Hassan B A 2004 Evolution of neural precursor selection: functional divergence of proneural proteins *Development* **131** 1679–89

[77] Madabushi S, Gross A K, Philippi A, Meng E C, Wensel T G and Lichtarge O 2004 Evolutionary trace of g protein-coupled receptors reveals clusters of residues that determine global and class-specific functions *J. Biol. Chem.* **279** 8126–32

[78] Hannenhalli S S and Russell R B 2000 analysis and prediction of functional sub-types from protein sequence alignments *J. Mol. Biol.* **303** 61–76

[79] Mirny L A and Gelfand M S 2002 Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors *J. Mol. Biol.* **321** 7–20

[80] Abascal F and Valencia A 2003 Automatic annotation of protein function based on family identification *Proteins 2003* **53** 683–92

[81] Mihalek I, Res I and Lichtarge O 2004 A family of evolution–entropy hybrid methods for ranking protein residues by importance *J. Mol. Biol.* **336** 1265–82

[82] Goh C-S, Bogan A A, Joachimiak M, Walther D and Cohen F E 2000 Co-evolution of proteins with their interaction partners *J. Mol. Biol.* **299** 283–93

[83] Goh C-S and Cohen F E 2002 Co-evolutionary analysis reveals insights into protein–protein interactions *J. Mol. Biol.* **324** 177–92
[84] Pazos F and Valencia A 2001 Similarity of phylogenetic trees as indicator of protein–protein interactions *Protein Eng.* **14** 609–14
[85] Pazos F and Valencia A 2002 In silico two-hybrid system for the selection of physically interacting protein pairs *Proteins: Struct. Funct. Genet.* **47** 219–27
[86] Tillier E R and Lui T W 2003 Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments *Bioinformatics* **19** 750–5
[87] Lockless S W and Ranganathan R 1999 Evolutionarily conserved pathways of energetic connectivity in protein families *Science* **286** 295–9
[88] Göbel U, Sander C, Scnneider R and Valencia A 1994 Correlated mutations and residue contacts in proteins *Proteins* **18** 309–17
[89] Pazos F, Helmer-Citterich M, Ausiello G and Valencia A 1997 Correlated mutations contain information about protein–protein interaction *J. Mol. Biol.* **271** 511–23
[90] Fodor A A and Aldrich R W 2004 Influence of conservation of amino acid covariance in multiple sequence alignments *Proteins: Struct. Funct. Bioinf.* **56** 211–21
[91] Fodor A A and Aldrich R W 2004 On evolutionary conservation of thermodynamic coupling in proteins *J. Biol. Chem.* **276** 19046–50
[92] Dekker J P, Fodor A, Aldrich R W and Yellen G 2004 A perturbation-based method for calculating explicit likelihood of evolutionary co-variance in multiple sequence alignments *Bioinformatics* **20** 1565–72
[93] Zhou H X and Shan Y 2001 Prediction of protein interaction sites from sequence profile and residue neighbor list *Proteins* **44** 336–43
[94] Gutteridge A, Bartlett G J and Thornton J M 2003 Using a neural network and spatial clustering to predict the location of active sites in enzymes *J. Mol. Biol.* **330** 719–34
[95] Yan C, Dobbs D and Honavar V 2004 A two-stage classifier for identification of protein–protein interface residues *Bioinformatics* **20** i371–78
[96] Res I, Mihalek I and Lichtarge O 2005 An evolution based classifier for prediction of protein interfaces without using protein structures *Bioinformatics* **21** 2496–501
[97] Koike A and Takagi T 2004 Prediction of protein–protein interaction sites using support vector machines *Protein Eng. Des. Sel.* **17** 165–73