Review in Advance first posted online on February 5, 2008. (Minor changes may still occur before final publication online and in print.)

Analyzing Protein Interaction Networks Using Structural Information

Christina Kiel,[1] Pedro Beltrao,[2] and Luis Serrano[1]

[1]EMBL-CRG Systems Biology Unit, Center de Regulacio Genomica, Barcelona 08003, Spain; email: christina.kiel@crg.es, luis.serrano@crg.es

[2]European Molecular Biology Laboratory, 69117 Heidelberg, Germany; email: beltrao@embl.de

Annu. Rev. Biochem. 2008. 77:5.1–5.27

The *Annual Review of Biochemistry* is online at biochem.annualreviews.org

This article's doi: 10.1146/annurev.biochem.77.062706.133317

0066-4154/08/0707-0001$20.00

Key Words

interface modeling, interaction types, protein complexes, structural proteomics

Abstract

Determining protein interaction networks and predicting network changes in time and space are crucial to understanding and modeling a biological system. In the past few years, the combination of experimental and computational tools has allowed great progress toward reaching this goal. Experimental methods include the large-scale determination of protein interactions using two-hybrid or pull-down analysis as well as proteomics. The latter one is especially valuable when changes in protein concentrations over time are recorded. Computational tools include methods to predict and validate protein interactions on the basis of structural information and bioinformatics tools that analyze and integrate data for the same purpose. In this review, we focus on the use of structural information in combination with computational tools to predict new protein interactions, to determine which interactions are compatible with each other, to obtain some functional insight into single and multiple mutations, and to estimate equilibrium and kinetic parameters. Finally, we discuss the importance of establishing criteria to biologically validate protein interactions.

Contents

1. INTRODUCTION

Determining protein-protein and protein-ligand (i.e., DNA, RNA, and small-molecule) interactions is essential if we want a systems biology understanding of living organisms and/or cells. Much effort has been made in the past years in experimentally determining protein-protein interactions on a large scale. Among the different methods used, we mention the two-hybrid approach with all its varieties [see Rual et al. (1), Uetz et al. (2), Ito et al. (3), and Stelzl et al. (4)]; the pull-down experiments using different tags (5–7); and more recently protein chips [8; reviewed in Kunk & Snyder (9)]. Each of the different methods have their advantages and disadvantages, which we do not discuss here [see Cusick et al. (10), Fields (11), and Berggard et al. (10–12)]. In parallel over many years different groups have attempted to predict protein-protein interactions using homology modeling and docking software. However, although we have seen significant progress in the field, the accuracy of methods is not good enough to attempt large-scale prediction of interaction networks [Aloy & Russell (17)]. With the advent and progress of different structural genomics projects, a new methodology, complementary to the above activities termed "interface modeling," was developed. This methodology depends on the availability of several good-quality three-dimensional (3D) structures, on correct (structure-based) sequence alignments, and on careful structural inspection of domains and sequences (16). Essentially, it needs the structure of a complex at high resolution, and it is based on the fact that proteins belonging to the same family if they interact they normally do it the same way and using similar positions (15), thus avoiding the docking problem (**Figure 1**) (18). Other simplifications are that only the interacting secondary structure elements are considered and the target sequences modeled on them (19, 61; G. Fernandez, P. Beltrao, L. Serrano, submitted for publication). Even simpler versions of this method just consider the nature of the residues, which form the complex interface and score the putative complex of related proteins (15). Complementary to this approach, other methodologies, such as the molecular dynamics of protein-peptide complexes, have been used to identify putative peptide ligands for a globular domain (110). However, even a successful prediction of an interaction from a biophysical point of view (i.e., the two proteins bind with nanomolar affinity) does not guarantee that the interaction is physiological. Thus, other computational

Sequence alignments: the arrangement of amino acid sequences from organisms in a way that aligns areas sharing common properties

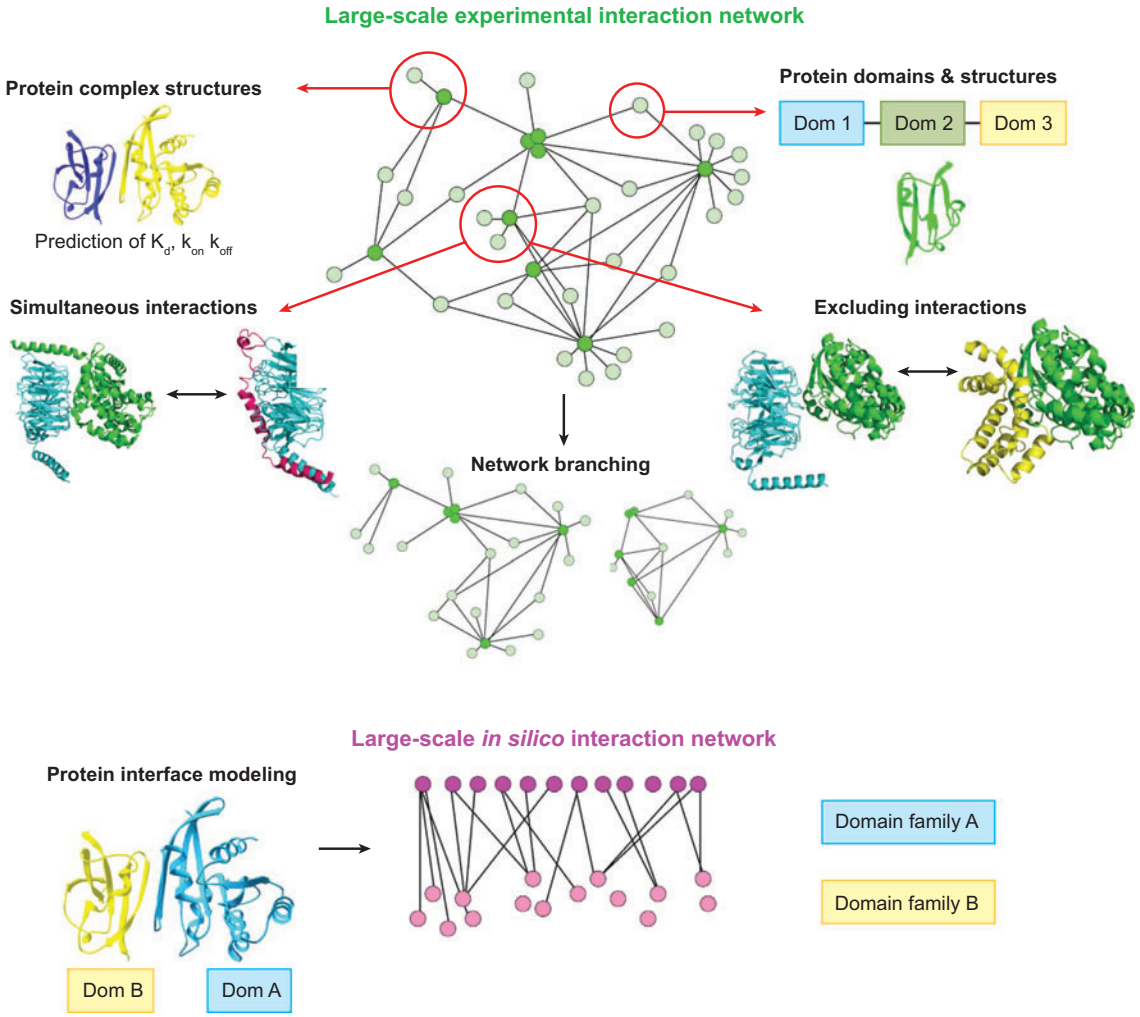


Figure 1

Summary of the main concepts discussed in this review. Structural biology in combination with computational tools can be used to validate interactions that have been found experimentally in a complex and to predict new protein interactions in silico. Furthermore, structural information is used to determine which interactions are compatible with each other and which ones exclude each other and to estimate equilibrium and kinetic constants.

tools are needed to add some biological credibility to the predicted interactions.

Information regarding protein interactions is often depicted in interaction network diagrams where usually proteins are represented by dots and connections to interacting proteins are shown as lines. This concept is very powerful in order to explore global properties of network topologies. However, treating proteins as dots neglects the important biophysical properties of proteins and protein complexes, which are crucial in mediating their cellular function: Do the interacting proteins form transient or stable protein complexes? Which domains mediate the interaction? What are the affinities and kinetic constants? Which interactions exclude each other and which ones can happen

Transient protein complex: short-lived protein complexes; the high dissociation rate constant makes binding usually weak

Stable protein complex: a group of two or more permanently associated proteins; they are a form of quaternary structure

UBD: ubiquin-like domain

simultaneously? What will be the effects of particular mutations, found in human populations, on the interactions?

Traditional high throughput methods do not answer these questions, and for example, they cannot distinguish between incompatible complexes sharing one or more components (13). This has prompted various groups to develop methodologies to distinguish them. For example, when doing pull-down experiments, one can tag all the elements of the complexes, and by finding particular associations, in principle, it should be possible to decide about incompatibilities (7). Also, another type of information, such as colocalization, can be used to distinguish between different complexes sharing common elements (14). Alternatively, partial or total answers to these questions can be obtained from structural information as pointed out by the pioneering work of Aloy & Russell (15, 17), which postulated that "Structural details can turn abstract system representation into models that more accurately reflect reality." Regarding which interactions exclude each other and which ones can happen simultaneously the M.B. Gerstein lab (104) was the first to apply structural information on a large-scale experimental network (the human proteome). By using a 3D-structural exclusion, they could distinguish overlapping from nonoverlapping interfaces and thus provided evolutionary insights into network evolution (104). Similarly, this information of conserved domain-domain interactions (and interface types) can be used to find possible binary interactions if experiments discover several proteins in a complex. Furthermore, the idea of testing whether domain interactions are simultaneously possible or whether they exclude each other can increase the information of a given network, allowing subdivision into mutually excluding complexes (A. Campagna, C. Kiel & L. Serrano, manuscript in preparation). With respect to the prediction of binding parameters, the pioneering work of Schreiber and coworkers (51) showed that, using structural information, it was possible obtain an estimate of the k_{on} for the formation of a protein complex. This work was later followed by other groups (19) and ours, showing that quite accurate estimates on k_{on} can be obtained for mutant variants and for members of the same families within protein complexes (C. Kiel, N. Kuemmerer, L. Serrano, manuscript in preparation). Baker's group showed that it was possible to obtain some reasonable estimate on K_d values for protein interactions (20). More recently our group has shown that for the Ras-Rbd family of protein complexes computer methods can predict quite accurate estimates of the K_d values (19). Finally protein design algorithms, such as FoldX and others (86–91) based on X-ray structures and interface models, can reliably predict the effect of mutations on protein complex stabilities, which allow interpretation of the effect of SNP variants on protein functionality and complex formation (21).

In this review, we introduce and summarize the main applications of structural information and interface modeling to predict and analyze protein interaction networks, as summarized above (**Figure 1**), and show the strengths and weaknesses of the methodology. We do not discuss other approaches for the prediction of protein-protein interactions based on docking (18) nor the nonhomology-based structural prediction methods (22). Similarly, we do not review current methodologies for structure prediction (23).

2. MACROVIEW OF STRUCTURE

A protein (specially in Eukaryotes) usually consists of many domains, and very often one domain can interact with different types of domains using different areas of its surface, as shown for the Ras-like G domain fold, which can interact with many different partner domains of various structures (**Figure 2*a***). However, proteins belonging to the same family [i.e., ubiquitin-like domain (UBD)] usually interact in a similar way with a particular protein family (i.e., Ras G proteins) (17) (**Figure 2*b***). In all Ras/UBD complexes,

the binding mode is similar and involves mainly two antiparallel ß-sheets of the UBD and the Ras domain, respectively, as well as the first helix of the UBD (19). However, a recent large-scale structural classification of protein-protein interfaces has shown that 24% of protein hetero-interactions between homologues associate in multiple orientations (24). Thus, further information on the target protein families is needed before we can assume that they will interact the same way as their homologues.

Structural proteomics efforts have already found 700–800 different folds (25) of the predicted 1000 domain folds in nature (26) and ~2000 of the predicted 10,000 domain-domain interaction types (25). It is estimated that more than 20 years is needed to obtain a representative structure for each interaction type (25). One important attempt to reach this goal is the so called "Pfam5000" structural genomics effort, which aims to solve the structure of the 5000 most important domain families found in the Pfam database (27). The identification of domain families from sequences is a very mature field in bioinformatics. In **Table 1**, we list the current Web tools and databases used to find domains on the basis of sequence similarity, protein order/disorder prediction, domain databases and classifications, and domain interaction type databases (28–48).

However, it is not the fold per se that determines whether two domains can interact, but rather epitopes on the surface of the domains (49) (**Figure 2*c***). Thus, although proteins of the same families interact through equivalent surfaces, there is no guarantee that any Ras protein will interact with an effector containing a UBD. It is often the case that one or two differences at key positions of the interaction surface are enough to prevent binding (i.e., Ras and Rap with RalGDS and Raf). In fact, it has been shown for Ras-effector interactions that the thermodynamic properties and kinetic properties can vary by orders of magnitude (50), as shown for Ras in complex with RalGDS-RA, Raf-RBD, and PLCe-RA2 (**Figure 2*d–f***). Depending on the residues on the UBD fold, the affinities in complex with Ras can be in the nano- or micromolar range. Enthalpy (ΔH) and entropy ($T\Delta S$) contributions to binding energy can both be favorable, or large favorable enthalpy contributions are compensated by small, but unfavorable, entropy contributions. Similarly, the contributions of association and dissociation rate constants (k_{on} and k_{off}) to affinity ($K_d = k_{off}/k_{on}$) vary significantly. Estimates for k_{on} can be readily obtained from the structure of the complexes (51, 52). It is not yet clear to what extent kinetic constants have an important biological role, or whether it is the K_d that is important.

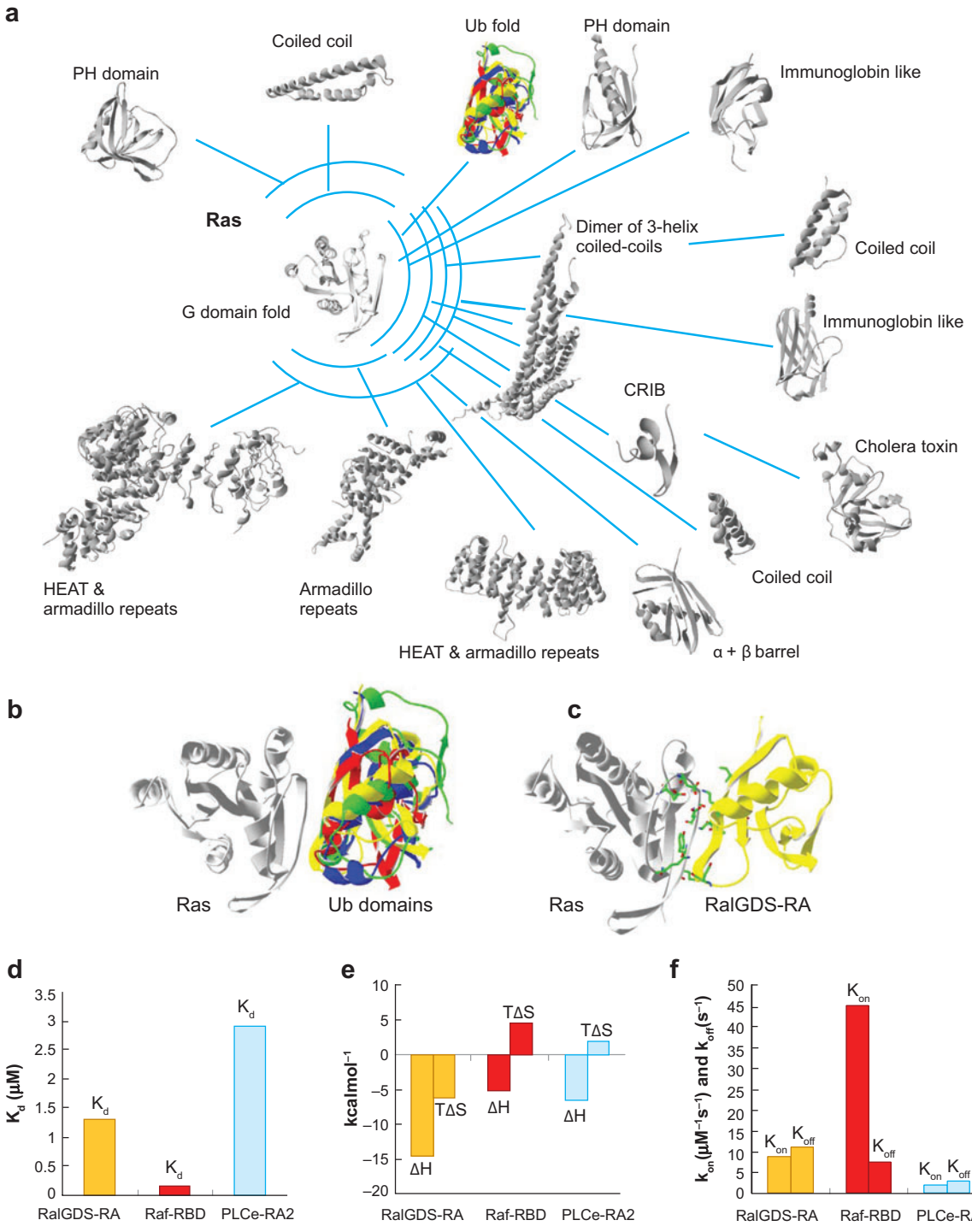
Table 1 Tools to analyze protein sequences, domains, and domain interaction types

Type and name of tool	References
Sequence similarity search	
BLAST/PSI-BLAST	Altschul et al. (28, 29)
Protein order and disorder prediction	
GlobPlot	Linding et al. (30)
IUPRed	Dosztány et al. (31, 32)
DisProt	Peng et al. (33)
VSL2	Peng et al. (148)
PrDOS	Ishida & Kinoshita (149)
POODLE-L	Hirose et al. (150)
Domain databases	
SMART	Schultz et al. (34), Letunic et al. (35)
Pfam	Bateman et al. (36)
PROSITE	Hulo et al. (37)
CDD	Marchler-Bauer et al. (38)
Domain classification	
CATH	Orengo et al. (39), Pearl et al. (40)
SCOP	Murzin et al. (41), Andreeva et al. (42)
Domain interaction type databases	
iPfam	Finn et al. (43)
3did	Stein et al. (44)
SCOPPI	Winter et al. (45)
PRISM	Ogmen et al. (46)
SNAPPI-DB	Jefferson et al. (47)
PIBASE	Davis & Sali (48)

In summary, although in the majority of the cases it is found that interaction surfaces tend to be conserved between domains belonging to the same families, it is the details of the interaction that determine the binding

RA: Ras-association domain

RBD: Ras-binding domain



affinity and kinetic properties. Thus, when predicting if two proteins will interact using the structural information of a complex involving related domains, an atomic model of the complex is needed. We discuss this process below with an emphasis on the difficulties associated with interface modeling.

3. HOMOLOGY INTERFACE MODELING

There are different methodologies to perform homology modeling [for reviews, see Goldsmith-Fischman & Honig (53) and Ginalski (54)], and different automatic servers are available, e.g., SWISS-MODEL, WHAT IF, and MODELLER (**http://swissmodel.expasy.org//SWISS-MODEL.html; http://swift.cmbi.kun.nl/WIWWWI/; http://www.salilab.org/modeller**) (55–57). When the sequence homology is large and the loops have the same length, homology modeling just involves replacing the residues, which are different in the template structure, with those of the sequence to be modeled. This is a side chain modeling problem, which has been tackled by a number of algorithms (58–60). Side chain modeling relies on the discretization of the conformational space of each amino acid into so-called rotamer states and on a search algorithm to find the best combination of these states. This is followed by energy evaluation to see if there is any structure incompatibility of the new sequence side chains with the rest of the protein, which could indicate that the backbone of the template structure should move to accommodate the new side chains. Sometimes even a single amino acid difference can slightly change the backbone. One example of this is a UBD sequence with a proline residue in a central position of a ß-strand or α-helix, which changes the backbone of these secondary structure elements (61) (see **Figure 3*a***). Another example is the stabilization of a long loop, which is involved in the interface between an E2 and an E3 RING (really interesting new gene) domain by a large bulky amino acid. A tryptophan residue in the E3 RING domain stabilizes the conformation of the loop, which is involved in a complex formation with its E2 partner. A RING sequence with a small residue at this position cannot be modeled reliably using this complex structure as a modeling template (62) because the loop is expected to have a different conformation (**Figure 3*b***). If placement of the new sequence on the structural template does not result in any major incompatibility, or if the problems lie far away from the interaction surface, then the model can be evaluated.

When the homology is low, threshold < 30%, [the so-called 30% rule (53), see the structural explanation for this by Chung & Subbiah (63)], and/or the loops have different lengths, the backbone is locally, or globally, different from that in the reference structure. For particular domain families, the sequence identity threshold may vary. In some

Binding affinity: the strength of noncovalent chemical binding between two proteins or ligands, measured by the dissociation constant (K_d) of the complex

Loop: loops in proteins are polypeptides connecting secondary structure elements, and they can vary in size and sequence between homologous proteins

Figure 2

Macroview of structure. (*a*) Different binding modes of Ras-effector interactions, where the G domain-like fold uses different areas of its surface to mediate binding to various domains. (*b*) Similar domains usually interact in a similar way, as shown by an overlay of four Ras-effector complexes. Ras is shown in gray, and the Ras binding domains of RalGDS, Raf, PLCe, and PI3-kinase are in yellow, red, blue, and green, respectively. (*c*) The detailed amino acid contacts in the interface of the two domains determine specificity and affinity toward other domains. Ras and the Ras-binding domain of RalGDS are shown. (*d*) The affinities of three Ras-binding domains in complex with Ras (49). (*e*) The thermodynamic properties of three Ras-binding domains in complex with Ras (49). (*f*) The kinetic properties of three Ras-binding domains in complex with Ras. Experimental binding information for panels *d*, *e*, and *f* was taken from Wohlgemuth et al. (49). Abbreviations: ΔH, enthalpy; k_{on}, association rate constant; k_{off}, dissociation rate constant; TΔS, entropy; Ub, ubiquitin.

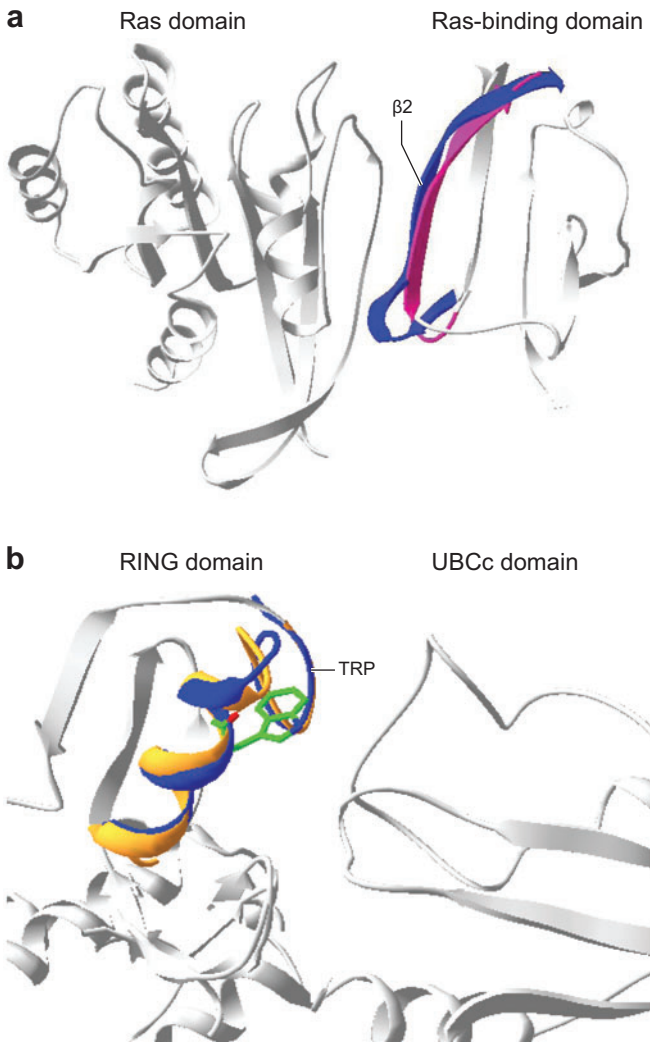


Figure 3

Examples of residues stabilizing a fold (backbone). (*a*) Overlay of β2 (β-strand 2) of two ubiquitin-like domains without and with a proline residue: β2 of Raf-RBD, Protein Data Bank (PDB) entry: 1GUA, and PLCe-RA2, (PDB entry: 2C5L). The β-strand of PLCe-RA2 has to bend to incorporate the proline residue. The complex of Ras and Raf-RBD is gray (PDB entry: 1GUA). (*b*) Stabilization of a loop by a tryptophan residue (W408) in RING domains (PDB entry: 1FBV). Ribbon representation of the cCbl-UBCH7 complex in gray (PDB entry: 1FBV). Included in this panel are the helix containing the TRP residue in 1FBV and the following loop as well as the helix of another RING domain (PDB entry: 1RMD) with a Cys residue at this position and the following loop involved in the interaction.

proteins families, very similar structures have low-sequence homology, and some clear sequence rules determine the local conformation (i.e., the SH3 family). Whereas in other cases, significant changes in conformation take place with the same homology level (i.e., the UBD family). In cases of low-sequence identity, different methodologies are used to model backbone movements: molecular dynamics (64), fragment libraries (65–67), or the possibility of building chimeras using different parts of proteins belonging to the same family (G. Fernandez, in preparation); but modeling requires a careful examination of the protein family to which the target belongs.

Loops frequently create a homology modeling problem. Insertion or deletion of one or more residues, or simply changing a critical residue in a loop, can result in large conformational changes that are difficult to predict. If possible, loops that are not involved in the interaction between proteins, or between protein and ligand, should not be modeled (16). When it is necessary to model them, the best solution is to build chimeras using loop information from other members of the same family. If this is not possible, then automatic methods like MODELLER (55) or WHAT IF (56) can be used, but the results should be regarded with caution.

3.1. Quality of Template Structures

When homology modeling is for sequences with a high level of sequence identity, success is dependent on the quality of the template structure, the quality of the rotamer library, the search engine, and the force field employed. Usually, a higher the resolution of the structure results in a better prediction. Ideally, structures below 2.2 Å resolution should be used. However, in some concrete cases where information from other members of the family is available, a lower resolution can be considered. One way to assess the quality of the structure of a complex is to calculate the interaction energy after refinement with a protein design algorithm. If the ΔΔG-binding energy

is not negative and is equal to or higher than the experimental data, the quality of the template is not good. In some cases, the problem could be local (a particular residue with bad, i.e., combinations of angles resulting in van der Waals clashes, φ and ψ angles, for example), whereas in others, it could be more global.

Further validation could be done by performing an *in silico* alanine-scanning mutagenesis with the original NMR structures and comparing the results with experimental mutagenesis data. If predicting a single alanine mutant is successful, this is an indication that the X-ray structure is of high quality (the opposite is not necessarily true because local conformational changes in response to the mutation could result in mispredictions). In the case of Ras effector interactions, alanine mutants were predicted with a correlation of 0.7/0.8 for the complexes of Ras in complex with RalGDS-RA and Raf-RBD (68).

A typical problem encountered when doing side chain replacement at the protein, or DNA, level is that the position of the CB (carbon atom ß) in the case of amino acids, or of the N9 or N1 atoms in the case of DNA, is not always constant with respect to the backbone. This means that placement of standard side chains on a protein or DNA backbone could result in large deviations at the tip of the residue or base. This problem can be solved by always superimposing structures on the CB atoms (or N9 or N1 in the case of DNA). However, this assumes that the deviations of these atoms with respect to standard side chains are due to structure determination artifacts and are not real. If the deviations are real and if the atoms have some small capacity for displacement, then the correct side chain conformation, which fits with a small displacement of the CB atom, may be missed.

In any case before doing any homology modeling, it is always recommended that the structural template be repaired. This means flipping the Asn, Gln, and His side chains back and forth 180° to see if they are correctly placed because in X-ray structures it is not possible to distinguish the CO and NH2 groups of Asn and Gln or the CD2 and ND1 atoms of His. Side chains should be moved slightly to eliminate van der Waals clashes, and if residues on the surface are part of the crystal contacts, other rotamers should be explored.

NMR structures should not be used unless they have been refined to very high resolution using dipolar couplings and other techniques (69, 70). If NMR structures are used, the recommended strategy is to repair them using the same protein design algorithm that will be used later for homology modeling and to select the structure with lowest energy.

MSA: multiple sequence alignment

3.2. Multiple Sequence Alignments: A Tree of Methods

One of the crucial steps for homology modeling that still requires further development is the alignment of the target sequence to model with the structural template. In particular, when sequence identity between the target sequence and template sequence is low (<30%), the accuracy of the alignment and the produced model are very weak (71).

Multiple sequence alignment (MSA) tools have been under development for decades (72, 74, 75). Wallace and colleagues (73) reported that, in 2005 alone, 20 new publications, describing new methods, were found in the literature. They showed that the different methodologies can be combined in a meta-alignment method (dubbed M-Coffee) by careful selection of independent methods (76).

Perhaps one of the most significant developments for the use of MSAs in homology modeling has been the incorporation of structural information [see the descriptions of 3DCoffee (77), Staccato (78), and SAlign (79)]. Given the robustness of protein structure against amino acid substitution, homologous proteins are more likely to retain structural similarity than sequence identity over time (80). For 3DCoffee, it was shown that

the inclusion of protein structures among distantly related sequences increased MSA 4% per added structure (77). The use of sequence alignment tools for the purpose of homology modeling of low target to template sequence identity was recently benchmarked (81). Dalton & Jackson (81) showed that when using the same modeling tool (modeller 8v2), 3D(T)Coffee + MODELLER perform best. Improvements were achieved for two- to four-template modeling, giving a significant advantage over one-template modeling.

Packages that bundle various programs and speed the creation of different modeling protocols are useful. One of the few examples of currently available packages is Biskit (82), a Python library for structural bioinformatics. Even using the best automatic sequence alignment methods, it is, nevertheless, necessary at the end to inspect the alignment manually using structural information to ensure that it is correct (83).

3.3. Limits of Structural Coverage

In 2005, Chandonia & Brenner (27) estimated that the available protein structures would cover from 35% to 51% of the proteomes of model organisms. This estimate was calculated as the fraction of proteins containing at least one domain belonging to a family that has one of its members resolved structurally. As we discussed above, to create an accurate model using homology modeling, stricter requirements are needed. Some domain families are more easily modeled than others. SH3 domains, for example, retain structural similarity for low-sequence identity. For this reason in a best-case scenario, SH3 domains serve as a good model to test how much structural coverage is achievable with current homology modeling methods. We set up an automatic modeling pipeline using Biskit (82) as a wrapper for 3DCoffee (77) and MODELLER (57) to test the extensibility of current structural coverage for modeling SH3 domains. We used single- and two-template modeling of seven SH3 domains of known structure to benchmark the procedure (see **Figure 4*a***). Using two templates and the best current practices, reasonable SH3 models were obtained when the average sequence identity of the templates was above 30%. At this threshold, 98% of the models had an average CA-RMSD (carbon α-root mean square deviation) of less than 1.5 Å, and 65% of the models had an average CA-RMSD below 1.0 Å. For single-template modeling, a higher identity threshold is required. Using single templates of 40% or better sequence identity, 88% of the models produced had a CA-RMSD below 1.5 Å, and 60% had a CA-RMSD below 1.0 Å. As expected, the deviations from the known structure were not uniformly distributed in the whole structure but were more likely to occur in the loop regions (see **Figure 4*b***). Therefore, the larger the number of structures of a particular complex and of its protein constituents, the larger the probability of success (19, 61). We tested the structural coverage for human SH3 domains using stringent cutoffs and the 79 nonredundant (at a 95% identity cutoff) SH3 domains of known X-ray structure that were available in the Protein Data Bank (PDB). Out of a total of 453 human SH3 domains, 64 (or 14%) have been solved, and 97 (or 22%) have at least 60% identity to known structures and should be possible to model with very high accuracy. Another 140 domains (31%) have at least 40% average identity to two of these templates, which also assures the likelihood of a very accurate model. Current structural coverage and automatic homology modeling methods allow for determination of close to 67% of all human SH3 domains (see **Figure 5**). This is, however, a very favorable case, which likely represents an upper bound of the current possibilities for protein modeling. For the majority of domains, structural coverage is scarce, and therefore, there are sequences that cannot be modeled reliably.

4. LEVELS OF PREDICTION

In this section, we discuss the advances in the use of emperical and energy potencials to

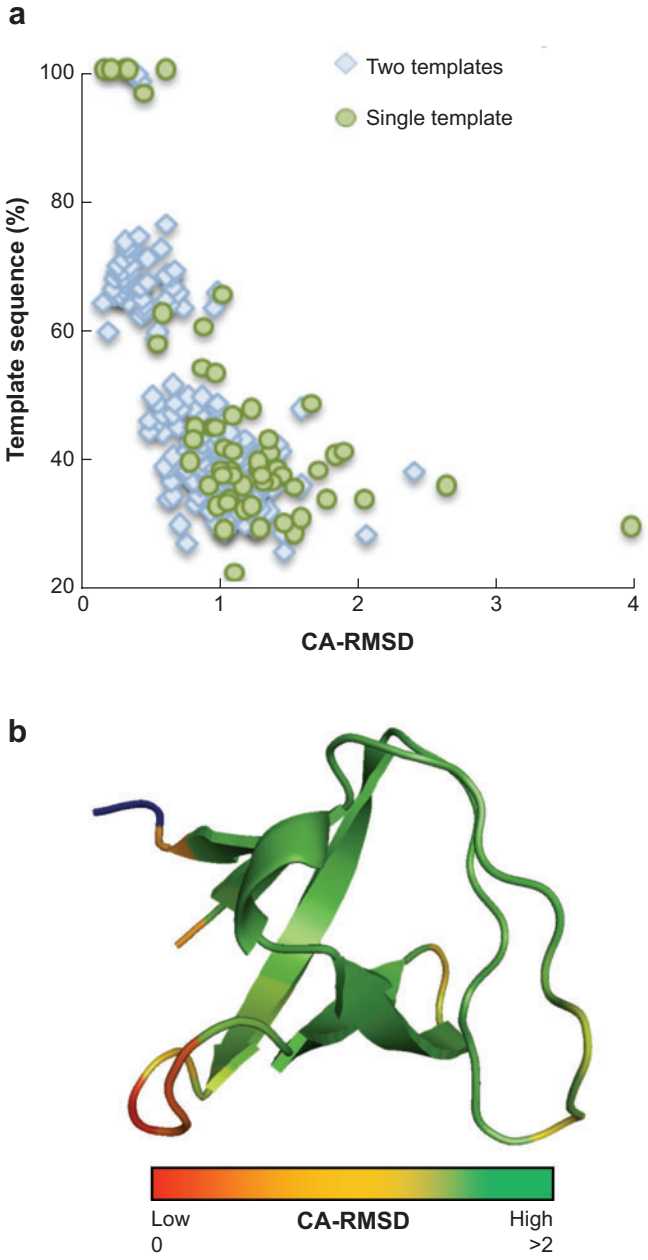


Figure 4

Automatic modeling of SH3 domains. (*a*) Seven SH3 domains of known structure were modeled by the current best practices for homology modeling [Dalton & Jackson (81)] using either single-template (*circles*) or two-template modeling (*diamonds*). For these seven domains, most of the models produced were of good quality, even for sequence identities close to 30%. (*b*) For each domain, the average CA-RMSD per residue can be plotted back onto the known structure to determine what regions are harder to model. For all domains, difficulties in modeling were generally located in the loop regions, primarily the N-scr loop, the distal loop, and the tip of the RT loop. In panel b, low = 0, and high = >2. Abbreviations: CA-RMSD, carbon α-root mean square deviation; RMSD, root mean square deviation.

predict protein interactions. We mention the analysis of domain-peptide interactions separately to emphasize some of the difficulties associated with the prediction of this type of interaction.

4.1. Protein-Protein Interactions

Assuming that similar sequences have a similar fold and that domains with a similar fold interact through the same surface (17), much progress has been made in predicting protein-protein interactions on the basis of structural information (17) and homology modeling (19, 61). In **Figure 6**, we show different template structures for use depending on the actual modeling problem. On the basis of the original NMR structure, template structures used in homology modeling can be generated. The modification of the original template structure depends on the application and on the sequence similarity of the domain to be modeled with the template.

- If the sequences are very similar, and the loops between secondary structure elements have identical lengths, the complete X-ray structure is used as a template structure (**Figure 6*a***) (68). The predictions are very accurate because they take into consideration regions at the edge of the interface, which contribute a small amount to the overall binding energy and thus are included.
- If sequences have different loop lengths than those in the original X-ray structure and the loops are not, or are minimally, involved in the interaction, 3D

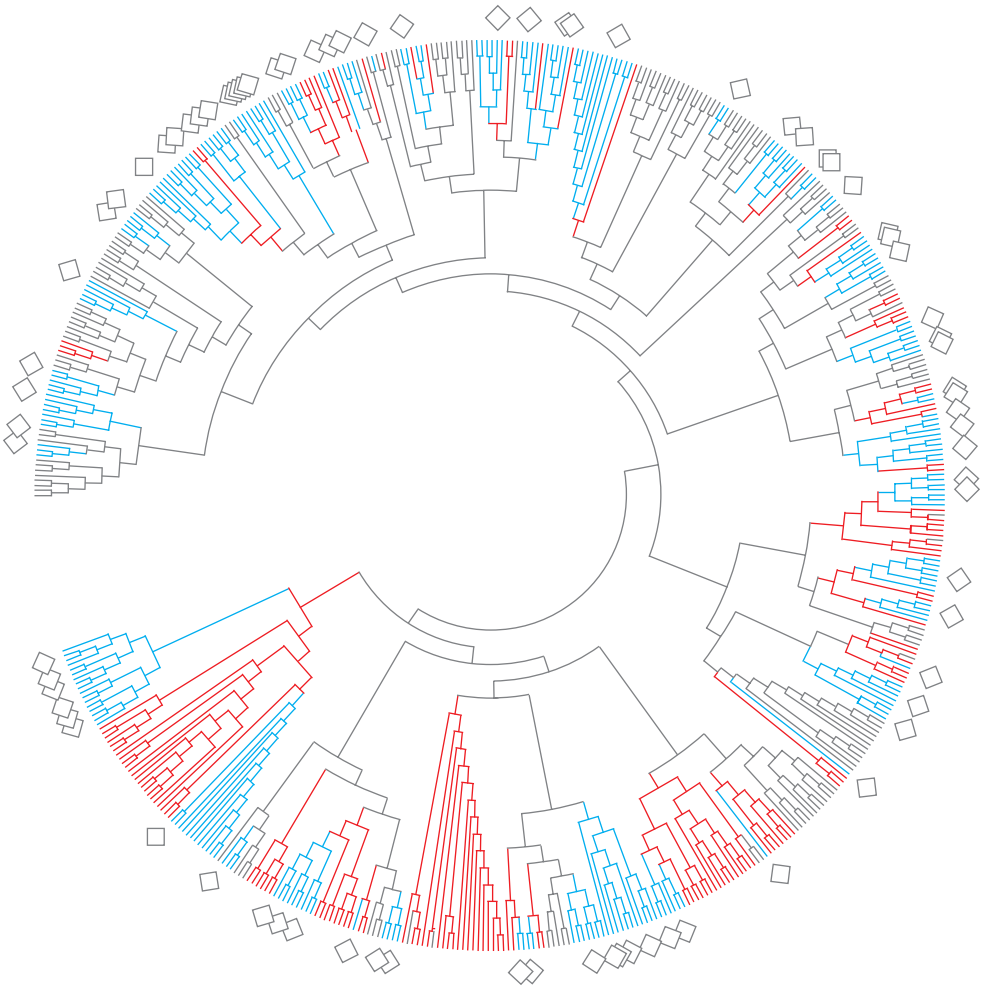


Figure 5

Modeling human SH3 domains. We built a phylogenetic tree for 453 human SH3 domains from the SMART database (**http://smart.embl.de**) and 79 nonredundant SH3 domains of known X-ray structure from the Protein Data Bank (**http://www.pdb.org/pdb/home/home.do**). Those domains that either have a known structure or that should be possible to model using available structures are indicated (*diamonds*).

structures are modified by deleting the structural parts not involved in the interaction (**Figure 6*b***) (19, 61).

- If sequences have different loops lengths and the loops contribute significantly to binding, the loops are generated using the WHAT IF library (56), or any other homology modeling tool, i.e., MODELLER (57), which contains loops of different lengths from a database of X-ray loop fragments (**Figure 6*c***). This was successful in the

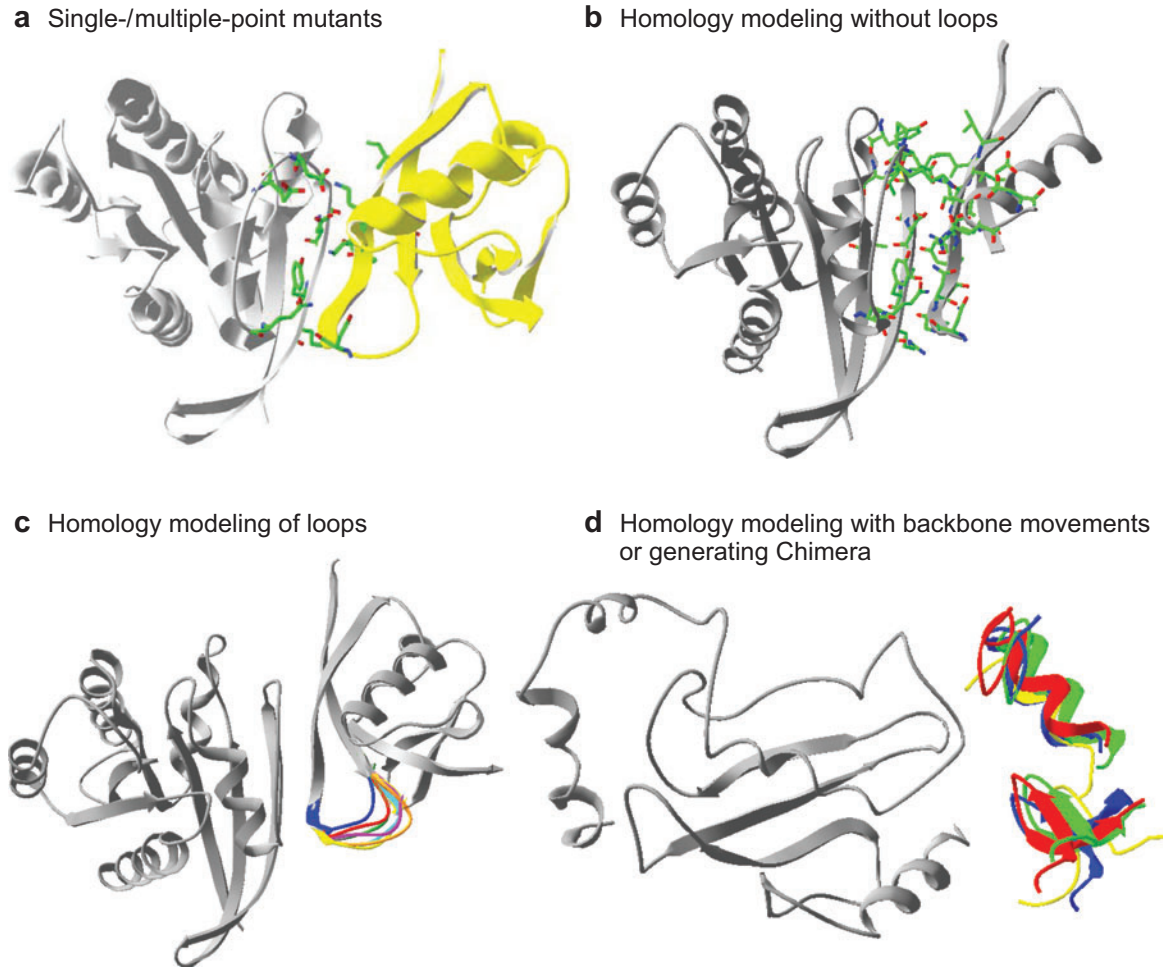


Figure 6

Template structures for homology modeling. (*a*) Complete 3D complex structures can be used if the sequences to be modeled have a similar loop lengths. (*b*) To model sequences with different loops lengths, which are not involved in the interaction, 3D complex structures are modified by deleting loops and secondary structural elements not involved in the interaction. (*c*) To model sequences with different loop lengths when the loops are involved in the interaction, loop template structures are generated using the WHAT IF library (see text). (*d*) If the sequences modeled are expected to have significant backbone changes, new chimera template structures are generated by superimposing the 3D complex structures with a 3D structure of a single domain of a similar sequence. Another possibility is to generate new template structures with a changed backbone using molecular dynamic simulations.

prediction of the Ras-effector interaction (61).

- If the structure of one of the partners to be modeled into the complex is known and differs from the template, or if the sequence analysis suggests important conformational changes at the interaction surface, then it is necessary to generate new templates (**Figure 6*d***). There are two ways of doing this. One can generate complex chimeras by superimposing on a 3D complex structure the known 3D structure of a single domain. Alternatively, if there is clear

TNF: tumor necrosis factor

evidence that particular sequences in a protein family determine certain specific local changes, a domain chimera is built by using different parts from proteins belonging to the same family (domain chimeras). This was successful for the prediction of SH3 domains with their target peptides (84).

Is there an ideal domain interaction type for the prediction of interactions on the basis of structural information? Ras proteins mediate their binding to effector domains using ß-sheet interactions and loops. The structural flexibility in these interfaces is low because of the backbone hydrogen bond (H-bond) interactions. In fact, the main structural changes occur in the helix α1 in the Ras-binding domain; these changes can have significantly different conformations depending on the complex. Thus, the accuracy of the predictions is very good, although only six different template structures have been used (61). The tumor necrosis factor (TNF) family ligand-receptor binding seems to be ideal for structure-based design because TNF family members adopt a very similar tertiary structure, and the diverse features of surface residues mediate the specificity between different TNF family members (85). Interactions through the formation of a β-sheet seem easy to predict, probably because the backbone-backbone H-bonds restrict the possibilities of slightly different conformations in various complexes. Surfaces that involve little or no main chain H-bonds are more problematic for the simple reason that side chain mutations could slightly change the interaction geometry between molecules A and B, and therefore the number of templates required for successful prediction increases enormously. The flexibility in protein-peptide complexes is much higher, and thus prediction methods for protein-peptide interactions are discussed separately, see below.

Large-scale predictions were done for the Ras-effector system (61). Because in this approach the backbone is kept constant and just the side chains are replaced, the problem of possible backbone movements is taken into consideration by using several template structures to model a sequence and to calculate the binding energy. Protein design algorithms, such as FoldX (86–88), usually predict changes in affinity accurately, but the interaction energies for different protein complexes are not easily transferable into affinity constants, as derived from quantitative binding experiments. However, when a set of 20 Ras proteins was modeled in complex with Ras and Rap, a low correlation between the experimental and calculated affinities was found (19). Therefore, a successful approach to analyze the interaction energies and to decide if two proteins interact on the basis of their interaction energies was to select the model with the lowest interaction energy generated using different template structures (19, 61). However, if the total energy (protein stability) of a sequence modeled with a particular template structure is very bad (has high energy) as a result of high van der Waals clashes, the model should be discarded because the result indicates that the sequence and the template structure are not compatible, and thus the result of homology modeling will not be reliable, although interaction energies might be favorable. After the best model, generated using different template structures, has been selected, one needs criteria for energy thresholds to decide whether this sequence will bind. To do so, energy thresholds are defined by calibrating energy sum values using experimental information (61). Using this threshold information, new binding and nonbinding domains can be successfully predicted, and the rest of the prediction is in the "twilight zone" (61). The prediction accuracy of predicting binding and nonbinding domains is very high (∼0.8) using this method (61).

Interface modeling, although accurate, is a time-consuming process, especially when predictions are done on a genome-wide scale. Consequently, a new method, the prediction

of protein-protein interactions on the basis of energy matrices, was successfully developed and applied to the prediction of Ras-effector interactions (93). In this method, position-specific energy matrices are generated, and using different Ras-effector X-ray template structures, all amino acids in the Ras-binding domain are sequentially mutated to all other amino acid residues, and the effect on binding energy is calculated. Then the precalculated matrices are used to score the binding of any Ras or effector sequences. Using these matrices, the sequences of putative Ras-binding domains are scanned quickly to calculate an energy sum value. By calibrating energy sum values with quantitative experimental binding data, thresholds are defined, and nonbinding domains are excluded quickly. Sequences that have energy sum values above this threshold are considered potential binding domains and can be further analyzed using homology interface modeling.

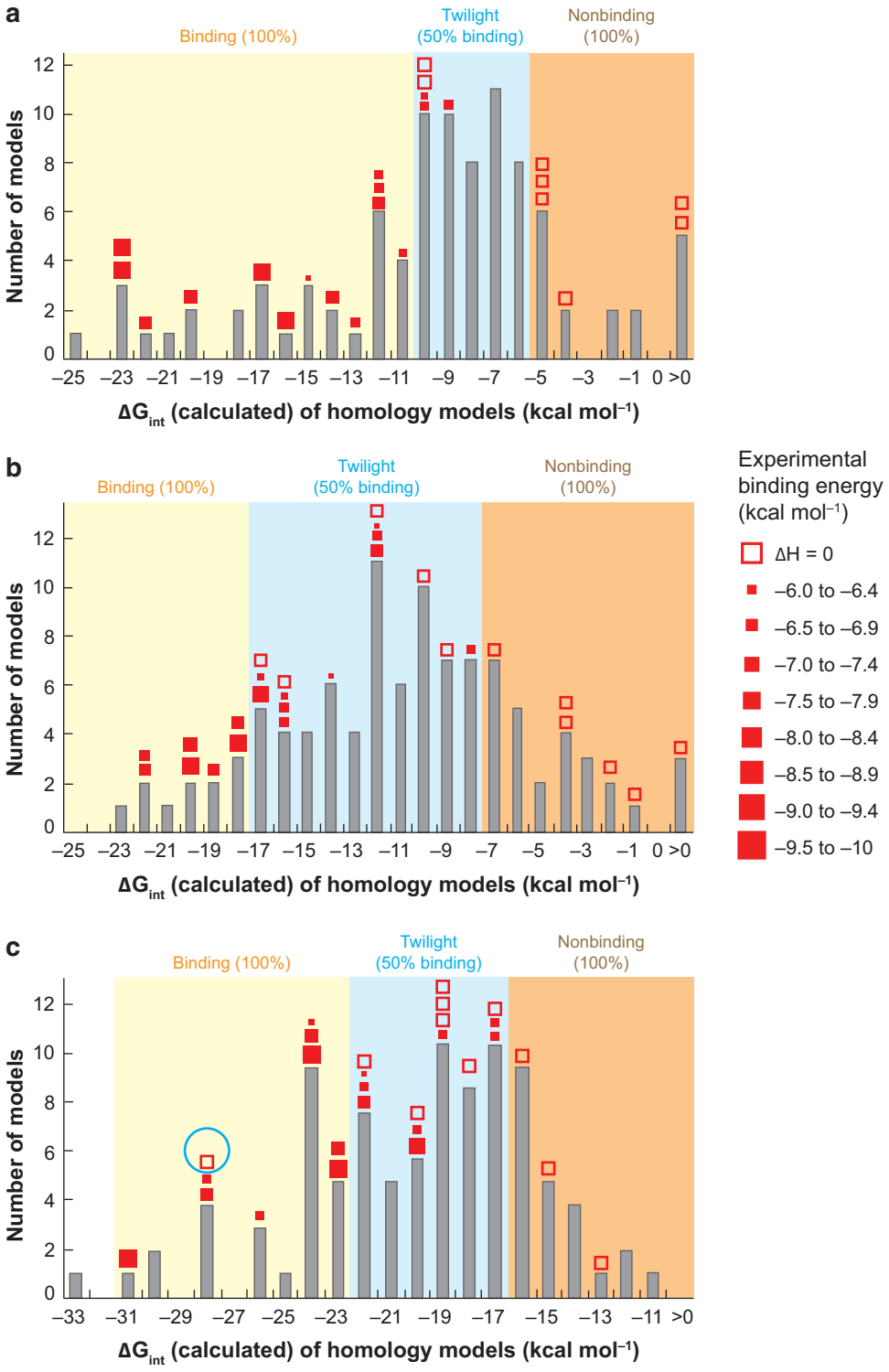
In **Figure 7**, the prediction success is compared—as judged from a range of energy thresholds—with a set of 50 UBD domains in complex with Ras and Rap using three different methods: homology modeling with loops, homology modeling without loops, and energy matrices. The highest discrimination power is obtained using homology models that explicitly take loop modeling into consideration (**Figure 7*a***). After adding experimental binding information, there are two clear thresholds for predicting binding and nonbinding domains, leaving only a very narrow twilight zone of 5 kcal/mol. Discrimination power decreases somewhat when modeling Ras-effector interactions if no loops are taken into consideration (**Figure 7*b***). Here, we also observe clear thresholds for binding and nonbinding domains; however, the area of twilight is little larger (10 kcal/mol). The results for predictions with position energy matrices using the test set of Ras/Rap-effector interactions (**Figure 7*c***) show a very good discriminative power for nonbinding domains. However, the false positives can occur at low-energy values, which makes defining the threshold for predicting binding domains difficult.

4.2. Domain-Peptide Interactions

The studies mentioned above focused on trying to predict domain-domain interactions and usually did not consider domain-peptide interactions. That is, the interaction is assumed to be between two folded structures with relatively large binding interfaces instead of interactions with linear peptides as is the case of several protein domains (SH3, SH2, WW, 14-3-3 domains). The binding specificity of peptide-binding domains can be characterized experimentally by many different approaches with the use of oriented peptide libraries. These peptide libraries can be presented either in phage display (94), spotted on cellulose membrane (95, 96), or allowed to interact with protein arrays containing the binding domains (97). These experimental approaches are time consuming and do not necessarily provide with an understanding of the structural properties that define the specificity of each domain. Developments in structure-based predictions of domain-peptide interactions are poised to circumvent these deficits. We discuss these developments, giving emphasis to the difficulties associated with the larger conformational variability of peptide ligands and the lack of specificity of peptide-binding domains.

Protein domain: an element of overall protein structure that is self-stabilizing and often folds independently of the rest of the protein chain

Some of the first attempts to predict domain-peptide interactions with the aid of some structural information were developed in the mid- to late 1990s (105). In this study, a simple energy potential function, specially developed for particular domain-peptide interactions, was fitted using structures of known complexes and experimentally determined binding energies. This approach requires significant knowledge and is not easily applicable to other cases. In the early 2000s, different groups built upon this idea to determine the binding specificity of domain families by analysis of complexes and known binding peptides for SH3 domains (106),



protein kinase domains (107), and SH2 domains (108). The structures of several domains of the same family, in complex with a peptide, were analyzed together with information on known binding peptides. The authors, mentioned above, determined the residues that are important for binding from the structural analysis and then used different methods to correlate the binding residues in the domain with the target residues in the peptide. With these rules, it is possible to use any protein domain of the family under study (SH3, kinase, and SH2), with similar enough sequences, and match key residues with predicted binding specificity. Predicting the binding specificity of new domains in this way can be very accurate but is only applicable for domains that are significantly similar to the ones already studied (109). By this means, current knowledge of binding specificity is extended to domains of identical sequences, but several complex structures and experimentally determined peptide-binding data must be available for the domain family chosen. These procedures are also related to the approach used by Aloy & Russell (15), described above, for Ras/RBD interactions.

More recently, our group and the lab of Wei Wang have used general purpose energy force fields to predict the binding specificity of peptide-binding domains (84, 110, 111; N. Sanchez, under review; G. Fernandez-Ballester, in preparation). These methods have the great advantage that no domain-specific information is required aside from a model of the complex for the domain under investigation. This might not be a difficult requirement because, in some cases, this model can be obtained from homology modeling, shown by Gregorio and colleagues, manuscript in preparation, and G. Fernandez-Ballester, manuscript in preparation. As described above for Ras-effector interactions, it is possible to use these force fields to do *in silico* mutagenesis of the target ligand. One advantage of predicting protein-peptide interactions is that the target ligand usually adopts an extended conformation, and another is that ligand positions can be assumed to be mostly independent. From this computational analysis, a position-specific scoring matrix (PSSM) is created containing the information on the preferred residues at each position in the ligand. Given that the target is a peptide ligand and not a folded domain, using these ligand matrices requires no domain information or modeling for the target protein analyzed.

The main difficulty in applying these approaches to protein-peptide predictions is the large conformation variability possible for the ligand. SH3 ligands, for example, are usually referred to as class I or class II ligands depending on their orientation (112). Soon after the initial peptide studies, a third group for less common types of ligands was suggested (113). Our own recent analysis of SH3 domain complexes retrieved 29 SH3 ligands of different conformations (G. Fernandez-Ballester, in preparation). This large variability makes it harder to evaluate the binding of putative targets because different amino acids might be accommodated with movements of the peptide backbone. As detailed above, it is possible to model each different conformation separately and analyze the energy of all conformations, but the problem becomes harder to tackle. Future improvements in domain-peptide predictions from structure have much to gain from programs capable of predicting backbone movements upon mutation. This can be achieved by incorporating molecular dynamics algorithms (110) or fragment libraries (65).

Figure 7

Prediction accuracy for a test set of 50 ubiquitin-like domains modeled in complex with Ras and the Ras-like protein, Rap1. (*a*) Homology modeling using loops. (*b*) Homology modeling without loops. (*c*) Prediction using position energy matrices. Abbreviation: ΔG_{int}, interaction energy.

Another disadvantage for target prediction of domain-peptide prediction, when compared to protein-protein predictions, is the smaller specificity defined by the binding surface. A domain-peptide interaction is usually determined by a small number of residues in the target peptide (114), which makes them difficult to identify in the proteome. However, many of these target sites are thought to occur in unstructured regions of the proteome (115). Interactions with unstructured protein segments might be functionally important for different reasons (e.g., decoupling of specificity and affinity, clustering of multiple binding sites, faster rates of association and dissociation), and many computational strategies have been developed to predict these sites of disorder [see the review by Radivojac et al. (116) and Table 1]. Therefore, improvements in binding accuracy can be achieved by restricting the search of putative binding sites to protein segments predicted to be intrinsically disordered (117).

4.3. Nonatomic Detail Prediction

In principle, predictions, which are based on sequence alone, are not be possible because only one residue change in the interface could lead to complete loss of binding affinity. However, aside from homology modeling, which quantitatively describes the interaction in atomic-level detail, nonatomic-detail methods have been successfully used to predict the interaction of proteins (15, 98). These methods, such as Interaction Prediction through Tertiary Structure (InterPreTS) (**http://www.russell.embl.de/cgi-bin/interprets2**) (99) and MULTIPROSPECTOR (100), use empirical pair potentials, which describe how well a homologous pair of sequences fit into a complex structure. Both were successfully applied to predict the specificities of large domain families, e.g., the complex between fibroblast growth factors and their receptors (15).

5. STRUCTURAL INFORMATION AS A TOOL TO ANALYZE PROTEIN INTERACTION NETWORKS

As discussed above, structural information in combination with protein design algorithms can be used to predict new protein interactions. However, there are other possible uses of structural information in the analysis of interaction networks. The relative interaction energy, which is based on complex protein structures, can be predicted in a fast and accurate way with existing protein design algorithms. Successful predictions of binding affinities for wild-type and mutant complexes have been carried out using the protein design algorithms FoldX (86–88) and Rosetta (89–91). Examples are the prediction of Ras-effector interactions (19, 61, 68) and alanine mutations at the interface of a member of the TNF-related apoptosis-inducing ligand (TRAIL) family in complex with its receptor, DR5 (92). Prediction of ubiquitin with ubiquitin-interacting motifs also gives the qualitatively correct trend (83). Structural information in combination with bioinformatic tools (118, 119) and/or protein design algorithms (21) can be used to predict the functional effect of human single-nucleotide polymorphisms.

As mentioned above protein design algorithms usually predict changes in affinity accurately. However, prediction of the absolute value for a binding constant or kinetic parameter is more difficult. Thus, the interaction energies for different protein complexes are not easily transferable into affinity constants, as derived from quantitative binding experiments. However, for proteins of the same family, and for complexes between proteins that do not involve conformational changes upon binding, rough estimates can be obtained for the binding constant (19, 20) and k_{on} values (19, 51) provided that a calibration with experimental data has been done first.

Finally, we should mention the possibility of using structural information to partition protein interaction networks into structurally compatible subnetworks. It is quite obvious now that many proteins do have more partners than surface available for interaction and therefore that some interactions must be mutually exclusive. In principle, if enough information is available regarding the complexes made by one protein or domain with several other proteins or domains, one could easily decide which interactions are simultaneously possible by doing a superimposition and looking for excluded surfaces (104). In the absence of structural information, it is possible to look for complexes involving homologue domains and, assuming that proteins of the same families will in general interact the same way, do the same exclusion exercise (A. Campagna, C. Kiel, & L. Serrano, manuscript in preparation). With more structures and complexes being deposited every day, the likelihood of dissecting protein interaction networks into functional subnetworks is becoming more pausible and will be one of the main contribuitions of structural biology to systems biology.

6. MINING FOR BIOLOGICAL CONTEXT

The structure-based methods described above predict binding affinities in the same way one would obtain them from in vitro assays. Knowledge of binding affinities alone is not sufficient to determine if proteins interact inside the cell. What determines binding in a living organism is a conjugation of factors, such as expression levels, localization, complex formation (i.e., scaffolding), posttranslation modifications, splicing forms, and association with small compounds. We need what one could call the biological context information or a way to predict it to make inroads into in vivo binding predictions. Some recent developments started to tackle this problem by using integrative probabilistic approaches that try to weight different information sources and by combining them with protein-binding specificity information (61, 120). We explore, below, some of the most commonly used methods to predict protein-protein interactions and ways to combine these into a single scoring function.

Protein fusion: in protein evolution analysis, a protein containing two sequence blocks that are homologous to proteins in other species

6.1. Sequence-Based Methods

Some early attempts to predict protein-protein association came from the early comparative genomics analyses. For example, it was observed that conserved proximity of two open reading frames correlates with an increase in likelihood of protein interaction between the coded proteins (121).

In similar fashion, it was shown that phylogenetic association of protein pairs also signals functional linkage. That is, if two proteins are always present or absent together (not necessarily in close vicinity in the genome) in many different species, then the two proteins are likely part of the same complex/pathway (122). Another sequence-based method relies on the determination of protein fusion events. In 1999, Marcotte and colleagues (123) showed that if two proteins are sometimes seen in some species fused into one contiguous protein, then these are very likely related in function and, therefore, also more likely to interact. These methods have the advantage that they only require the simple analysis of a large number of genomes, but they do not directly predict protein interactions but instead functional association. Another disadvantage is that these methods are not very effective in eukaryotic species, given that they have a more complex genome structure and fewer of these genomes are available to study.

A different method to predict protein interaction from sequence information was developed by Pazos and colleagues (124, 125). Analyzing alignments of interacting proteins, the authors showed that correlated mutations, between the two proteins, are identifiable signals for protein-protein interaction. This method not only identifies directly

protein-protein association, but it also determines the protein-binding regions.

Since these works by Pazos et al., many other metrics of correlated mutations have been proposed [reviewed by Halperin et al. (126)], although most have been directed at studies of intramolecular interactions instead of prediction of intermolecular contacts. Halperin and colleagues (126) tested the capacity of these different measures to predict protein-protein contacts and found most to be weak predictors.

A related method to the study of correlated mutations is the analysis of the coevolution of entire sequences known as "mirror tree" methods, pioneered by Goh and colleagues in 2000 (127) and further improved later, see the review by Shoemaker & Panchenko (128). In this approach, proteins from interacting protein families are aligned, and the correlation of the obtained pairwise distances is used as a signal of coevolution.

In both the correlated mutations and mirror tree methods, it is assumed that the evolution of the protein sequences of the interacting families will be mostly driven by the correlated changes in the binding epitopes. Given that most of variance in protein evolutionary rates can be explained by the level of protein expression (129), it is possible that the predictive power of these approaches might be limited (130).

6.2. Domain Interaction Propensity

From the analysis of the different interaction networks experimentally determined, one can extract the likelihood that any two given protein domains might interact. This knowledge of domain interaction propensities can then be used to direct the prediction of new protein-protein interactions (101–103). These approaches were further improved by advances in the statistical analysis (131) with the current top performing algorithm being the InSite program (132). A great advantage of these approaches is that they not only predict protein-protein interactions but also directly inform on the putative domains involved.

6.3. Graph Theory Methods

One early approach in the study of large cellular interaction maps was to simplify the information into a graph form. Each component was symbolized as a node, and each interaction as an edge in the graph (133). This graph abstraction is, in many respects, too large a simplification of what we already know of proteins and cellular functions, but it allowed for a vast number of studies regarding interaction networks (133–136). One interesting observation coming from these graph studies is that protein interactions, or edges, can be predicted just by looking at the graphs of current incomplete interactomes (137, 138). Some recent advances in protein-protein interaction prediction have come from trying to do comparative graph analysis between the interactomes of different species (139, 140). This type of approach, dubbed comparative interactomics (141), can be thought of as the analogy to comparative genomics.

6.4. Integration of Different Methods

The proliferation of experimental and computational methods to study protein-protein interactions has prompted comparative studies of the different approaches (142). These analyses have shown that the different experimental and computational methods are not overlapping, and all suffer from low accuracy and low coverage when benchmarked against a trusted dataset. Also, it was demonstrated that interactions that were observed in more than one of the analyses were more likely to be a true interactions. From these first efforts to compare the different methods came then the idea that more reliable information can be obtained from the combination of different experimental and computational observations.

One year after the comparative analysis from von Mering and colleagues (142), two different groups provided the first examples of a statistical combination of the various approaches with a Bayesian framework (14, 143). This strategy was used in 2005 to integrate different information sources to predict human protein interactions with considerable success (10,000 predictions with a 20% false-positive rate) (144). Given the interdependences that occur between the different datasets used, there is a limit to the benefit obtained from this type of integration (145).

These probabilistic weighting schemes can be used to increase the reliability of interactions determined by the prediction of binding specificities. Recently this approach was used to improve the prediction of human kinase targets (120). In this work, known phosphorylation targets were linked to the most likely kinase by combining information on binding specificity with prediction of functional interactions taken from the STRING database (146). In our lab, we have used a naive Bayes network to weigh different interaction predictors with a set of in vivo interactions retrieved from the Human Protein Reference Database (147). This combined predictor was then used to attribute confidence scores to putative Ras/RBDs (61).

We believe that effort should be made by the community to establish prediction servers that are constantly updated to integrate meaningful datasets and computation methods. These servers in combination with putative binding specificity should allow us to predict biological pathways with great accuracy.

7. LIMITATIONS

Structure-based prediction of protein-protein interactions is becoming a mature field. The recent advances in protein design algorithms and MSAs, in combination with other bioinformatic approaches, allow genome-wide prediction for particular domain interactions. However, there are some serious limitations that preclude its broad general use. The more important one is the lack of enough structural templates in many cases, both at the level of isolated domains and also at the level of protein complexes. The quality of the predictions is strictly related to the abundance of different templates that explore the conformational space available for a particular complex. Molecular dynamics and chimeras can partly compensate for it, but they require some expertise and are time consuming. It is important that, in the future, structural genomics project scientists focus on filling the structural gaps in families. Although it does not make too much sense to determine the structure of a domain if there is already a PDB record of a closely related protein, filling the structural gaps is critical to solve structures of sequences with low homology.

Even with the best methodology (experimental and computational), a detailed bioinformatic or experimental biological validation should be done. Any computationally or experimentally determined K_d is meaningless in the absence of a biological context. Thus, a micromolar K_d could be relevant if the concentration is high or if one of the two proteins is localized. Similarly, a nanomolar K_d is meaningless if the two proteins never see each other.

SUMMARY POINTS

1. Methods to predict protein interactions on the basis of structural information can complement large-scale experimental protein interaction networks.
2. All prediction methods are based on the finding that structurally similar domains usually interact in a similar way.

3. However, it is not the fold per se which determines whether two proteins can interact, but rather certain amino acid residues on the surface of this conserved fold determine the affinity and specificity between the domains.
4. Homology modeling is used to generate complex models for sequences that are homologous to the template (X-ray) structure, and interaction energies can be calculated using protein design force fields.
5. Energy matrices contain the binding energy contributions for all 20 amino acid residues for every position in the interface. They can be used to quickly scan sequences and evaluate whether they are potential binding domains and need to be modeled, and they can exclude sequences that cannot bind.
6. Structural information can also be used to partition interaction networks into structurally compatible subnetworks.
7. Crude estimates of binding (equilibrium and kinetic) parameters as well as functional consequences of point mutations can be obtained from structural analysis.
8. Bioinformatics data integration helps validate predictions and provides further knowledge about expression levels and localization.
9. The main limitations of structure-based predictions are the quality of template structures, slightly different binding modes, and backbone changes of the participating domains.

FUTURE ISSUES

1. In the area of homology/interface modeling, the following important issues need to be solved: backbone moves, loop modeling, automatization of template selection (which sequences are compatible with a specific template structure?).
2. Structure-based sequence alignments need to be optimized.
3. Current structural proteomics approaches will solve the structure of more protein complexes and thus increase the number of template structures (~10,000 domain-domain interaction types are predicted, so far only ~2,000 are known).
4. The existing methods to predict domains based on the amino acid sequence need to be optimized by taking secondary structure prediction and structural information into account.
5. Simulations have to be performed to integrate protein interaction network changes (time and localization).
6. Prediction of thermodynamic and kinetic parameters (K_d, k_{on}, k_{off}) on the basis of structure need to be further developed.
7. The roles of different K_ds/specificity for signal transduction and signaling flow have to be studied.
8. There is a need for bioinformatic servers, which are capable of integrating predicted binding affinity with other experimental resources to determine biologically relevant protein interactions.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENT

We thank the EU for finantial support (Interaction Proteome, grant number LSHG-CT-2003-505520).

LITERATURE CITED

1. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. 2005. *Nature* 437:1173–78
2. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. 2000. *Nature* 403:623–27
3. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, et al. 2001. *Proc. Natl. Acad. Sci. USA* 98:4569–74
4. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. 2005. *Cell* 122:957–68
5. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. 2002. *Nature* 415:141–47
6. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al. 2002. *Nature* 415:180–83
7. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, et al. 2006. *Nature* 440:631–36
8. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, et al. 2001. *Science* 293:2101–5
9. Kung LA, Snyder M. 2006. *Nat. Rev. Mol. Cell Biol.* 7:617–22
10. Cusick ME, Klitgord N, Vidal M, Hill DE. 2005. *Hum. Mol. Genet.* 14:R171–81
11. Fields S. 2005. *FEBS J.* 272:5391–99
12. Berggard T, Linse S, James P. 2007. *Proteomics* 7:2833–42
13. Devos D, Russell RB. 2007. *Curr. Opin. Struct. Biol.* 17:370–77
14. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, et al. 2003. *Science* 302:449–53
15. Aloy P, Russell RB. 2002. *Proc. Natl. Acad. Sci. USA* 99:5896–901
16. Kiel C, Serrano L. 2008. In *Structural Proteomics*, ed. J Sussman, I Silman, pp. 29–46. Singapore: World Sci. In press
17. Aloy P, Russell RB. 2006. *Nat. Rev. Mol. Cell Biol.* 3:188–97
18. Lensink MF, Mendez R, Wodak SJ. 2007. *Proteins* 69:704–18
19. Kiel C, Wohlgemuth S, Rousseau F, Schymkowitz J, Ferkinghoff-Borg J, et al. 2005. *J. Mol. Biol.* 348:759–75
20. Kortemme T, Joachimiak LA, Bullock AN, Schuler AD, Stoddard BL, Baker D. 2004. *Nat. Struct. Mol. Biol.* 11:371–79
21. Pey AL, Stricher F, Serrano L, Martinez A. 2007. *Am. J. Hum. Genet.* 81:1006–24
22. Zhang Y. 2007. *Proteins* 69:108–17
23. Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A. 2007. *Proteins* 69:3–9
24. Kim WK, Henschel A, Winter C, Schroeder M. 2006. *PLoS Comp. Biol.* 2:1151–64
25. Aloy P, Russell RB. 2004. *Nat. Biotechnol.* 22:1317–21
26. Chothia C. 1992. *Nature* 357:543–44
27. Chandonia J-M, Brenner SE. 2005. *Proteins* 58:166–79
28. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. *J. Mol. Biol.* 215:403–10
29. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. 1997. *Nucleic Acids Res.* 25:3389–402

30. Linding R, Russell RB, Neduva V, Gibson TJ. 2003. *Nucleic Acids Res.* 31:3701–8
31. Dosztányi Z, Csizmók V, Tompa P, Simon I. 2005. *J. Mol. Biol.* 347:827–39
32. Dosztányi Z, Csizmók V, Tompa P, Simon I. 2005. *Bioinformatics* 21:3433–34
33. Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK, et al. 2005. *J. Bioinform. Comput. Biol.* 3:35–60
34. Schultz J, Milpetz F, Bork P, Ponting CP. 1998. *Proc. Natl. Acad. Sci. USA* 95:5857–64
35. Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, et al. 2006. *Nucleic Acids Res.* 34:D257–60
36. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. 2004. *Nucleic Acids Res.* 32:D138–41
37. Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, et al. 2006. *Nucleic Acids Res.* 34:D227–30
38. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, et al. 2005. *Nucleic Acids Res.* 33:D192–96
39. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, et al. 1997. *Structure* 5:1093–108
40. Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, et al. 2005. *Nucleic Acids Res.* 33:D247–51
41. Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. *J. Mol. Biol.* 247:536–40
42. Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, et al. 2004. *Nucleic Acids Res.* 32:D226–29
43. Finn RD, Marshall M, Bateman A. 2005. *Bioinformatics* 21:410–12
44. Stein A, Russell RB, Aloy P. 2005. *Nucleic Acids Res.* 33:D413–17
45. Winter C, Henschel A, Kim WK, Schroeder M. 2006. *Nucleic Acids Res.* 34:D310–14
46. Ogmen U, Keskin O, Aytuna AS, Nussinov R, Gursoy A. 2005. *Nucleic Acids Res.* 33:W331–36
47. Jefferson ER, Walsh TP, Roberts TJ, Barton GJ. 2007. *Nucleic Acids Res.* 35:D580–89
48. Davis FP, Sali A. 2005. *Bioinformatics* 21:1901–7
49. Wohlgemuth S, Kiel C, Kraemer A, Serrano L, Wittinghofer F, et al. 2005. *J. Mol. Biol.* 348:741–58
50. Kiel C, Serrano L. 2007. *Curr. Chem. Biol.* 1:215–25
51. Selzer T, Albeck S, Schreiber G. 2000. *Nat. Struct. Biol.* 7:537–41
52. Kiel C, Selzer T, Shaul Y, Schreiber G, Herrmann C. 2004. *Proc. Nat. Acad. Sci. USA* 101:9223–28
53. Goldsmith-Fischmann S, Honig B. 2003. *Protein Sci.* 12:1813–21
54. Ginalski K. 2006. *Curr. Opin. Struct. Biol.* 16:172–77
55. Schwede T, Kopp J, Guex N, Peitsch MC. 2003. *Nucleic Acids Res.* 31:3381–85
56. Vriend G. 1990. *J. Mol. Graph.* 8:52–56
57. Sali A, Blundell TL. 1993. *J. Mol. Biol.* 234:779–815
58. Mendes J, Baptista AM, Carrondo MA, Soares CM. 1999. *Proteins* 37:530–43
59. Allen BD, Mayo SL. 2006. *J. Comput. Chem.* 27:1071–75
60. Santana R, Larranaga P, Lozano JA. 2007. *Artif. Intell. Med.* 39:49–63
61. Kiel C, Foglierini M, Kuemmerer N, Beltrao N, Serrano L. 2007. *J. Mol. Biol.* 370:1020–32
62. Zheng N, Wang P, Jeffrey PD, Pavletich NP. 2000. *Cell* 102:533–39
63. Chung SY, Subbiah S. 1996. *Structure* 4:1123–27
64. Holyoake J, Caulfeild V, Baldwin SA, Sansom MS. 2006. *Biophys. J.* 91:L84–86
65. Simons KT, Bonneaum R, Ruczinski I, Baker D. 1999. *Proteins* 3(Suppl.):171–76
66. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, et al. 2005. *Proteins* 61(Suppl.):128–34

67. Yarov-Yarovoy V, Schonbrun J, Baker D. 2006. *Proteins* 62:1010–25
68. Kiel C, Serrano L, Herrmann C. 2004. *J. Mol. Biol.* 340:1039–58
69. Wedemeyer WJ, Baker D. 2003. *Proteins* 53:262–72
70. Jensen PR, Axelsen JB, Lerche MH, Poulsen FM. 2004. *J. Biomol. NMR* 28:31–41
71. Kopp J, Schwede T. 2004. *Pharmacogenomics* 5:405–16
72. Lipman DJ, Altschul SF, Kececioglu JD. 1989. *Proc. Natl. Acad. Sci. USA* 86:4412–15
73. Wallace IM, Blackshields G, Higgins DG. 2005. *Curr. Opin. Struct. Biol.* 15:261–66
74. Edgar RC. 2004. *Nucleic Acids Res.* 32:1792–97
75. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. 2005. *Genome Res.* 15:330–40
76. Wallace IM, O'Sullivan O, Higgins DG, Notredame C. 2006. *Nucleic Acids Res.* 34:1692–99
77. O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C. 2004. *J. Mol. Biol.* 340:385–95
78. Shatsky M, Dror O, Schneidman-Duhovny D, Nussinov R, Wolfson HJ. 2004. *Nucleic Acids Res.* 32:W503–7
79. Madhusudhan MS, Marti-Renom MA, Sanchez R, Sali A. 2006. *Protein Eng. Des. Sel.* 19:129–33
80. Chothia C, Lesk AM. 1986. *EMBO J.* 5:823–26
81. Dalton JA, Jackson RM. 2007. *Bioinformatics* 23:1901–8
82. Grunberg R, Nilges M, Leckner J. 2007. *Bioinformatics* 23:769–70
83. Kiel C, Serrano L. 2006. *J. Mol. Biol.* 355:821–44
84. Musi V, Birdsall B, Fernandez-Ballester G, Guerrini R, Salvatori S, et al. 2006. *Protein Sci.* 4:795–807
85. Zhang G. 2004. *Curr. Opin. Struct. Biol.* 14:1–7
86. Guerois R, Nielsen JE, Serrano L. 2002. *J. Mol. Biol.* 320:369–87
87. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, et al. 2005. *Nucleic Acids Res.* 33:W382–88
88. Schymkowitz JWH, Rousseau F, Martins IC, Ferkinghoff-Borg J, Stricher F, et al. 2005. *Proc. Natl. Acad. Sci. USA* 102:10147–52
89. Simons KT, Kooperberg C, Huang E, Baker D. 1997. *J. Mol. Biol.* 268:209–25
90. Simons KT, Ruczinski I, Kooperberg C, Fox B, Bystroff C, et al. 1999. *Proteins* 34:82–95
91. Kortemme T, Baker D. 2002. *Proc. Natl. Acad. Sci. USA* 99:14116–21
92. Van der Sloot AM, Tur V, Szegezdi E, Mullally MM, Cool RH, et al. 2006. *Proc. Natl. Acad. Sci. USA* 103:8634–39
93. Kiel C, Serrano L. 2007. *Bioinformatics* 23:2226–30
94. Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, et al. 2002. *Science* 295:321–24
95. Landgraf C, Panni S, Montecchi-Palazzi L, Castagnoli L, Schneider-Mergener J, et al. 2004. *PLoS Biol.* 2:E14
96. Huang H, Li L, Wu C, Schibli D, Colwill K, et al. 2007. *Mol. Cell. Proteomics*. In press
97. Jones RB, Gordus A, Krall JA, MacBeath G. 2006. *Nature* 439:168–74
98. Lu L, Arakaki AK, Lu H, Skolnick J. 2003. *Genome Res.* 13:1146–54
99. Aloy P, Russell RB. 2003. *Bioinformatics* 19:161–62
100. Lu L, Lu H, Skolnick J. 2002. *Proteins* 49:350–64
101. Sprinzak E, Margalit H. 2001. *J. Mol. Biol.* 311:681–92
102. Wojcik J, Schachter V. 2001. *Bioinformatics* 17(Suppl. 1):296–305
103. Deng M, Mehta S, Sun F, Chen T. 2002. *Genome Res.* 12:1540–48
104. Kim PM, Lu LJ, Xia Y, Gerstein MB. 2006. *Science* 314:1938–41
105. Zhou Y, Abagyan R. 1998. *Fold Des.* 3:513–22

106. Brannetti B, Via A, Cestra G, Cesareni G, Helmer-Citterich M. 2000. *J. Mol. Biol.* 298:313–28
107. Brinkworth RI, Breinl RA, Kobe B. 2003. *Proc. Natl. Acad. Sci. USA* 100:74–79
108. Sheinerman FB, Al-Lazikani B, Honig B. 2003. *J. Mol. Biol.* 334:823–41
109. Ferraro E, Via A, Ausiello G, Helmer-Citterich M. 2006. *Bioinformatics* 22:2333–39
110. Hou T, Chen K, McLaughlin WA, Lu B, Wang W. 2006. *PLoS Comput. Biol.* 2:e1
111. McLaughlin WA, Hou T, Wang W. 2006. *J. Mol. Biol.* 357:1322–34
112. Zarrinpar A, Bhattacharyya RP, Lim WA. 2003. *Sci. STKE* 2003:RE8
113. Cesareni G, Panni S, Nardelli G, Castagnoli L. 2002. *FEBS Lett.* 513:38–44
114. Pawson T, Raina M, Nash P. 2002. *FEBS Lett.* 513:2–10
115. Fuxreiter M, Tompa P, Simon I. 2007. *Bioinformatics* 23:950–56
116. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK. 2007. *Biophys. J.* 92:1439–56
117. Beltrao P, Serrano L. 2005. *PLoS Comput. Biol.* 3:e26
118. Money S. 2006. *Bioinformatics* 6:44–56
119. Torkamani A, Schork NJ. 2007. *Bioinformatics* 23:2918–25
120. Linding R, Jensen LJ, Ostheimer GJ, van Vugt MA, Jorgensen C, et al. 2007. *Cell* 129:1415–26
121. Dandekar T, Snel B, Huynen M, Bork P. 1998. *Trends Biochem. Sci.* 23:324–28
122. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. *Proc. Natl. Acad. Sci. USA* 96:4285–88
123. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, et al. 1999. *Science* 285:751–53
124. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A. 1997. *J. Mol. Biol.* 271:511–23
125. Pazos F, Valencia A. 2002. *Proteins* 47:219–27
126. Halperin I, Wolfson H, Nussinov R. 2006. *Proteins* 63:832–45
127. Goh CS, Bogan AA, Joachimiak M, Walther D, Cohen FE. 2000. *J. Mol. Biol.* 299:283–93
128. Shoemaker BA, Panchenko AR. 2007. *PLoS Comput. Biol.* 3:e43
129. Drummond DA, Raval A, Wilke CO. 2006. *Mol. Biol. Evol.* 23:327–37
130. Hakes L, Lovell SC, Oliver SG, Robertson DL. 2007. *Proc. Natl. Acad. Sci. USA* 104:7999–8004
131. Riley R, Lee C, Sabatti C, Eisenberg D. 2005. *Genome Biol.* 6:R89
132. Wang H, Segal E, Ben-Hur A, Li QR, Vidal M, Koller D. 2007. *Genome Biol.* 8:R192
133. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL. 2000. *Nature* 407:651–64
134. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001. *Nature* 411:41–42
135. Wagner A. 2001. *Mol. Biol. Evol.* 18:1283–92
136. Wuchty S, Oltvai ZN, Barabasi AL. 2003. *Nat. Genet.* 35:176–79
137. Goldberg DS, Roth FP. 2003. *Proc. Natl. Acad. Sci. USA* 100:4372–76
138. King AD, Przulj N, Jurisica I. 2004. *Bioinformatics* 20:3013–20
139. Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, et al. 2003. *Proc. Natl. Acad. Sci. USA* 100:11394–99
140. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, et al. 2005. *Proc. Natl. Acad. Sci. USA* 102:1974–79
141. Cesareni G, Ceol A, Gavrila C, Palazzi LM, Persico M, et al. 2005. *FEBS Lett.* 579:1828–33
142. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, et al. 2002. *Nature* 417:399–403
143. Troyanskaya OG, Dolinski K, Owen AB, Altman RB, Botstein D. 2003. *Proc. Natl. Acad. Sci. USA* 100:8348–53
144. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, et al. 2005. *Nat. Biotechnol.* 23:951–59

145. Lu LJ, Xia Y, Paccanaro A, Yu H, Gerstein M. 2005. *Genome Res.* 15:945–53
146. Snel B, Lehmann G, Bork P, Huynen MA. 2000. *Nucleic Acids Res.* 28:3442–44
147. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, et al. 2003. *Genome Res.* 13:2363–71
148. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. 2006. *BMC Bioinformatics* 7:208
149. Ishida T, Kinoshita K. 2007. *Nucleic Acids Res.* 35:(Web Server Issue) W460-4
150. Hirose S, Shimizu K, Kanai S, Kuroda Y, Noguchi T. 2007. *Bioinformatics* 23:2046–53