Computational methods for the prediction of protein interactions

Alfonso Valencia* and Florencio Pazos†

Establishing protein interaction networks is crucial for understanding cellular operations. Detailed knowledge of the 'interactome', the full network of protein–protein interactions, in model cellular systems should provide new insights into the structure and properties of these systems. Parallel to the first massive application of experimental techniques to the determination of protein interaction networks and protein complexes, the first computational methods, based on sequence and genomic information, have emerged.

Addresses
*Protein Design Group, National Center for Biotechnology,
CNB-CSIC, Cantoblanco, 28049 Madrid, Spain;
e-mail: valencia@cnb.uam.es
†ALMA Bioinformática, Centro Empresarial Euronova, Ronda de Poniente,
4 Tres Cantos, 28760 Madrid, Spain;
e-mail: pazos@almabioinfo.com

Current Opinion in Structural Biology 2002, **12**:368–373

0959-440X/02/$ – see front matter

Abbreviations

i2h	*in silico* two-hybrid
MSA	multiple sequence alignment

Introduction

The molecular bases of cellular operations are largely sustained by different types of interactions among proteins. However, only recently has it become possible to combine the traditional study of proteins as isolated entities with the analysis of large protein interaction networks. This is of particular interest as many of the properties of complex systems seem to be more closely determined by their interactions than by the characteristics of their individual components. Furthermore, recent findings — including the fact that *Caenorhabditis elegans* and humans have a similar number of genes, and the marked similarity in the sequences of human and mouse genes — suggest that species differences cannot be accounted for by the individual properties of their component genes, but rather by the relationships between them. The study of protein interaction networks is important not only from a theoretical stance but also in terms of potential practical applications, because it might enable new drugs to be developed that can specifically interrupt or modulate protein interactions, instead of simply targeting a given protein's complete set of functions.

An impressive set of experimental techniques has been developed for the systematic analysis of protein interactions, including yeast two-hybrid-based methods [1], identification by mass spectrometry of isolated protein complexes [2•,3], protein chips [4•] and hybrid approaches [5]. The aim of all of these techniques is to obtain the full protein interaction network for simple cellular systems, such as yeast [2•,3,6••,7••] and *Helicobacter pylori* [8•]. And, although the limits of resolution of these approaches are open to discussion [9,10••], they do nevertheless promise much for the future.

In parallel, a number of computational methods have been developed for the prediction of protein interactions from genomic information [11,12•], extending into the prediction of the residues that participate in the interacting surfaces. Here, we describe the five computational techniques available for the prediction of interaction partners and examine their range of applicability. In addition, we analyze new trends in the determination of interacting surfaces on the basis of sequence information.

Computational methods for the prediction of interaction partners

Presence or absence of genes in related species

This method is based on the pattern of the presence or absence of a given gene in a set of genomes, that is, determining in which organisms the gene is present and in which it is not (*phylogenetic profiles* method; Figure 1a). Similarity of phylogenetic profiles might then be interpreted as being indicative of the functional need for corresponding proteins to be simultaneously present in order to perform a given function together. However, although this similarity may suggest a related functional role, a direct physical interaction between the proteins is not necessarily implied [13,14]. The main limitations of this approach lie in the fact that it can only be applied to complete genomes (as only then is it possible to rule out the absence of a given gene). Similarly, it cannot be used with the essential proteins that are common to most organisms.

Conservation of gene neighborhood

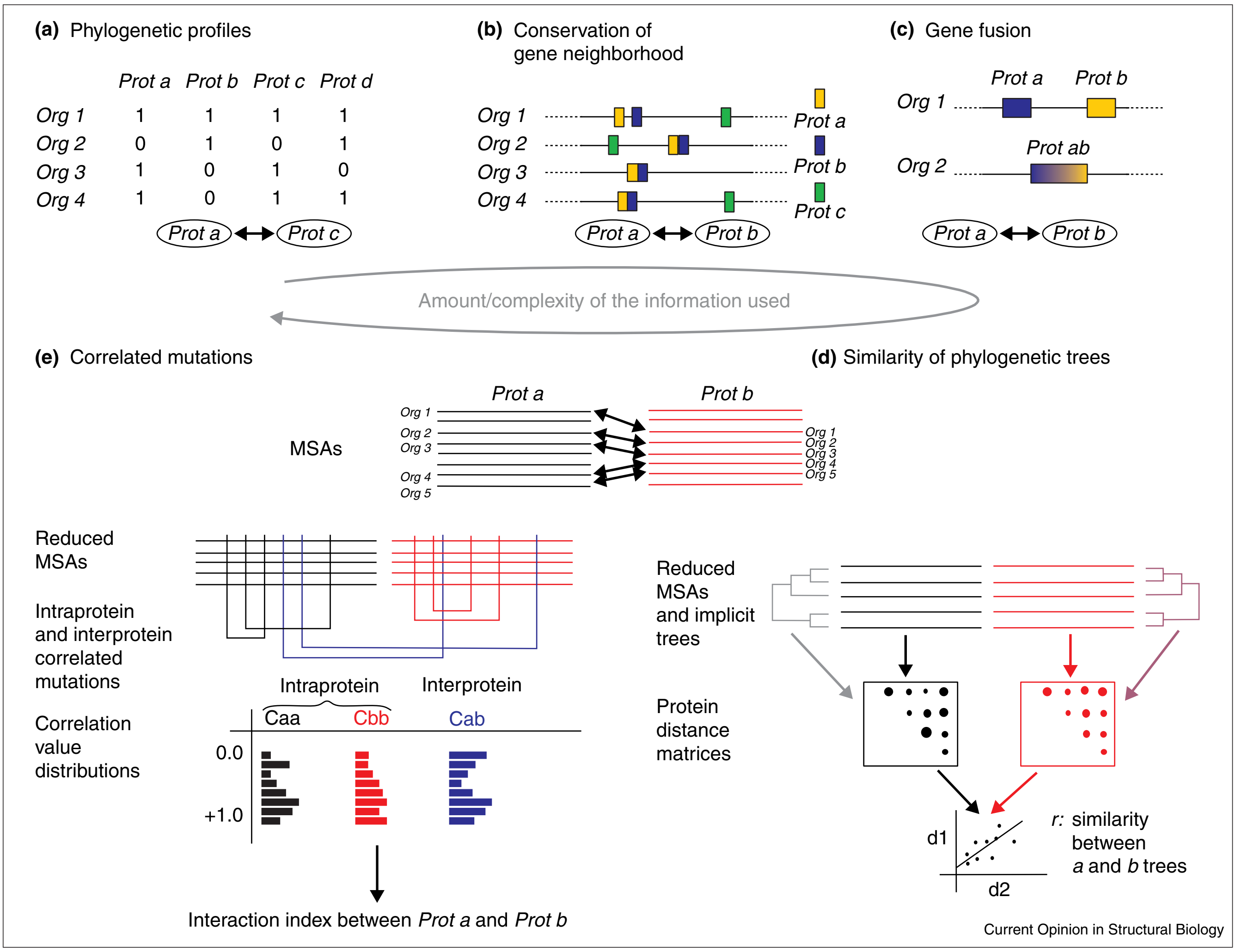
The organization of bacterial genomes into regions that tend to code for functionally related proteins, such as operons, is a well-known fact. This neighborhood relationship becomes even more relevant when it is conserved in different species [15]. The adjacency of genes in various bacterial genomes has been used to predict functional relationships between the corresponding proteins [16,17] (Figure 1b).

One of the main limitations of this method is that it is only directly applicable to bacteria, in which the genome order is a relevant property.

Gene fusion events

Interactions between proteins can be deduced from the presence in different genomes of the same protein domains, which either form part of a single polypeptide chain (multidomain protein) or act as independent proteins (single domains) (Figure 1c). Methods based on recursive

Figure 1


(a) Phylogenetic profiles
Prot a Prot b Prot c Prot d
Org 1 1 1 1 1
Org 2 0 1 0 1
Org 3 1 0 1 0
Org 4 1 0 1 1
Prot a ↔ Prot c
(b) Conservation of gene neighborhood
Org 1
Org 2
Org 3
Org 4
Prot a
Prot b
Prot c
Prot a ↔ Prot b
(c) Gene fusion
Prot a Prot b
Org 1
Prot ab
Org 2
Prot a ↔ Prot b
Amount/complexity of the information used
(e) Correlated mutations
(d) Similarity of phylogenetic trees
Prot a
Prot b
MSAs
Org 1
Org 2
Org 3
Org 4
Org 5
Reduced MSAs
Intraprotein and interprotein correlated mutations
Correlation value distributions
Intraprotein
Interprotein
Caa
Cbb
Cab
0.0
+1.0
Interaction index between Prot a and Prot b
Reduced MSAs and implicit trees
Protein distance matrices
d1
d2
r: similarity between a and b trees
Current Opinion in Structural Biology

Methods for predicting protein interaction partners from genomic and sequence information. The methods are presented according to the amount of information they include, ranging from simple patterns of gene presence in genomes to detailed sequence information (amino acids in each position) in protein families. **(a)** Phylogenetic profiles [13,14]. A profile is constructed for each protein (*Prot a–Prot d*), recording its presence (1) or absence (0) in a set of organisms (*Org 1–Org 4*). Pairs of proteins with identical (or similar) phylogenetic profiles are predicted to interact (*Prot a* and *Prot c* in this case). **(b)** Conservation of gene neighborhood [16,17]. Proteins whose genes are physically close in the genomes of various organisms are predicted to interact (*Prot a* and *Prot b*). **(c)** Gene fusion [18,19]. Two proteins of a given organism (*Prot a* and *Prot b* of *Org 1*) are predict to interact if they form part of a single protein in other organisms (*Org 2*). **(d)** Similarity of phylogenetic trees (*mirrortree*) [24•,25•]. To obtain a quantitative indicator of the interaction between two proteins (*Prot a* and *Prot b*), the MSAs of both proteins are reduced to the set of organisms common to the two proteins (*Org 1–Org 5*). Each of the reduced alignments is used to construct the corresponding intersequence distance matrix. These matrices are commonly used to construct the corresponding phylogenetic trees. Finally, the linear correlation between these distance matrices is calculated. High correlation values are interpreted as indicative of the similarity between phylogenetic trees and hence are taken as predicted interactions. **(e)** Correlated mutations (*i2h*) [31•]. The first step (reduction of the MSAs to a set of common organisms) is the same as that described for the *mirrortree* method (d). A correlation coefficient is calculated for every pair of residues. The pairs are divided into three sets: two for the intraprotein pairs (Caa and Cbb; pairs of positions within *Prot a* and within *Prot b*) and one for the interprotein pairs (Cab; one position from *Prot a* and one from *Prot b*). The distributions of correlation values are recorded for these three sets. The ‘interaction index’ is calculated by comparing the distribution of interprotein correlations with the two distributions of intraprotein correlations [31•].

sequence searches and multiple sequence alignments (MSAs) have been combined in order to detect such domain fusion events [18,19]. It has also been shown that fusion events are particularly common in metabolic proteins [20]. By definition, this approach is restricted to shared domains in distinct proteins, a phenomenon whose true extent is still unclear [21], especially in prokaryotic organisms.

Similarity of phylogenetic trees *(mirrortree)*

In a number of closely studied cases, it has been possible to show that interacting protein pairs co-evolve, for example, insulin and its receptors [22], and dockerins and cohexins [23]. In such cases, the corresponding phylogenetic trees of the interacting proteins show a greater degree of similarity (symmetry) than noninteracting proteins would be expected to show. For the two domains of phosphoglycerate kinase, Goh *et al.* [24•] quantified the similarity of their phylogenetic trees as the linear correlation between the distance matrices used to construct the trees. This approach (the *mirrortree* method) was extended [25•] to large sets of interacting proteins and protein domains, for which the value of the correlation between the distance matrices of pairs of proteins was found to be a good indicator of their probability of interaction (Figure 1d).

It seems that the process of co-evolution could lead, in the limit, to the simultaneous loss of both proteins in some organisms, an observation that forms the basis of the phylogenetic profiles method discussed above (Figure 1a). To this extent, the 'phylogenetic profiles' method might be considered a simplification of the *mirrortree* procedure, as the former does not take the structure of the trees (length and order of the branches) into account in the same way as the *mirrortree* method does by analyzing the information implicit in the protein sequence distance matrices.

The main limitation of the *mirrortree* method is the need to obtain good quality, complete MSAs for the two proteins. These alignments should include sequences from the same species for the two proteins under consideration (Figure 1d).

In silico two-hybrid method

The co-evolution of interacting proteins can be followed more closely by quantifying the degree of co-variation between pairs of residues from these proteins (correlated mutations). These positions may correspond to compensatory mutations that stabilize the mutations in one protein with changes in the other. Information about correlated mutations in single proteins has been used in particular to predict proximal pairs of residues [26,27], to discriminate structural models derived by threading [28] and to drive *ab initio* folding simulations [29].

For certain proteins, correlated mutations have been demonstrated to be able to select the correct structural arrangement of two proteins based on the accumulation of signals in the proximity of interacting surfaces [30]. This relationship between correlated residues and interacting surfaces has been extended to the prediction of interacting protein pairs based on the differential accumulation of correlated mutations between the interacting partners (interprotein correlated mutations) and within the individual proteins (intraprotein correlated mutations) [31•] (Figure 1e).

As in the case of the *mirrortree* method, the main limitation of the *in silico* two-hybrid (*i2h*) approach is the need for complete alignments with a good coverage of species common to the two proteins under study. This limitation arises as a direct consequence of the hypothesis of co-evolution, which naturally requires the simultaneous study of the corresponding protein pairs in each genome. On a more positive note, however, this method, based on the compensatory mutation of residues that are expected to lie physically close to each other, should provide a better prediction of physical interactions than the other methods (Figure 1a–d), which are based on general genomic information and tend to mix direct physical and indirect functional relationships.

Comparison of the computational methods

Unfortunately, a definitive evaluation of any of these methods cannot yet be undertaken, because the availability of collections of interacting proteins is still highly limited (as is an accurate understanding of those proteins that do not interact). Current efforts to develop databases of protein interactions [32•,33,34•] and to establish standards for the exchange of information between these databases (e.g. 'Intact' project, EC V Framework Program; http://www.ebi.ac.uk:80/msd/Temblor/Temblor1.html) will, however, play a key role in the evolution of prediction techniques.

Complementary to these efforts, various data-mining procedures are emerging for the automatic extraction of information about protein interactions from the vast amount of accumulated bibliographic information (for a review, see [35,36]). Although these systems face considerable technical challenges — including the absence of standard protein and gene names, and the complexity of the functional relationships between interacting proteins — they are already starting to provide evidence of their capabilities.

Before this large collection of well-documented interactions becomes available, current efforts at evaluation must be based on general functional characteristics, such as key words describing function, class of cellular function and so on, which provide very little information about protein interactions and are, in fact, more closely concerned with functional relationships. A further aspect that concerns us is the degree of coverage the prediction methods can provide of the possible set of interactions, though their actual extent is still unknown. Huynen *et al.* [12•] compared the three methods based on genomic information (Figure 1a–c). In their analysis, the method based on gene order (Figure 1b) could be applied to 37% of the *Mycoplasma genitalium* genes, whereas the phylogenetic profiles method (Figure 1a) and the method based on gene fusion (Figure 1c) could only be applied to 11% and 6%, respectively. The combination of the three methods yielded predictions for 50% of *M. genitalium* genes, with just a small degree of overlap in the techniques. With respect to the accuracy of this test set, the percentage of pairs predicted by the three methods that either present a physical interaction, belong to the same macromolecular complex, form part of the same pathway or are implicated in the same process are: 78% for 'gene fusion' (with no false positives), 80% for 'conservation of gene order' and 63% for 'phylogenetic profiles'. The percentages for physical interactions only are 56%, 30% and 23%, respectively.

Recently, the phylogenetic profiles, *mirrortree* and *i2h* methods (Figure 1a,d,e) were applied to the *Escherichia coli* genome (D De Juan, F Pazos, A Valencia, unpublished data). Starting with 38 fully sequenced genomes for building the orthologous tables and MSAs, and requiring a minimum of 14 common sequences for *mirrortree* and *i2h*, it was possible to apply the methods to a common set of more than 480 000 pairs of proteins, which covers 1318 genes (more than 30% of the 4289 *E. coli* genes). The preliminary results show that these three methods are independent and that there is a relationship between each one of their scores and the possibility of physical or functional interaction, measured as coincidence of Swiss-Prot keywords or copresence in metabolic pathways.

Prediction of the molecular basis of protein interactions

Activity to develop computational methods that can predict interactions between proteins of known three-dimensional structure (the docking problem, see [37•] for a recent review) has been intense. These docking methods, however, are only applicable to the small fraction of complexes for which the structures of the two interacting proteins are known. Interestingly, a set of new computational methods can now address the problem of the prediction of interacting surfaces in the absence of complete information about the corresponding structures of the binding proteins.

Initial approaches have been based on the observed properties of the statistical composition of interacting surfaces in terms of residue types (polarity, charge, etc.) and on the structure of the surfaces [38–40]. Two recently published methods [41•,42•] encode some of these characteristics in neural networks in order to predict binding regions in individual proteins of known structure. The accuracy of these methods is around 70% for the prediction of interactions at the residue level.

A second type of method addresses the prediction of interacting residues in the absence of structural information. The first reported application determines the distribution of positions that show family-dependent patterns of conservation in MSAs ('tree determinants' [43]). The relationships between these positions and the binding surfaces have been demonstrated for a number of systems [43–47], and the resulting predictions have been experimentally validated in at least two cases [48,49].

A promising alternative to that described above is the use of information about correlated mutations in order to highlight the interaction sites in binding proteins. In this case, it is possible to interpret the information used to predict interacting partners (*i2h* method; Figure 1e) in terms of the physical proximity between pairs of positions subject to evolutionary compensation on the surface of interacting proteins [30].

Conclusions

Cellular function can only be understood by considering the individual properties of cellular components (proteins, genes, etc.) in the context of their complex relationships. It is therefore unsurprising that the study of these interactions and complexes is establishing itself as the main task in the 'post-genomic' era [50].

By calling on the accumulation of genomic information, the first computational techniques for predicting protein interaction networks are emerging. The five methods described here are based on a wide range of ideas and, although they are yet to be perfected, they should prove to be more than competent complements to the various experimental approaches already developed for the determination of interactions, particularly given the shortcomings that also exist for these experimental methods [9,10••].

The combination of experimental and theoretical data could, for the first time, provide complete information about interaction networks, thereby allowing studies to be undertaken of the distribution and number of interactions, the presence of key nodes in the networks, tolerance to perturbations and differences in network organization from one organism to another. Indeed, initial analyses of these networks have revealed interesting new properties of biological interaction networks [51•,52•,53], which may have major practical consequences for the design of new drugs and constitute the foundations of the new 'system biology'.

Update

Aloy and Russell [54] have proposed a method for predicting the specificity of interactions in families of interacting proteins. Their approach is based on the structure of homologous complexes and relies on the concept of tree determinants described previously [43–49].

Sprinzak and Margalit [55] analyzed the distribution of well-characterized sequence domains in interacting protein pairs. This information is used to search for putative new interacting pairs with similar domain composition.

Fraser *et al.* [56] have quantified the relationship between evolutionary rates, fitness and sequence co-evolution in a large set of experimentally proposed yeast interaction networks. Their statistical approach shows that the more connected nodes in the interaction network evolve at lower rates, possibly because they are subject to a stronger pressure to co-evolve with their interaction partners. This study lays the evolutionary foundation for the methods described in [25•,31•].

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

• of special interest
•• of outstanding interest

1. Fields S, Song O: **A novel genetic system to detect protein-protein interactions.** *Nature* 1989, **340**:245-246.

2. Gavin AC, Bösche M, Krause R, Grandi P, Marcioch M, Bauer A,
• Schultz J, Rick JM, Michon AM, Cruciat CM *et al.*: **Functional organisation of the yeast proteome by systematic analysis of protein complexes.** *Nature* 2002, **415**:141-147.

One of the first massive experimental determinations of multiprotein complexes. Previous studies used yeast two-hybrid-based approaches [1] for the massive detection of binary (two-protein) interactions in complete proteomes. In this study, the authors tried to detect multiprotein complexes in the yeast proteome using tandem affinity purification followed by mass spectrometry analysis (TAP/MS).

3. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K *et al.*: **Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry.** *Nature* 2002, **415**:180-183.

4. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N,
• Jansen R, Bidlingmaier S, Houfek T *et al.*: **Global analysis of protein activities using proteome chips.** *Science* 2001, **293**:2101-2105.

The first protein chip with a complete eukaryotic proteome. The chip was used to scan for protein–protein and protein–cofactor interactions.

5. Tong AH, Drees B, Nardelli G, Bader GD, Brannetti B, Castagnoli L, Evangelista M, Ferracuti S, Nelson B, Paoluzi S *et al.*: **A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules.** *Science* 2002, **295**:321-324.

6. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K,
•• Kuhara S, Sakaki Y: **Toward a protein-protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins.** *Proc Natl Acad Sci USA* 2000, **97**:1143-1147.

See annotation to [7••].

7. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR,
•• Lockshon D, Narayan V, Srinivasan M, Pochart P *et al.*: **A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*.** *Nature* 2000, **403**:623-631.

Two groups [6••,7••] carried out the first massive determination of protein interactions in the complete yeast proteome, using different technical implementations of the yeast two-hybrid methodology [1].

8. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, Simon S,
• Lenzen G, Petel F, Wojcik J, Schächter V *et al.*: **The protein-protein interaction map of *Helicobacter pylori*.** *Nature* 2001, **409**:211-215.

The first massive determination of interactions in a bacterial genome based on an implementation of the yeast two-hybrid technique [1], which provides a score associated with each predicted interaction.

9. Lakey JH, Raggett EM: **Measuring protein-protein interactions.** *Curr Opin Struct Biol* 1998, **8**:119-123.

10. Legrain P, Wojcik J, Gauthier JM: **Protein-protein interaction maps:
•• a lead towards cellular functions.** *Trends Genet* 2001, **17**:346-352.

An interesting discussion of the limitations of methods for the massive determination of protein interactions, highlighting the low degree of overlap between the interactions determined by different methods.

11. Galperin MY, Koonin EV: **Who's your neighbor? New computational approaches for functional genomics.** *Nat Biotechol* 2000, **18**:609-613.

12. Huynen M, Snel B, Lathe W, Bork P: **Predicting protein function by
• genomic context: quantitative evaluation and qualitative inferences.** *Genome Res* 2000, **10**:1204-1210.

The first comparison of computational methods for the prediction of interaction partners using genomic information. The three methods compared are: 'phylogenetic profiles', based on comparing the patterns of the presence or absence of genes in genomes to predict interaction, originally described in [13]; 'conservation of gene order' in different species [16]; and 'gene fusion' events [18,19] detected in MSAs.

13. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles.** *Proc Natl Acad Sci USA* 1999, **96**:4285-4288.

14. Gaasterland T, Ragan MA: **Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes.** *Microb Comp Genomics* 1998, **3**:199-217.

15. Tamames J, Casari G, Ouzounis C, Valencia A: **Conserved clusters of functionally related genes in two bacterial genomes.** *J Mol Evol* 1997, **44**:66-73.

16. Dandekar T, Snel B, Huynen M, Bork P: **Conservation of gene order: a fingerprint of proteins that physically interact.** *Trends Biochem Sci* 1998, **23**:324-328.

17. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N: **Use of contiguity on the chromosome to predict functional coupling.** *In Silico Biol* 1999, **1**:93-108.

18. Marcotte EM, Pellegrini M, Ho-Leung N, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences.** *Science* 1999, **285**:751-753.

19. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events.** *Nature* 1999, **402**:86-90.

20. Tsoka S, Ouzounis CA: **Prediction of protein interactions: metabolic enzymes are frequently involved in gene fusion.** *Nat Genet* 2000, **26**:141-142.

21. Sprinzak E, Margalit H: **Correlated sequence-signatures as markers of protein-protein interactions.** *J Mol Biol* 2001, **311**:681-692.

22. Fryxell KJ: **The coevolution of gene family trees.** *Trends Genet* 1996, **12**:364-369.

23. Pages S, Belaich A, Belaich JP, Morag E, Lamed R, Shoham Y, Bayer EA: **Species-specificity of the cohesin-dockerin interaction between *Clostridium thermocellum* and *Clostridium cellulolyticum*: prediction of specificity determinants of the dockerin domain.** *Proteins* 1997, **29**:517-527.

24. Goh C-S, Bogan AA, Joachimiak M, Walther D, Cohen FE:
• **Co-evolution of proteins with their interaction partners.** *J Mol Biol* 2000, **299**:283-293.

See annotation to [25•].

25. Pazos F, Valencia A: **Similarity of phylogenetic trees as indicator of
• protein-protein interaction.** *Protein Eng* 2001, **14**:609-614.

The relationship between 'similarity of phylogenetic trees' and protein interaction is studied in two articles [24•,25•]. Both used a very similar quantification of the similarity between phylogenetic trees. In the first article [24•], this measurement was applied to the two domains of phosphoglycerate kinase, whereas in the second [25•], the method was applied to large sets of domains and proteins, including more than 67 000 pairs of *E. coli* proteins.

26. Göbel U, Sander C, Schneider R, Valencia A: **Correlated mutations and residue contacts in proteins.** *Proteins* 1994, **18**:309-317.

27. Olmea O, Valencia A: **Improving contact predictions by the combination of correlated mutations and other sources of sequence information.** *Fold Des* 1997, **2**:S25-S32.

28. Olmea O, Rost B, Valencia A: **Effective use of sequence correlation and conservation in fold recognition.** *J Mol Biol* 1999, **293**:1221-1239.

29. Ortiz A, Kolinski A, Rotkiewicz P, Ilkowski B, Skolnick J: ***Ab initio* folding of proteins using restraints derived from evolutionary information.** *Proteins* 1999, **S3**:177-185.

30. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A: **Correlated mutations contain information about protein-protein interaction.** *J Mol Biol* 1997, **271**:511-523.

31. Pazos F, Valencia A: ***In silico* two-hybrid system for the
• selection of physically interacting protein pairs.** *Proteins* 2002, **47**:219-227.

This study can be seen as an extension of previous observations on the relationship between correlated mutations and protein–protein interfaces [30]. In this article, correlated mutations were used for the prediction of pairs of interacting proteins. The method was applied to large sets of domains and proteins.

32. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D:
• **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.** *Nucleic Acids Res* 2002, **30**:303-305.

See annotation to [34•].

33. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW: **BIND—The Biomolecular Interaction Network Database.** *Nucleic Acids Res* 2001, **29**:242-245.

34. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G,
• Helmer-Citterich M, Cesareni G: **MINT: a Molecular INTeraction database.** *FEBS Lett* 2001, **25674**:1-6.

This paper, together with [32•], is one of the first attempts to construct annotated databases of protein interactions, which can be used, among other things, to train and test methods for predicting protein interactions.

35. Blaschke C, Valencia A: **The peculiarities of the molecular biology nomenclature thwart the progress in information extraction.** *IEEE Intel Sys* 2002, in press.

36. Blaschke C, Hirschman L, Valencia A: **Information extraction in molecular biology.** *Brief Bioinform* 2002, in press.

37. Smith GR, Sternberg MJE: **Prediction of protein-protein**
• **interactions by docking methods.** *Curr Opin Struct Biol* 2002, **12**:28-35.

A very recent review of methods for physical docking. This paper includes a comparative table with the available docking programs and their web addresses. An interesting discussion on the protein recognition mechanism is also included.

38. Jones S, Thornton JM: **Analysis of protein-protein interaction sites using surface patches.** *J Mol Biol* 1997, **272**:121-132.

39. Lo Conte L, Chothia C, Janin J: **The atomic structure of protein-protein recognition sites.** *J Mol Biol* 1999, **285**:2177-2198.

40. Valdar WS, Thornton JM: **Protein-protein interfaces: analysis of amino acid conservation in homodimers.** *Proteins* 2001, **42**:108-124.

41. Zhou H X, Shan Y: **Prediction of protein interaction sites from**
• **sequence profile and residue neighbor list.** *Proteins* 2001, **44**:336-343.

See annotation to [42•].

42. Fariselli P, Pazos F, Valencia A, Casadio R: **Prediction of**
• **protein-protein interaction sites in heterocomplexes with neural networks.** *Eur J Biochem* 2002, **269**:1356-1361.

These studies [41•,42•] used a similar methodology to predict regions of a protein implicated in interactions with others. Neural networks are fed with the sequence profile of the residues that form surface patches and are trained to predict whether these patches form part of interacting surfaces or not.

43. Casari G, Sander C, Valencia A: **A method to predict functional residues in proteins.** *Nat Struct Biol* 1995, **2**:171-178.

44. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families.** *J Mol Biol* 1996, **257**:342-358.

45. Andrade MA, Casari G, Sander C, Valencia A: **Classification of protein families and detection of the determinant residues with an improved self-organizing map.** *Biol Cybern* 1997, **76**:441-450.

46. Pereira-Leal JB, Seabra MC: **Evolution of the Rab family of small GTP-binding proteins.** *J Mol Biol* 2001, **313**:889-901.

47. Caffrey DR, O'Neill LA, Shields DC: **A method to predict residues conferring functional differences between related proteins: application to MAP kinase pathways.** *Protein Sci* 2000, **9**:655-670.

48. Stenmark H, Valencia A, Martinez O, Ulrich O, Goud B, Zerial M: **Distinct structural elements of rab5 define its functional specificity.** *EMBO J* 1994, **13**:575-583.

49. Bauer B, Mirey G, Vetter IR, Garcia-Ranea JA, Valencia A, Wittinghofer A, Camonis JH, Cool RH: **Effector recognition by the small GTP-binding proteins Ras and Ral.** *J Biol Chem* 1999, **274**:17763-17770.

50. Walhout AJ, Vidal M: **Protein interaction maps for model organisms.** *Nat Rev Mol Cell Biol* 2001, **2**:55-62.

51. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large scale**
• **organisation of metabolic networks.** *Nature* 2000, **407**:651-653.

See annotation to [52•].

52. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality**
• **in protein networks.** *Nature* 2001, **411**:41-42.

The authors of these studies [51•,52•] constructed protein interaction networks based on experimentally determined and predicted interactions (proteins forming part of the same metabolic network, for example) and examined the organization of such networks. They showed that, in these biological networks, the number of connections is not uniformly distributed, with a few proteins acting as key nodes and concentrating many connections.

53. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L: **Integrated genomic and proteomic analyses of a systematically perturbed metabolic network.** *Science* 2001, **292**:929-934.

54. Aloy P, Russell RB: **Interrogating protein interaction networks through structural biology.** *Proc Natl Acad Sci USA* 2002, **99**:5896-5901.

55. Sprinzak E, Margalit H: **Correlated sequence-signatures as markers of protein-protein interactions.** *J Mol Biol* 2001, **311**:681-692.

56. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW: **Evolutionary rate in the protein interaction network.** *Science* 2002, **296**:750-752.