A novel, comprehensive method to detect and predict protein–protein interactions applied to the study of vesicular trafficking

Christof Winter, Thorsten Baust, Bernard Hoflack and Michael Schroeder
Biotechnological Centre, Dresden University of Technology, 01307 Dresden, Germany

Abstract: *Motivation.* Computational methods to predict protein–protein interactions are of great need. They can help to formulate hypotheses, guide experimental research and serve as additional measures to assess the quality of data obtained in high-throughput interaction experiments. Here, we describe a fully automated three-step procedure to predict and confirm protein–protein interactions. By maximising the information from text mining of the biomedical literature, data from interaction databases, and from available protein structures, we aim at generating a comprehensive picture of known and novel potential interactions between a given set of proteins. *Results.* A recent proteomics assay to identify the protein machinery involved in vesicular trafficking between the biosynthetic and the endosomal compartments revealed 35 proteins that were found as part of membrane coats on liposomes. When applying our method to this data set, we are able to reconstruct most of the interactions known to the molecular biologist. In addition, we predict novel interactions, among these potential linkers of the AP-1 and the Arp2/3 complex to membrane-bound proteins as well as a potential GTPase–GTPase effector interaction. *Conclusions.* Our method allows for a comprehensive network reconstruction that can assist the molecular biologist. Predicted interactions are backed up by structural or experimental evidence and can be inferred at varying levels of confidence. Our method pinpoints existing key interactions and can facilitate the generation of hypotheses.

Keywords: Protein interaction, text mining, protein structure, interaction prediction, membrane traffic.

Introduction

Protein interactions. Protein–protein interactions are fundamental to almost all cellular processes. In addition, nearly every major process in a cell is believed to be carried out by assemblies of ten or more protein molecules [1]. Identification of putative binding partners of a protein is therefore a desirable ambition. It can contribute to understand how such complex molecular machines are organised and how their parts work together. While much effort has been put in large-scale experiments to identify protein–protein interactions in yeast, worm, fly, and human on a genome-wide extent [2–9], the false positive rates of such approaches are estimated to be as high as 50% [10, 11]. Moreover, the intersection of large-scale interaction data sets with those derived from the literature is surprisingly small [8]. By predicting and assessing protein interactions, computational

approaches can help in separating false positive from true positive ones. Sequence-based methods for the prediction of protein–protein interactions include gene context conservation [12], synthetic lethality [13], phylogenetic profiling [14, 15] or co-evolution of gene expression [16]. Structural approaches have focused on the study of protein complexes of known structure [17]. Various databases of binding sites and interfaces between proteins and their domains exist [18–20]. The modelling of interactions using structural templates of sufficient similarity was employed by Aloy and colleagues to successfully model the yeast exosome and some 100 yeast complexes [21, 22].

Additionally, much knowledge on interactions is stored in abstracts that are publicly available in literature databases such as PubMed. While the expert normally is aware of the relevant literature concerning his field, this knowledge remains hidden to the non-specialist who just encounters a couple of genes, for example as a result of a microarray experiment. Text mining can provide access to theoretically all protein interactions hidden in the biomedical literature [23].

We will apply the above methods and others to reconstruct networks relevant for vesicular trafficking. While several recent studies made use of homologous interactions in other species to assess and predict protein interactions [24–27], no study has so far used sequence similarity to literature interactions obtained from text mining to our knowledge.

Vesicle coats and adaptor proteins. Vesicles are small, membrane-enclosed containers that mediate transport between cellular compartments. The formation of vesicles and the selective incorporation of cargo molecules are both mediated by protein coats, which are recruited onto the cytosolic side of the forming vesicle. In the case of clathrin-coated vesicles, the cargo transmembrane proteins are linked via adaptor proteins (APs) with structural coats such as clathrin [28]. AP-1 mediates the transport of selected transmembrane proteins that cycle between the trans-Golgi network, endosomes and the plasma membrane. How APs interact with other molecular components of the complex coat machinery remains largely unclear.

Materials and Methods

As resources, we combine protein–protein interactions derived from text mining of the biomedical literature, available protein interaction databases, and structural domain interactions. Aim of our method is to obtain all known interactions between a given set of proteins, and to further predict new interactions that are likely to be present within this set. The overall approach is a procedure of three steps. It is summarised in Figure 1 and described in detail below.

Collection of literature-based interactions. In the first step, we start collecting known interactions from the literature. We use NetPro, an expert curated and annotated database containing $\sim$100,000 protein–protein interactions [29]. These were extracted from PubMed abstracts by a semi-automated method and then cross-checked by human experts. For every

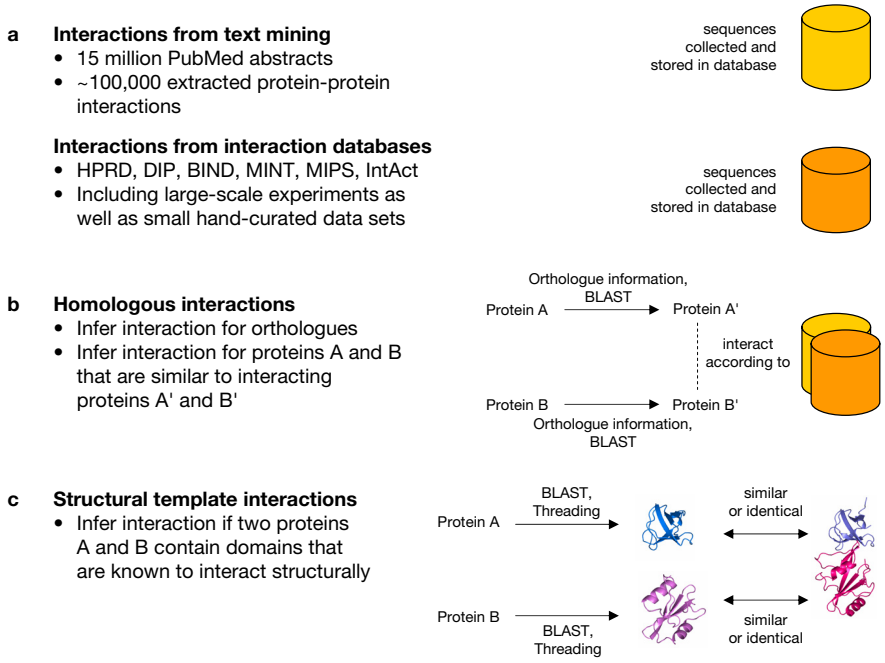

Figure 1. A three-step approach to gather all knowledge about protein–protein interactions. **a.** First, text mining is used to collect interactions from the literature. Sequences of the involved proteins are stored in a database. In a similar manner, available interaction databases are integrated and their protein sequences stored. **b.** Second, for a given protein pair A and B, a search for homologous interactions is performed using orthologue information and BLAST. Thus, interactions between homologous proteins in other species or similar proteins can serve as templates to predict an interaction between the original proteins A and B. **c.** Third, similarity to structural templates of interacting domains is used to predict interactions. To this end, we employ both sequence similarity (BLAST) and structural similarity (Threading) measures.

interaction, NetPro lists the two involved gene identifiers, species, the abstract sentences documenting the interaction, an interaction verb and an interaction nature. The interaction nature can be direct or indirect, where interactions verbs such as *binds to* classify an interaction as direct, verbs such as *colocalises with* as indirect. For every protein in NetPro, we collect its sequence from the NCBI Protein Database (Figure 1a).

Collection of interactions from interaction databases. We complement our collected literature-derived interactions with data from various interaction databases. Protein–protein interaction sets are obtained from HPRD [30], DIP [31], BIND [32], MINT [33], MIPS [34], and IntAct [35]. Again, sequences of the interacting proteins are collected and stored (Figure 1a). We are aware of potential false positives introduced by high-throughput interactions screens [10]. Since our approach aims at maximum sensitivity, we do not apply

any filter, but rather record the experimental origin of an interaction finding as confidence criterion.

Identification of interactions for a given data set. To identify known interaction in a given data set, we simply query our databases of collected interactions. The result is a protein–protein interaction network where an interaction either was described in the literature or stored in an interaction database. In a second step, we expand this network. Orthologues, i.e. homologues in other species, of proteins in our given data set are identified using the NCBI HomoloGene Database (Release 46.1). Our collected interactions are again checked for the orthologues. Following the idea of *interologs* described in [36], we predict a putative interaction between two proteins if they have interacting homologues in another species. We further extend this idea of homologous interactions by performing a BLAST search with our data set against the collected interaction sequences. Thereby, we are able to find a template protein pair A' and B' known to interact, where A' and B' are similar to two proteins A and B from our data set (Figure 1b). If the similarity (e.g. measured in sequence identity) is sufficient, we infer a putative interaction between A and B. In order to obtain a score reflecting the reliability of the prediction, we calculate the joint percentage identity. For a protein pair (A, B), this score is defined as $min(I_A, I_B)$ where I_A, I_B are the sequence percentage identities between the protein pair and the template. In this study, we require a joint percentage identity of at least 40%.

Structure-based interaction prediction. The third step of our approach is shown in Figure 1c. To predict interactions on the basis of known structures, we use SCOPPI [18], a database of domain–domain interactions and their interfaces derived from all multi-domain proteins in the Protein Data Bank [37]. Domains are defined by SCOP, the Structural Classification of Proteins. Domain residues within a distance of 5 Å to another domain are considered interacting, thus being in accord to other interface definitions [38]. As structural interaction templates for our predictions, we use a subset of SCOPPI obeying the following filter criteria: 1) interacting domains are required to be on different polypeptide chains, 2) interface size (defined as change in accessible surface area, ΔASA, calculated with Naccess) $\geq$ 600 Å^2 to filter out unspecific interfaces, 3) exclusion of homo-dimers to avoid false positive predictions between highly similar proteins. Two proteins are predicted to potentially interact if they contain domains that are known to interact structurally, according to the SCOPPI subset described above. To assign these domains to a given protein sequence, we employ both sequence similarity and structure prediction methods: First, we perform a BLAST search against a database containing the defined SCOPPI subset. By using a sequence identity cut-off at 40%, we ensure that the assigned domains have a similar structure to our structural interaction templates. Second, all protein sequences are threaded with GenTHREADER [39], considering hits with p-values $<$ 0.001 only. This procedure assigns probable SCOP domains being part of GenTHREADER's fold library to each protein sequence. As above, we assign a score of $min(I_A, I_B)$ for the sequence identities between the protein pair and the template, and $max(p_A, p_B)$ where p_A, p_B are the threading p-values, respectively.

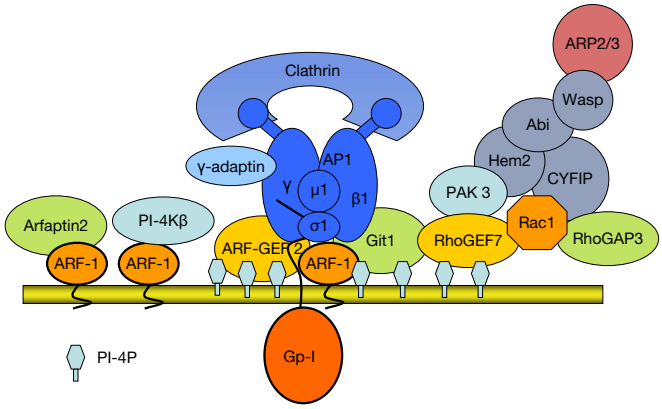

Figure 2. Theoretical model of the adaptor protein 1 (AP-1) related machinery. Several small GTPases along with their GAP and GEF effectors are involved. Some of the depicted interactions are known, while others are still presumptions [41].

Data set origin: Experimental identification of coat proteins. The data set of proteins used in our study was previously obtained by a collaborating group [40]. Aim of this study was to identify cytosolic proteins that are involved in the adaptor protein 1 (AP-1) coat assembly. The result comprises 35 murine proteins that could be selectively recruited onto liposomes that exhibit cytoplasmic domains of AP-1 cargo molecules. Among these, the AP-1 complex, clathrin, several GTPases and their effectors as well as an actin nucleation machinery was found. Table 1 shows the 35 proteins identified by mass spectrometry. How these proteins spatially arrange on liposome membranes is still speculative. Here, our method can help to suggest possible interactions and thus aid to formulate hypotheses concerning the recruited molecular machinery.

Results and discussion

Construction of an interaction network for proteins from clathrin-coated vesicles. As a use case for our method, we choose the data set described above. It contains 35 proteins shown in Table 1. Figure 2 shows the putative spatial arrangement of a subset drawn by the collaborating expert biologist.

Figure 3 shows the resulting network after taking the steps described in detail in Materials and Methods. Interactions that could be found from the literature or interaction databases are shown in red. Homologous interactions in close species (human, rat) are also red. Predicted interactions, inferred by orthology in remote species or by sequence similarity, are depicted in yellow. In addition, indirect literature interactions are dashed, whereas direct ones are solid lines. Blue lines indicate predictions based on structural templates.

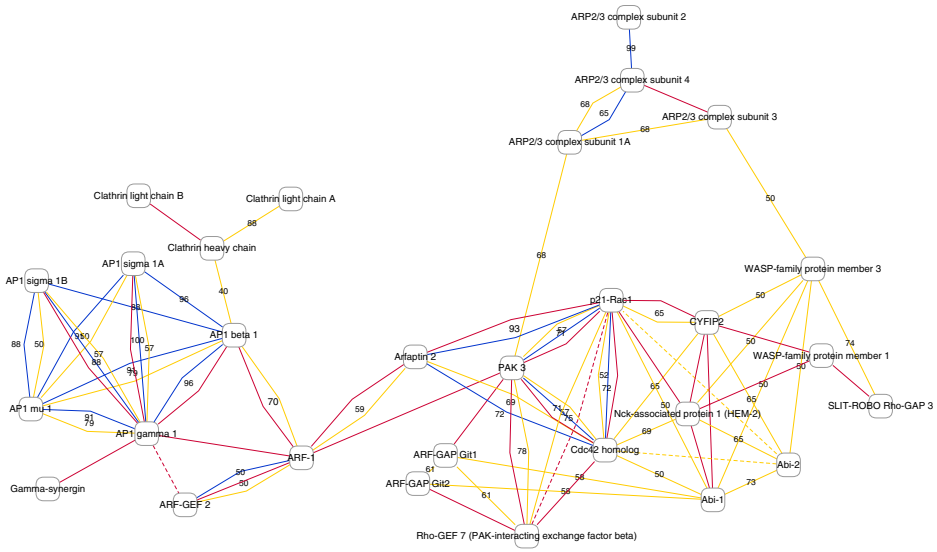

Figure 3. Constructed interaction network for the AP-1 complex. Interactions known by literature or interaction database are depicted in red, predictions based on these are yellow, and blue lines represent predictions based on 3D structural templates. Numbers indicate the sequence identity in percent that the predicted interaction shares with the template. The higher this number, the more reliable the prediction. For the sake of clarity, the cut-off is set at 50% here, except for the discussed example of the AP-1–Clatrhin example.

In case of yellow and blue lines, numbers specify the joint sequence identity percentage score of the prediction. The result allows for the generation of several new hypotheses about the molecular machinery around AP-1. In the following, three predicted examples will be discussed in detail.

Literature- and interaction databases approach. The ability to blast against a multitude of literature interactions represents a quite powerful tool. First, it allows for the detection of seamless grades of similarity. Second, it deals with the problem of synonyms in a very elegant way. Once text mining matched a protein name to a protein entity, we do not further rely on synonyms but on the sequence which—in combination with the species—unambigously described the protein entity.

PAK3 is a potential interactor of the Arp2/3 complex subunit 1A. The literature-derived search predicts, for example, the interaction between murine p21-activated kinase 3 (PAK3) and the Actin-related protein 2/3 complex subunit 1A. Basis of this prediction is a literature-documented interaction between human p21-activated kinase 1 (PAK1) and human Arp2/3 subunit 1B, reported in [42]. Overall sequence identity is 69% between the kinases, and 83% between the Arp2/3 subunits. We cannot be not sure if this interaction

is indeed true, but our method provides evidence that an interaction is likely. This would suggest that PAK3 functions in the given data set of mouse proteins in a similar manner as PAK1, namely by phosphorylating the Arp2/3 complex, thus influencing vesicle motility. Further inspection of this example reveals additional support for the prediction. In the abstract of [42], we learn that PAK1 phosphorylates p41-Arc (another name for the Arp2/3 subunit) on threonine 21. As we check the alignment, we find threonine 21 present in a well-conserved region in both proteins.

Text mining challenges: Clathrin should be linked to the AP-1 complex subunit beta. General problems of text mining still affect our approach. If interactions are not extracted in the first place, we lack this information and hence cannot infer any similar predictions. In our study, this occurs in case of the AP-1—Clathrin interaction. It is has long been known that Clathrin is associated with adaptor proteins on clathrin-coated vesicles that mediate traffic of between intracellular compartments. The physical interaction between Clathrin and the beta 1 and beta 2 subunits of the AP complexes was first described 1993 in [43]. However, neither the incorporated interaction databases, nor the literature-based NetPro database contain this particular interaction. For NetPro, the reason seems obvious. The relevant sentence in the abstract of [43] states: *"It was found that, in the absence of all the other AP subunits, beta 1 and beta 2 interact with clathrin."*. The fact that the interaction partners are just described as "beta1" and "beta2" makes it extremely hard for an algorithm to reason that these two are actually AP subunits. As we are lacking this interaction, and since no such interaction could be inferred from structural templates, we cannot connect Clathrin with the AP-1 complex in the interaction network. There is, however, a homologous interaction our method detects: According to interaction database DIP [31], yeast beta-adaptin homolog APL2 interacts with yeast clathrin heavy chain 1 [44]. The BLAST search of our method picks up the similarity between the yeast and mouse orthologues of AP-1 beta subunit (40 % identities, 62 % positives) and Clathrin heavy chain (49 % identities, 70 % positives). On this basis, our approach predicts a potential interaction between the two murine proteins. In this case, it turns out that the prediction is correct, with the known, but missing interaction validating our prediction.

Structural template-based approach. Structures are available for a considerable number of the AP1-related proteins (Table 1). The first stage easily detects these templates by sequence identity search via BLAST. If at 40% sequence identity cut-off no domain structures are found, we employ threading to assign domains. For every protein pair, we check if the SCOPPI database lists any of the assigned domains as interacting. If so, we mark these two proteins as potentially interacting (for details, see Materials and Methods).

Since crystal structures are available for the AP-1 and the Arp2/3 complex, the structural templates-based approach connects the subunits according to their contacts (blue lines in Figure 3). One interesting candidate is the predicted connection between the Rho GTPase CDC42 and Arfaptin, an effector of the Arf GTPase.

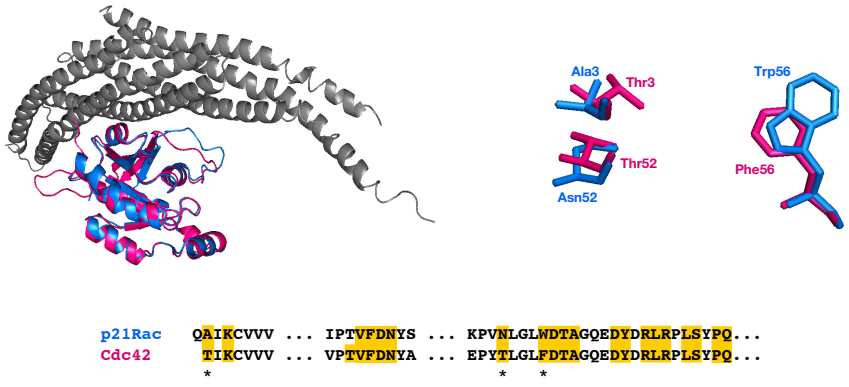

Figure 4. Structural modelling of the CDC42–Arfaptin 1 interaction. Left: The known complex structure p21Rac1 (blue) interacting with Arfaptin 1 (grey) serves as a template. CDC42 (red) and Rac1 (blue) structurally align reasonably well (RMSD 0.8). Bottom: Residues contributing to the binding site in the complex. Interacting residues are marked yellow, mismatching interacting residues are marked with a star. Right: Interface mismatches (Ala/Thr, Asn/Thr, Trp/Phe) after structural superposition of CDC42 and Rac1. Although being different amino acids, their structural side chain arrangement is similar.

The small Rho GTPase CDC42 is a potential interactor of Arfaptin 1. The small GTPases Rho, Rac and CDC42 are regulators of actin structures, cell adhesion and motility. Here, we predict the interaction between CDC42 and Arfaptin 1. As template, we use the crystal structure of RAC1-GDP in complex with Arfaptin (PDB ID 1i4l, [45]). Figure 4 shows the superposition of CDC42 and Rac1 with an RMSD of 0.8. The interfacial residues of both GTPases, as defined to be within 5 Angstrom distance to the Arfaptin, are aligned and highlighted in yellow. Closer examination of the three mismatches (Ala/Thr, Asn/Thr, and Trp/Phe) in the interface reveals that all three residues align reasonably well in the superposition of CDC42 and Rac1. We therefore have reason to believe that at least from a steric point of view the interaction is feasible.

Reliability scores and evidence. Our approach generates reliability scores as well as supporting evidence for the protein–protein interactions predicted. For structure-based predictions, we provide a structural template as well as confidence scores. These are sequence identity percentages and/or threading p-values. For literature-derived predictions, we provide statements from PubMed articles which explicitly document details of the interaction. In addition, joint sequence identity scores are available for every predicted interaction. These are 100% for known interactions described in the literature and lower for decreasing degrees of potential homology. An introduction of various cut-off levels could account for the different interaction nature of proteins, e.g. a stricter level for GTPases, and a more relaxed level for unspecific protein interactions. If the binding site is known, the matches of interfacial residues serve as additional parameters for the quality of the predicted inter-

action. The more conserved the interface, the more likely the interaction.

Limitations of method. Our method shares the common limitations of interaction prediction methods. Technical false positives (i.e. those due to the method) are likely, especially for predictions with low joint sequence percentage. Biological false positives (i.e. interactions that could be observed in vitro, but have no biological relevance, because the two proteins are not expressed in the same tissues or compartments, or not at the same time) can at least in this study be ruled out due to the experimental setup used to produce the data set. It ensures that the tested proteins are within close proximity, thus displaying a considerable potential to form interactions.

Another problem is that of the specificity of our predictions. Small GTPases and their effectors (such as GTPase activating and guanine nucleotide exchanging proteins) are good examples for forming specific interactions [46]. The problem can be addressed by a rigorous sequence identity threshold, as suggested above.

Although a considerable number of proteins in our study are known by structure, and although our method has access to $\sim$25,000 different domain interaction templates, we are not able to link the AP-1 or the Arp2/3 complex to any other protein in the data set by merely using structural information. This points at the problem that there are still comparatively few multi-domain structures available that can serve as modelling templates for interactions. As a positive outlook, we observe a supra-linear growth of these templates.

Evaluation of method. It is difficult to assess predicted interactions by other means than the biological experiment. The main problem is estimation of a false positive rate. How can one be sure that two proteins do *not* interact? Simple absence of the interaction in reliable data sources is not sufficient—the interaction might just not have been discovered yet. The *closed-world assumption*, i.e. interactions not known are also not true, does not hold for biology. Our predicted interactions are currently being tested by our collaborating group. The result of these experiments will allow for a thorough evaluation of our method.

Summary

We propose a fully automated method for the retrieval and prediction of protein–protein interactions. By merging information from literature abstracts stored in PubMed, interaction databases, and structures in the Protein Data Bank PDB we obtain a comprehensive picture on documented interactions. On the basis of this knowledge, we can construct an interaction network for any given data set, and further extend it predicted interactions at various confidence levels based on sequence or structural similarity to known interaction templates.

Applied to a data set of proteins that form coat complexes on vesicle membranes, our method identifies almost all relevant interactions. Further interactions are predicted, among them potential linkers for the AP1- and the Arp2/3 complex. Therewith, we provide po-

tential interaction candidates for further experimental testing. By incorporating the whole spectrum of text mining interactions described in the biomedical literature, data stored in interaction databases, and all structurally known domain–domain interactions, our method ensures a comprehensive network reconstruction that can assist the molecular biologist. Applying it on a genome-wide scale, we can further scale up this network to a systems biology level that provides a view on a whole interactome, thus providing a valuable tool for the life sciences.

Acknowledgements. Funding by EFRE project CODI no. 4212/04-07 is kindly acknowledged.

References

[1] Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92**, 291–294 (1998).

[2] Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623–627 (2000).

[3] Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**, 4569–4574 (2001).

[4] Gavin, A.-C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).

[5] Ho, Y. *et al.* Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature* **415**, 180–183 (2002).

[6] Giot, L. *et al.* A protein interaction map of Drosophila melanogaster. *Science* **302**, 1727–1736 (2003).

[7] Li, S. *et al.* A map of the interactome network of the metazoan C. elegans. *Science* **303**, 540–543 (2004).

[8] Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).

[9] Gavin, A. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* (2006).

[10] von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).

[11] Sprinzak, E., Sattath, S. & Margalit, H. How reliable are experimental protein-protein interaction data? *J Mol Biol* **327**, 919–923 (2003).

[12] Galperin, M. Y. & Koonin, E. V. Who's your neighbor? new computational approaches for functional genomics. *Nat Biotechnol* **18**, 609–613 (2000).

[13] Tong, A. H. Y. *et al.* Global mapping of the yeast genetic interaction network. *Science* **303**, 808–813 (2004).

[14] Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**, 4285–4288 (1999).

[15] Sun, J. *et al.* Refined phylogenetic profiles method for predicting protein-protein interactions. *Bioinformatics* **21**, 3409–3415 (2005). Evaluation Studies.

[16] Fraser, H. B., Hirsh, A. E., Wall, D. P. & Eisen, M. B. Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci U S A* **101**, 9033–9038 (2004).

[17] Aloy, P. & Russell, R. Structural systems biology: modelling protein interactions. *Nat Rev Mol Cell Biol* **7**, 188–197 (2006).

[18] Winter, C., Henschel, A., Kim, W. K. & Schroeder, M. SCOPPI: A Structural Classification of Protein–Protein Interfaces. *Nucleic Acids Res* **34**, 310–314 (2006).

[19] Davis, F. & Sali, A. PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics* (2005).

[20] Stein, A., Russell, R. B. & Aloy, P. 3DID: interacting protein domains of known three-dimensional structure. *Nucleic Acids Res* **33**, 413–417 (2005).

[21] Aloy, P. *et al.* A complex prediction: three-dimensional model of the yeast exosome. *EMBO Rep* **3**, 628–635 (2002).

[22] Aloy, P. *et al.* Structure-based assembly of protein complexes in yeast. *Science* **303**, 2026–2029 (2004).

[23] Jensen, L. J., Saric, J. & Bork, P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet* **7**, 119–129 (2006).

[24] Ben-Hur, A. & Noble, W. S. Kernel methods for predicting protein-protein interactions. *Bioinformatics* **21 Suppl 1**, i38–46 (2005).

[25] Espadaler, J., Romero-Isart, O., Jackson, R. M. & Oliva, B. Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. *Bioinformatics* **21**, 3360–3368 (2005). Evaluation Studies.

[26] Kim, W. K., Bolser, D. M. & Park, J. H. Large-scale co-evolution analysis of protein structural interlogues using the global protein structural interactome map (PSIMAP). *Bioinformatics* **20**, 1138–1150 (2004).

[27] Han, D.-S., Kim, H.-S., Jang, W.-H., Lee, S.-D. & Suh, J.-K. Prespi: a domain combination based prediction system for protein-protein interaction. *Nucleic Acids Res* **32**, 6312–6320 (2004).

[28] Owen, D. J., Collins, B. M. & Evans, P. R. Adaptors for clathrin coats: structure and function. *Annu Rev Cell Dev Biol* **20**, 153–191 (2004).

[29] `http://www.molecularconnections.com`.

[30] Mishra, G. R. *et al.* Human protein reference database–2006 update. *Nucleic Acids Res* **34**, 411–414 (2006).

[31] Salwinski, L. *et al.* The database of interacting proteins: 2004 update. *Nucleic Acids Res* **32**, 449–451 (2004).

[32] Bader, G. & Hogue, C. Bind–a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics* **16**, 465–77 (2000).

[33] Zanzoni, A. *et al.* Mint: a molecular interaction database. *FEBS Lett* **513**, 135–140 (2002).

[34] Pagel, P. *et al.* The mips mammalian protein-protein interaction database. *Bioinformatics* (2004). JOURNAL ARTICLE.

[35] Hermjakob, H. *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res* **32**, 452–455 (2004).

[36] Walhout, A. J. *et al.* Protein interaction mapping in c. elegans using proteins involved in vulval development. *Science* **287**, 116–122 (2000).

[37] Berman, H. M. *et al.* The protein data bank. *Nucleic Acids Res* **28**, 235–242 (2000).

[38] Tsai, C. J., Lin, S. L., Wolfson, H. J. & Nussinov, R. A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. *J Mol Biol* **260**, 604–620 (1996).

[39] Jones, D. T. Genthreader: an efficient and reliable protein fold recognition method for genomic sequences. *J Mol Biol* **287**, 797–815 (1999).

[40] Baust, T., Czupalla, C., Krause, E., Bourel-Bonnet, L. & Hoflack, B. Proteomic analysis of adaptor protein 1a coats selectively assembled on liposomes. *Proc Natl Acad Sci U S A* (2006).

[41] Hoflack, B. Personal communication.

[42] Vadlamudi, R. K., Li, F., Barnes, C. J., Bagheri-Yarmand, R. & Kumar, R. p41-Arc subunit of human Arp2/3 complex is a p21-activated kinase-1-interacting substrate. *EMBO Rep* **5**, 154–160 (2004).

[43] Gallusser, A. & Kirchhausen, T. The beta 1 and beta 2 subunits of the AP complexes are the clathrin coat assembly components. *EMBO J* **12**, 5237–5244 (1993).

[44] Yeung, B. G., Phan, H. L. & Payne, G. S. Adaptor complex-independent clathrin function in yeast. *Mol Biol Cell* **10**, 3643–3659 (1999).

[45] Tarricone, C. *et al.* The structural basis of Arfaptin-mediated cross-talk between Rac and Arf signalling pathways. *Nature* **411**, 215–219 (2001).

[46] Hall, A. (ed.) *GTPases* (Oxford University Press, 2000).