Structural systems biology: modelling protein interactions

Patrick Aloy[*‡] and Robert B. Russell[‡]*

Abstract | Much of systems biology aims to predict the behaviour of biological systems on the basis of the set of molecules involved. Understanding the interactions between these molecules is therefore crucial to such efforts. Although many thousands of interactions are known, precise molecular details are available for only a tiny fraction of them. The difficulties that are involved in experimentally determining atomic structures for interacting proteins make predictive methods essential for progress. Structural details can ultimately turn abstract system representations into models that more accurately reflect biological reality.

Structural genomics
Initiatives to solve X-ray or NMR structures in a high-throughput manner. They are usually focused on a single organism, pathway or disease, or are aimed at providing a complete set of protein folds (by solving representative structures, on the basis of which all other structures can be modelled).

Homology modelling
A method of protein-structure prediction that uses a known structure as a modelling template for a homologue that has been identified on the basis of sequence similarity.

Systems biology means different things to different people[1]. There are those who see it as a logical continuation of functional genomics — that is, carrying out experiments on the genome scale with the aim of understanding how the whole is greater than the sum of its parts (see, for example, REFS 2,3). Others see it as a branch of mathematical biology (see, for example, REFS 4,5), which consists of the study of small systems for which sufficient parameters have been measured to allow simulations of how the molecules function together to achieve a particular outcome. In our view, it is both of these things. Molecular biology is no longer dominated by studies of single macromolecules — studying pathways, complexes or even entire organisms is now the norm.

Genome-sequencing projects have provided a near complete list of the components that are present in an organism, and post-genomic projects have aimed to catalogue the relationships between them. Systems biology is mainly about making sense of these relationships when they are considered together. For example, understanding metabolic and signalling pathways or gene-regulatory networks relies on a detailed knowledge of protein–metabolite, protein–protein and protein–nucleic-acid interactions.

A full understanding of how molecules interact comes only from three-dimensional (3D) structures, as they provide crucial atomic details about binding. Knowledge of these details allows the more rational design of experiments to disrupt an interaction and therefore to perturb any system in which the interaction is involved.

Structural-genomics initiatives and the advancing pace of structural biology mean that it is increasingly rare to find a single protein for which no structural information is available or for which structural information is not readily accessible by straightforward homology modelling[6]. It is probable that a near-complete structural picture will be available for most of the proteins in any given organism soon. However, structural biology remains limited in terms of what it can deliver, and still struggles with the structures that are of the most relevance to systems biology — that is, those in which two or more macromolecules are in contact. Large protein complexes or whole systems require years of study for a detailed structural understanding to be reached. To address this problem, several new techniques have emerged to predict and model the structures of interacting proteins. We review these here, and discuss how they are already having an impact on our understanding of complex biological systems.

What makes structural biology so hard?

Determining the 3D structures of proteins has been hard work since the beginning. The first X-ray structures took decades to solve (see, for example, REF. 7). However, the situation has markedly progressed, and now individual protein structures can be determined in a matter of days when sufficient material is available. Modern overexpression and purification procedures can usually supply sufficient material for structural studies on a single protein, but obtaining sufficient material can be an enormous problem

*Institució Catalana de Recerca i Estudis Avançats (ICREA) and Institute for Research in Biomedicine (IRB), Parc Científic de Barcelona, Josep Samitier 1–5, 08028 Barcelona, Spain.
‡European Molecular Biology Laboratory (EMBL), Meyerhofstrasse 1, 69117 Heidelberg, Germany.
Correspondence to R.B.R.
e-mail: russell@embl.de

doi:10.1038/nrm1859

Interactome
The protein-interaction equivalent of the genome. It denotes the set of interactions that occur in an organism.

for large complexes. The reasons are fairly simple: complex assembly, although not well understood, is something that requires precise control and timing in the cell, and this is not easy to reproduce in a laboratory setting. Working with large complexes therefore usually involves years of tinkering with systems to obtain material from natural sources and growing many hundreds or thousands of litres of culture (see, for example, REFS 8,9). This is necessary because milligram concentrations of material are often needed to attempt the difficult task of growing crystals that will diffract to a high resolution — a task that is also more difficult for complexes than for individual proteins.

Reassuringly, several advances are beginning to address these problems. For example, attempts to express the subunits of a complex together in various organisms have shown promise[10–12], and improvements in both crystallization techniques and synchrotron radiation facilities mean that smaller amounts of material can be used to solve large structures. Elsewhere, relatively new techniques such as cryo-electron microscopy (for example, REF. 13), which can reconstruct structures from samples at very low temperatures, can provide lower-resolution structures for large complexes using much smaller amounts of material.

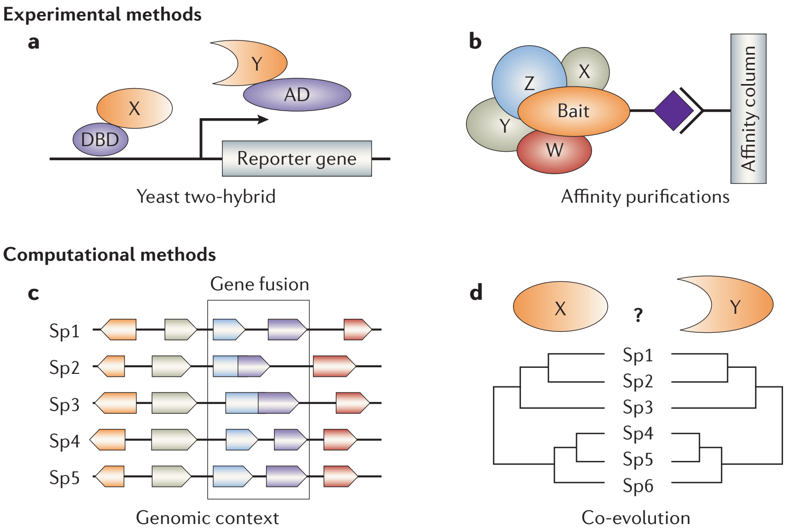
However, there is still a large gap between the number of complexes that are thought to exist on the basis of data from, for example, two-hybrid[14–18] or affinity-purification[19–21] techniques and the number for which experimental 3D structures are available. And this gap is growing with the arrival of the first drafts of the human interactome[22,23]. This has essentially defined the next generation of structure prediction — rather than focusing on single proteins, prediction techniques must now tackle whole complexes or systems to have the most impact in biology[24,25].

What and how to predict

There are many prediction challenges in the protein-interaction world. Perhaps the most obvious is simply to predict ‘who interacts with whom’ (BOX 1). The first drafts of whole-organism interactomes from high-throughput protein-interaction approaches are still far from complete[26,27], and can therefore be complemented by computational predictions. The late 1990s saw the emergence of several methods that have this aim. Perhaps the best known are those that are based around ‘genomic context’. Here, the unifying theme is to propose interactions between proteins for which there is evidence of an association, because of similarities either in how they are placed relative to each other in the hundreds of known genome sequences or in their expression profile[28–30] (BOX 1). For example, genes that lie in the same bacterial operon often encode proteins that are functionally associated.

Functional associations that are derived from a genomic context do not necessarily imply a direct physical interaction between two molecules. Proteins at opposite ends of a single pathway or complex can give the same signal as those in tight, direct, physical contact. Moreover, errors in the underlying genome or expression data can also lead to false predictions or to interactions being missed. To overcome these problems, several groups are developing methods to combine several types of interaction data quantitatively (genome context, expression or other)[31,32], whilst also considering the accuracy of each dataset. The result is an overall confidence score for each interaction, and higher scores are more likely to indicate direct physical contacts. The challenge for the future is to somehow make more biophysical inferences from such diverse data sources, for example, by correlating association scores with real physical measurements like dissociation constants[33], or to develop specific high-throughput quantitative assays to measure such constants directly[34].

Box 1 | Uncovering protein interactions

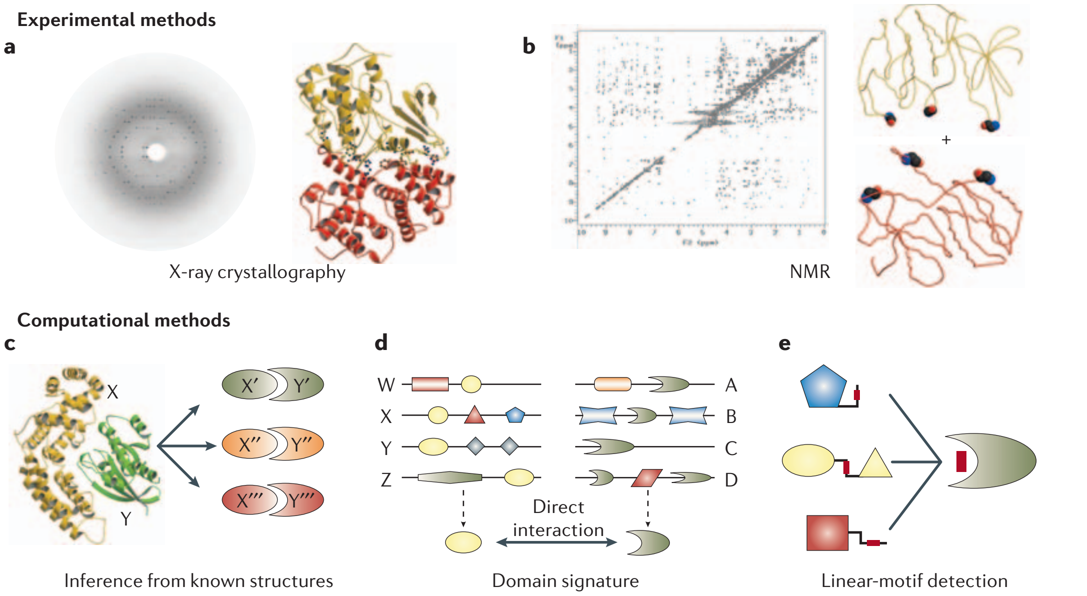


Many efforts have been undertaken to provide comprehensive lists of protein–protein interactions and uncover the secrets behind cell networks. Experimental methods include chemical crosslinking, chemical footprinting, protein arrays, fluorescence resonance energy transfer and, more recently, fluorescence cross-correlation spectroscopy[95]. However, the most widely used systems remain the yeast two-hybrid system and affinity purifications. The idea behind the yeast two-hybrid system is simple (see figure, part **a**). In the most common variant, a bifunctional transcription factor (usually GAL4) is split into its DNA-binding domain (DBD) and its activation domain (AD). Each segment is then fused to a protein of interest (X and Y) and if these two proteins interact, the activity of the transcription factor is reconstituted. The system has been scaled up and applied in genome-scale screens[14–18,22,23]. For affinity purification (see figure, part **b**), a protein of interest (bait) is tagged with a molecular label (dark purple in the figure) to allow easy purification. The tagged protein is then co-purified together with its interacting partners (W–Z), which are usually identified by mass spectrometry. This strategy has also been applied on a genome scale[19–21].

Protein–protein interactions can also be predicted computationally, with accuracies that are comparable to those of large-scale experiments[27]. Most of the computational efforts are based on the comparison of complete genome sequences. For example, if two protein-coding genes are found to be separate in one species (Sp) and fused to form a single gene in another (see figure, part **c**), a physical interaction is probable[30,96]. Other methods consider only the two proteins of interest (X and Y). For example, they would predict that the two proteins interact, or are functionally related, if they show a similar pattern of evolution across several species[97] (see figure, part **d**).

The many thousands of protein interactions that have been discovered experimentally have been compiled into publicly available databases[98–100]. Several groups are now developing methods to integrate interaction data quantitatively with other types of information, such as gene-expression profiles, subcellular localizations or literature mining[31,32].

Box 2 | Determining how molecules interact



Even when high-throughput screens provide a complete list of all of the interactions in a cell, the molecular details of these interactions are sparse. Variations of these high-throughput methods have been developed to allow a finer determination of the interacting domains by replicating the experiments with smaller constructs[16]. However, high-resolution three-dimensional (3D) structures of interacting proteins are still the best source of information, as they provide an atomic description of the binding interfaces. X-ray crystallography (see figure, part **a**), which is the most common technique, provides atomic-resolution models for proteins and complexes. Although it has no size limit, it is difficult to obtain sufficient material for the crystallization of large complexes. NMR (see figure, part **b**) is limited to proteins that have ~300 residues, but it has an important role in defining interaction interfaces between proteins for which the 3D structures are known (for example, see REF. 101).

The structures of interacting proteins can be modelled computationally if structures have been previously determined for suitable homologous proteins. New approaches have been developed to test whether interactions between homologous proteins can be modelled on the basis of an interaction of known structure[48,49] (see figure, part **c**).

In the absence of any known or predicted structural information, it is possible to look in a pair of interacting proteins for domain or sequence signatures that might mediate binding. Several methods have been developed that look for over-represented pairs of domains or motifs in interacting proteins[37–39,51]. These pairs not only provide an explanation for how interactions are mediated, but can be used to predict direct interactions between proteins that have been more loosely linked (for example, in gene-expression studies). These methods have also predicted several new domain–domain[37–39] (see figure, part **d**) and domain–motif[51] (see figure, part **e**) pairs.

Chemical crosslinking
The process of chemically joining two molecules using a covalent bond. Chemical agents are used to determine near-neighbour relationships, to analyse protein structure, and to provide information on the distance between interacting molecules.

Chemical footprinting
A method that takes advantage of chemical labelling to study protein–protein and protein–DNA interactions, by identifying the exact residues or DNA signature to which a protein binds.

Protein arrays
Solid-phase, ligand-binding assays that use immobilized proteins on different surfaces (for example, glass or membranes). Bound proteins are normally identified using specific antibodies.

Fluorescence resonance energy transfer
(FRET). The process of energy transfer between two fluorophores, which can be used to measure protein interactions *in vivo*. It can be used to determine the distance between two molecules or between two attachment positions in a macromolecule.

Fluorescence cross-correlation spectroscopy
(FCCS). A technique that detects the synchronous movement of two biomolecules with different fluorescent labels. It can be applied to live cells.

Both experimental and computational protein-interaction discovery/prediction methods can miss real interactions (false negatives) or identify others that are incorrect (false positives). Estimating rates for these is difficult, as there is still no 'gold standard' for positive interactions (protein pairs that are known to interact) or, more importantly, for negative interactions (protein pairs that are known not to interact). Nevertheless, using imperfect benchmark interaction sets, estimates of 30–60% false positives and 40–80% false negatives have been assigned to high-throughput studies that have used two-hybrid or affinity-purification techniques[27,35]. Roughly the same range of values is applicable when testing computational approaches[27,36], although these approaches also suffer from the lack of a definitive benchmark.

The above methods, and indeed the experimental techniques that identify interactions, say little about the molecular details of the association. However, it is sometimes possible to pinpoint finer details, such as the domains or segments of the proteins that are mediating the interaction. It is possible to narrow down the interacting parts of two proteins experimentally by repeating experiments with smaller constructs[16], and recurring 'domain signatures' in pairs of interacting proteins can be identified computationally[37–39] (BOX 2). If a pair of domains is identified repeatedly in interacting pairs of proteins, then these domains are probably mediating the interaction. Lists of these signatures can then be used either to predict interactions or to propose how newly determined interactions might be occurring. Approaches such as these usefully narrow down the parts of larger proteins that are involved in an interaction, but they still fall short of providing the atomic detail of the interface that is needed for a deeper understanding of what is going on.

Several groups have developed methods to predict atomic details for a pair of interacting proteins (BOX 2). Classic 'docking' approaches attempt to find the best docked complex on the basis of shape or electrostatic complementarity between protein surfaces[40,41]. To be accurate, the docking method generally needs high-resolution structures of the interacting proteins, and usually requires some idea of where the binding sites are from mutagenesis or other experiments. In the past, this method was only rarely applicable owing to the paucity of knowledge of both 3D structures and protein–protein interactions. However, the growth in both structure and interaction databases, and techniques such as NMR that can identify interaction surfaces on known structures, means that this method is experiencing something of a rebirth. Innovative new methods have been developed to combine docking with chemical-shift NMR experiments[42,43] (for example, HADDOCK; see Further information), and encouraging results have also been obtained by combining docking with mutagenesis studies (see, for example, REF. 44) and X-ray crystallography (see, for example, REF. 45). For the greatest applicability in the future, docking techniques will need to work with modelled protein structures: the increased pace of structure determination has led to representative structures being available for many single globular proteins or domains, but it will take many more decades for a full set of experimental structures to become available. It would also be advantageous if new docking methods could predict interactions between proteins — that is, if they could indicate whether or not a pair of proteins interact. Current methods only search for the optimal fit between two proteins, without attempting to distinguish pairs of proteins that interact from those that do not. To our knowledge, a move in this direction has not been attemped, although there is a feeling in the community that this is too great a challenge at present. Until the methods can reliably estimate binding free energy, predicting interactions might not be possible (A. Bonvin and R. Jackson, personal communication).

However, docking is not always necessary. There are now many thousands of interactions for which structural data are available[35,46], which means that it is increasingly possible to model structures for protein interactions on the basis of those that have been seen previously. Similar to most modelling efforts, the accuracy depends greatly on the degree of sequence identity between the target and the template onto which it is being modelled. When modelling an interaction, the choice of template is all the more crucial because the use of the wrong template can produce results that indicate that proteins interact through the wrong interface. This is roughly analogous to modelling a single protein on another that has a different fold. Encouragingly, however, when sequence similarity is high (for example, greater than 25–30% sequence identity) proteins are highly likely to interact in the same way[46], although exceptions are possible[47]. There are cases in which interactions are structurally similar despite there being no sequence similarity; the trick is to find them.

The past five years have seen the emergence of a new class of techniques that model interacting structures by homology (for example, InterPReTS[48] and MULTIPROSPECTOR[49]; see Further information). The idea is simple — use protein–protein complexes for which coordinate data are available to model interactions between their homologues. The methods are based on techniques that have been borrowed from protein-fold recognition (or threading) — namely empirical pair potentials — which are used to assess how well a homologous pair of sequences 'fit' onto a previously determined structure of a complex. The principles of native interaction interfaces are learned from known structures, and these are used to test new interfaces that have been modelled on the basis of homology. In this way, it has become possible to predict specificities for large protein families (for example, those between fibroblast growth factors (FGFs) and their receptors[48]) and to predict interactions on a genome scale by applying these techniques to all of the possible interacting proteins[36]. However, these approaches are far from perfect, and they suffer if the interactions involve conformational changes at the interface, or if the modelled interfaces contain insertions or deletions with respect to the template that are not accurately modelled.

The above methods usually assume that the proteins will interact using two relatively large interfaces (that is, domain–domain interactions). However, it is well established that many interactions, particularly those of lower affinity, are, in fact, mediated by one domain binding to a small stretch of polypeptide in another protein. These types of interaction are difficult to detect and study computationally or experimentally, because they often involve unstructured parts of the polypeptide chain that become ordered only on binding[50]. Interactions that involve phosphorylation events fall into this category, as do hundreds of short peptide sequences that are known to bind to particular domains (for example, Src-homology-3 (SH3) domains bind the motif PXXP, where X is any amino acid). There are likely to be more such interactions in nature than those that are known at present[51]. Peptides that associate with a particular domain often share a consensus sequence pattern or linear motif that can, in principle, be useful for predicting new interaction sites for cases in which the sites have not been determined experimentally. The Eukaryotic Linear Motif (ELM) resource (see Further information) provides the means to do this for several hundreds of known motifs[52]. There are also several other initiatives that aim to predict phosphorylation sites, such as NetPhos[53] and PhosphoELM[54] (see Further information), by deriving principles from known sites for particular kinases. Structural information can also have a role in finding new interaction sites for domain–motif pairs that are already known (for example, SH3-domain–PXXP-motif interactions). Recent years have seen the development of several new approaches that use the structures of known protein–peptide-ligand complexes to identify potential new ligands in genomes (see, for example, REFS 55,56). Such approaches have already predicted putative new binding sites for Src-homology-2 (SH2) domains, SH3 domains, WW domains and PDZ domains (see iSPOT in the Further information).

Despite the clear advances, it is worth remembering that all of the methods that predict these types of

Protein-fold recognition (or threading)
A method of protein-structure prediction that attempts to find a suitable template on which to model a protein of unknown structure regardless of any sequence similarity (because dissimilar sequences can adopt similar protein folds). The sequence being queried is fitted, or threaded, onto a library of known structures to find out which one is most compatible (as measured by various structural criteria — for example, how well hydrophobic residues are buried).

SH3 domain
(Src-homology-3 domain). A protein of about 50 amino acids that recognizes and binds to sequences that are rich in proline residues.

SH2 domain
(Src-homology-2 domain). A protein motif that recognizes and binds to tyrosine-phosphorylated sequences, and thereby has a key role in relaying cascades of signal transduction.

WW domain
A protein-interaction domain that is characterized by a pair of tryptophan residues that are 20–22 amino acids apart, and an invariant proline residue within a region of 40 amino acids. WW domains interact with proline-rich regions, including those containing phosphoserine or phosphothreonine.

PDZ domain
(postsynaptic-density protein of 95 kDa, Discs large, Zona occludens-1). A protein-interaction domain that often occurs in scaffolding proteins and is named after the founding members of this protein family.

interaction — whether they are structure-based or otherwise — are error prone. Often, only a handful of experimentally verified interaction sites are available, which makes deriving general principles — and indeed even error rates — difficult. Moreover, most interaction motifs have only a handful of important residues, which makes it probable that interactions will be predicted by chance for non-functional sequences. Some interaction sites, such as those for the cyclin-dependent kinases (CDKs), have a relatively complex motif (SPXR for CDKs) and can therefore be predicted with some confidence. However, others, such as those for tyrosine kinases, have little in common apart from a phosphorylated tyrosine and are therefore nearly impossible to predict from sequence alone.

All interactions details — whether identified experimentally or predicted computationally — need to be considered together with protein context. Proteins can be expressed at different times during the cell cycle, and in different tissues or cellular compartments. It is certainly possible to see interactions experimentally or to predict or model interactions between proteins that do not, in fact, ever 'see' each other in nature[48]. Computational and experimental results both need to allow for the fact that an *in vitro* interaction might have no *in vivo* meaning.

There are not only many types of protein–protein interaction, but many different strengths of interaction. There is an affinity range of more than ten orders of magnitude across functionally relevant interactions in the cell (K_d values are typically between ~10^{-14}–10^{-5}M; REF. 57). The techniques discussed above can give some crude insights into affinities on the basis of the interaction type (for example, domain–domain interactions are normally tighter than domain–peptide interactions, and phosphorylated peptides usually bind more strongly than other peptides), but accurate values are difficult to obtain experimentally or theoretically. The development of generalized systems to determine or predict kinetic parameters for protein interactions (for example, K_d, k_{on} and k_{off} values) is certainly an important challenge for the future.

Complex complexes

Macromolecular complexes are the foundation of biological activity in cells — they are the tiny machines that carry out most of the textbook processes. For example, replication, transcription, splicing, translation and metabolism are all carried out by a series of molecular machines. Many complexes have now been identified in the sense that their components have been determined, but 3D structural information is available for only a few of them. It is therefore timely to develop approaches that can determine the structures of complexes on the basis of the structural information that is available for the individual subunits and their interactions, and to couple this approach with any available low-resolution structural information (for example, from electron microscopy).

RAD51
The early steps of recombination involving homologous pairing and strand exchange are promoted by proteins of the RecA/RAD51 family of recombinases in all organisms. Human RAD51 is a relatively small protein, but it is functional as a long helical polymer that is made up of hundreds of monomers.

X-ray crystallographers — perhaps frustrated by the technical problems that are involved in solving the structures of large complexes — often use data from other sources to obtain the best possible structural model. For example, Nagai and co-workers determined the crystal structures of two subcomplexes from human small nuclear ribonucleoprotein particles (snRNPs) and, in the absence of structural information for the complete seven-component snRNP ring complex, managed to build a single model that was consistent with the protein-interaction, mutagenesis and electron-microscopy data that were available at the time[58]. Now, such approaches are essential when working with large complexes. The structures of the bacteriophage-T4 baseplate[59], RAD51 (REF. 60), the ribosome[61] and actin–myosin fibres[62] have all been elucidated using these hybrid approaches (BOX 3).

Computational biologists have since taken up the challenge of constructing complexes from their component parts. The relatively high proportion of structures that are available for individual subunits makes the use of docking methods an attractive possibility. For example, Nussinov, Wolfson and co-workers have developed a multi-docking procedure, in which docking results are first considered for the components in a pairwise fashion and are then combined to generate the most coherent structures for a complex[63]. Our own research is focused on first finding suitable 3D templates on which to model binary interactions, and then combining them in a similar fashion[24,25].

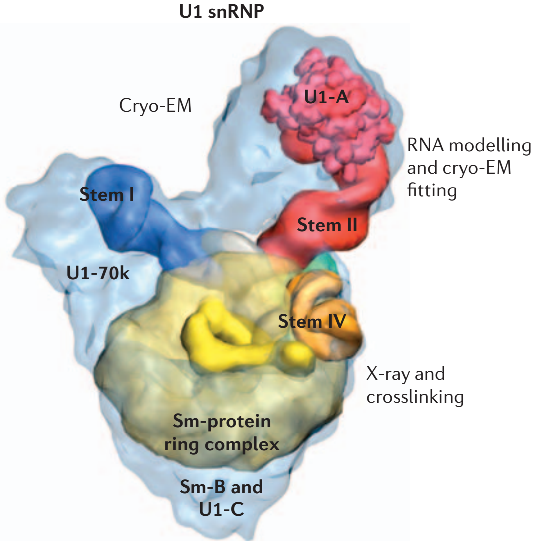
More structured pathways

Pathways have long been a convenient way of summarizing the results of many hundreds of experiments in order to chart the flow of signals or metabolites in a cell. The past decade has prompted the creation of several databases of metabolic and signalling pathways (for example, BioCyc[64], Kyoto Encyclopedia of Genes and Genomes (KEGG)[65], BioCarta, Signal Transduction Knowledge Environment (STKE) and Reactome[66]; see Further information). In general, these resources represent the relationships between molecules in a cell either as reactions or as activation or inhibition events. Although some of them try to capture specific details of the interaction (for example, phosphorylation sites), like most systems-biology data sets, they generally lack functional details about what an arrow between two proteins actually means.

Known structures, predicted interactions and predicted binding sites can greatly illuminate the understanding of a pathway. FIGURE 1 shows the known or predicted structural details for a part of the FGF signalling pathway. Crystal structures are available for the interaction of FGF1 and FGF2 with two of the three extracellular domains of the receptors FGFR1 and FGFR2a, respectively (see, for example, REFS 67,68). Crystal structures are also available for FGF10–FGFR2b (REF. 69) and FGF1–FGFR3c (REF. 70). This structural information covers only a small fraction of the possible interactions between these families of ligands and receptors. There are over 30 FGF homologues in humans and at least 7 different receptors, and the specificities are not well established. There is therefore a need for interaction-modelling techniques (see, for example, REFS 48,49) to predict the precise pairings.

Box 3 | Hybrid methods for determining the structures of complex assemblies

In the absence of atomic-resolution data, approximate atomic models of complex assemblies can be derived using a combination of several lower-resolution techniques. For example, a three-dimensional (3D) structure of the small nuclear ribonucleoprotein particles (snRNPs) that bind to pre-messenger RNA to form the spliceosome has been proposed by integrating different data types[102]. The figure shows the modelling strategy that was used to obtain the structure of the U1 snRNP. High-resolution structures of the Sm proteins and U1-A were determined by X-ray crystallography, and the relative positions of the Sm proteins inside the ring (yellow) were determined by crosslinking studies. The RNA pieces were modelled according to the known binding interactions (stem II, red; stem IV, orange; and stem I, dark blue). The total volume was determined by cryo-electron microscopy (cryo-EM; light blue), and all the available 3D models were fitted into the cryo-EM map. Finally, the approximate volumes of the proteins U1-70k, Sm-B and U1-C, for which 3D structures are not available, were used to predict their location in the cryo-EM map.



Hybrid approaches have also been used to propose structures for other big macromolecular complexes, such as the *Saccharomyces cerevisiae* exosome[103], the bacteriophage-T4 baseplate[59] and RAD51 (REF. 60). Hybrid methods are also well suited to the study of complex dynamics, because cryo-EM can often capture complexes in different conformations. For example, substantial rearrangements in the components of the ribosome are needed to account for the large differences that are observed between the tight and loose conformations of the 70S ribosome of *Escherichia coli*[61]. From a more functional perspective, fitting the atomic models of actin and myosin into a 14-Å-resolution cryo-EM map has revealed the molecular details of the actin–myosin interaction[62].

The figure was kindly provided by H. Stark, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany.

The binding of ligands to FGFRs leads to receptor dimerization and the subsequent activation of the kinase domain that is located inside the cell and is linked to the ligand-binding domains by a transmembrane helix. Crystal structures are available for the dimeric form of the intracellular tyrosine-kinase domain of FGFR1 (REF. 71) (FIG. 1). This structure, when combined with knowledge of how other kinases bind ligands and ATP, might help in the identification of suitable target phosphorylation sites for this kinase, both intramolecular and intermolecular. Autophosphorylation of the kinase (see, for example, REF. 72) leads to the recruitment of several other proteins — such as FRS2 (FGFR substrate-2) and SHC1 (SH2-domain-containing transforming protein-1) — through their phosphotyrosine binding (PTB) or SH2 domains, and to the subsequent phosphorylation of sites on these proteins by the receptor tyrosine kinase[73,74]. These sites are then recognized by the SH2 domains of other proteins, including GRB2 (growth-factor-receptor-bound protein-2; REF. 75) and SHP2 (SH2-domain-containing tyrosine phosphatase-2; REF. 76). Crystal structures or models are available for many of these PTB and SH2 domains (see, for example, REFS 77–79), but modelling is required to infer how they interact with their substrates (that is, modelling on the basis of other PTB- or SH2-domain structures that have been solved in complex with their peptide substrates; see, for example, REF. 80). GRB2 and SHP2 then bind to other proteins further downstream in the cascade. For example, the C-terminal SH3 domain of GRB2 binds to proline-rich segments in the C terminus of SOS1 (son-of-sevenless-1; REF. 81) — an interaction that can be modelled on the basis of other SH3–peptide complexes (see, for example, REF. 82) (FIG. 1). SOS1 binds to Ras, and a human crystal structure is available for this interaction[83]. Ras then goes on to bind to Raf1, and this interaction can be modelled on the basis of a related crystal structure[84]. The model reveals that the binding of Raf1 and SOS1 to Ras are mutally exclusive, because they bind to the same part of the Ras molecule (FIG. 1). Similar details can be gleaned for much of the rest of this pathway (and several others), from the receptor all the way to the nucleus. The quest for structural understanding can also highlight important missing details — for example, the precise location of phosphorylation sites is often not known, which necessitates methods to predict them (see above).

Combining pathways with 3D details ultimately makes them more useful for systems biology. If the nature of an interaction is known (for example, domain–domain versus domain–peptide), then it is easier to estimate the affinity of the association. Structures can also give information on the order of events in a pathway, by indicating which interactions cannot occur simultaneously owing to a common binding interface and by indicating, for example, that SH2 binding must be preceded by tyrosine phosphorylation. It also provides a more rational basis for deciding how to interfere with a pathway in order to study it or to treat a particular disease. Chemically tractable targets, such as protein kinases, can be considered in relation to their

Exosome
A protein complex found in eukaryotes and archae that has 3′→5′ exonuclease activity and is involved in RNA processing and degradation.

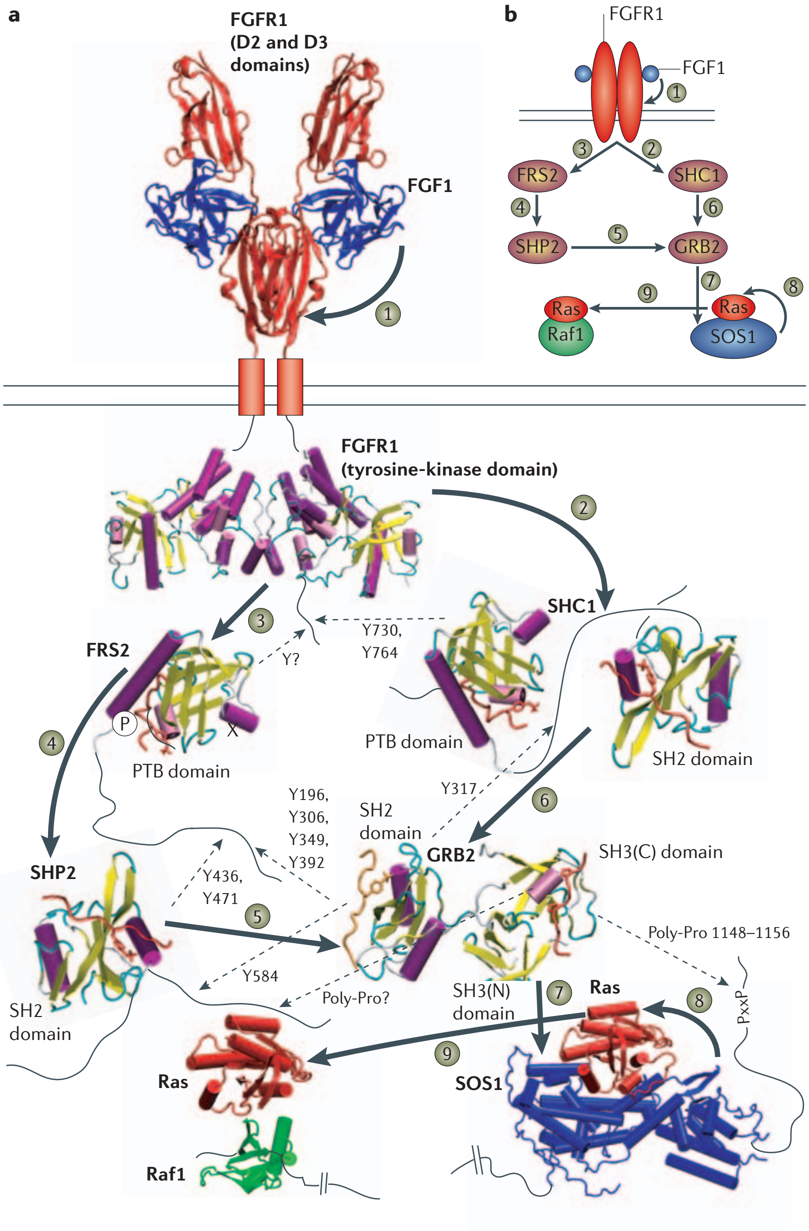


Figure 1 | **Structural details of part of the fibroblast-growth-factor signalling pathway. a** | The structural details for part of the fibroblast growth factor (FGF) pathway. **b** | The same pathway shown schematically. In part **a**, the structures of three protein complexes are shown (FGF1–FGF-receptor-1 (FGFR1) (blue–red), son-of-sevenless-1 (SOS1)–Ras (blue–red) and Ras–Raf1 (red–green)). The structures that are involved in domain–peptide or phosphorylation interactions are coloured according to the secondary structure of the domain (α-helices are in purple, β-strands are in yellow, and turns are in light blue), and the interacting peptides are coloured red or orange, or are shown schematically. Solid black arrows denote activation events, whereas dashed arrows indicate an interaction between a domain of one protein and a particular region of another (if known, the labels on these arrows indicate the residues that the domain binds). The steps are as follows. Step 1, FGF1 binds FGFR1, which dimerizes and autophosphorylates. Step 2, the SH2-domain-containing transforming protein-1 (SHC1) phosphotyrosine-binding (PTB) domain binds phosphotyrosine (pTyr) on FGFR1, and FGFR1 phosphorylates SHC1. Step 3, the FGFR substrate-2 (FRS2) PTB domain binds pTyr on FGFR1 and FGFR1 phosphorylates FRS2. Step 4, the SH2-domain-containing tyrosine phosphatase-2 (SHP2) SH2 domain binds pTyr on FRS2. Step 5, the growth-factor-receptor-bound protein-2 (GRB2) Src-homology-2 (SH2) domain binds pTyr on SHP2 (SHP2 is possibly phosphorylated by FGFR1) and its C-terminal Src-homology-3 (SH3) domain binds SHP2. Step 6, the GRB2 SH2 domain binds pTyr on SHC1. Step 7, the C-terminal SH3 domain of GRB2 binds the SOS1 Pro-rich region (GRB2 can also bind to FRS2). Step 8, the SOS1 globular domains bind and activate Ras. Step 9, Ras binds Raf1, which results in mitogen-activated-protein-kinase recruitment. The Protein Data Bank accession codes that were used to create the structures shown in this figure are: 1EVT for the FGF1–FGFR1 complex; 1IRS, which was used to model the PTB domains of FRS2 and SHC1; 1FMK, which was used to model the SH2 domains of SHP2, GRB2 and SHC1; 1GRI and 1N5Z, which were used to model the N- and C-terminal SH3 domains of GRB2, respectively; 1BKD for the Ras–SOS1 complex; and 1C1Y, which was used to model the Ras–Raf1 complex.

promiscuity (in terms of the number of substrates they have), and less tractable interactions, such as large protein–protein interfaces, can be avoided.

The previous example illustrates how known structures, when combined with modelling, can provide insights into the interacting components of a well-studied pathway. However, a more tantalizing possibility is to use interaction modelling/prediction as a means to propose new pathway elements, or indeed pathways that are completely new. A combination of interaction data and gene-expression information can indicate sets of proteins that function together in some way. Combining this information with methods to predict protein interactions can be used to give insights into the molecular details of the interactions, and therefore to give some guidance regarding the order of events (see above). Such approaches might also allow us to determine whether clusters of interacting proteins correspond to a single large complex or to a set of proteins that belong to a pathway.

A crucial part of studying any system is to consider the system *in vivo*. Strictly speaking, no pathway truly exists as an independent entity in nature. For example, a metabolic pathway is a representation of a set of co-localized proteins — with various concentrations, interaction partners and 3D structures — that behave a certain way in the presence of particular metabolites. Signalling pathways are also somewhat artificial, and are often a collection of molecules that have been selected from a much larger network to illustrate a particular aspect of biological function. For example, the Wnt (Wingless/Int-1)[85] and glycogen-synthase-kinase-3 (REF. 86) pathways are highly overlapping and differ mainly in terms of the molecule that is central to the biological principle being discussed. A structural perspective on these systems should allow us to step away from these rather artificial, discrete divisions, and to move towards a more systematic definition of the pathways in a cell.

From networks to real life

Interaction networks provide a convenient framework for understanding complex biological systems and the study of their inherent properties has proven extremely useful. However, these properties are sometimes too abstract to be readily applicable to biology and, again, the networks lack structural details.

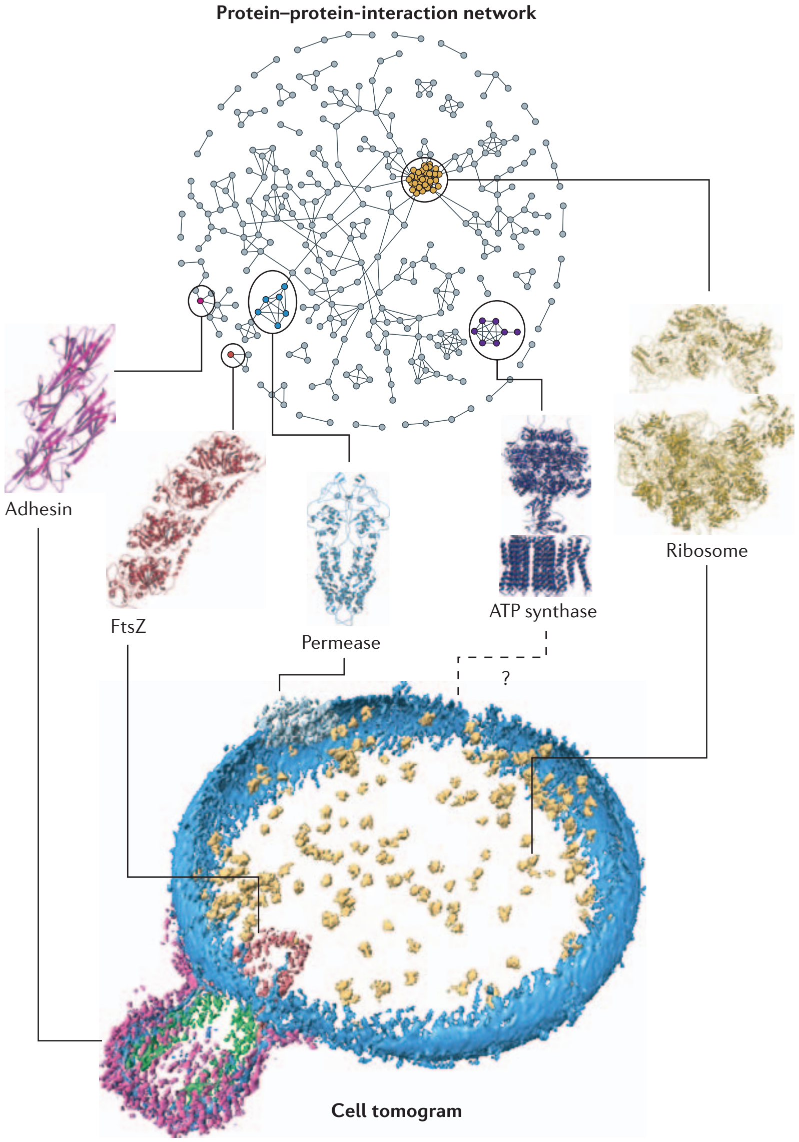


Figure 2 | **Moving from abstract networks to real cells.** A schematic figure that shows the relationship between the interaction network for the bacterium *Mycoplasma pneumoniae* and a whole-cell tomogram. The protein–protein-interaction network was derived using STRING (high-confidence interactions only; see Further information)[32]. The middle of the figure shows a few large complexes that can be homology modelled on the basis of equivalent structures that have been determined for other species, and the lines show where some of these structures can be found in the tomogram. Others, for example, ATP synthase, still need to be located. The tomogram shows the rendered surface for several cell entities — the cell membrane (dark blue), ribosomes (yellow), transport vesicles (which contain permeases; light blue), the cytoskeleton (which is mainly composed of the tubulin homologue FtsZ (pink); pink and green), and adhesion proteins (purple). The tomogram was kindly provided by A. Frangakis (European Molecular Biology Laboratory (EMBL), Heidelberg, Germany). The Protein Data Bank accession codes that were used to make this figure are: 1R17 for adhesin; 1FA0 for FtsZ; 1PFM for permase; 1C17 for the ATP synthase; and 1FFK and 1FJT for the ribosome.

Soon after the first genome-scale protein-interaction maps became available, certain trends became clear. Perhaps the best known is that biological networks tend to be scale free[87] — that is, the number of connections per molecule is not distributed randomly. Instead, they follow a power-law distribution such that most nodes have only a few connections, and a small number of 'hubs' are highly connected. The deletion of such hubs is often lethal, which is logical because something so centrally connected probably affects many crucial cellular processes[88]. Thinking in terms of structure can have a role in rationalizing these hubs, as well as other abstract properties of the network. For example, it is difficult to imagine how a protein can interact with over 100 partners unless it has a global housekeeping function (for example, it could be involved in post-translational modification, refolding or degradation) and uses the same surface in all the interactions. Indeed, this is the case for many of these hubs: they are often proteins that are responsible for the correct functioning of the entire cell, not only a particular process. These include chaperones and components of the protein-synthesis or -degradation machinery, whereas proteins in the network periphery, with few interaction partners, are more often their targets. However, in some cases, there seem to be no general trends: some hubs bind to dozens of other proteins without any obvious pattern to explain how or why. It is also possible that some of the most well-connected hubs might correspond to experimental artefacts[89].

By combining interaction maps and expression data, it has recently been argued that some of these highly connected molecules are 'party hubs', which tend to interact with other proteins continuously throughout the cell cycle, whereas others are 'date hubs', which make specific connections at particular times with different partners[90]. The structural rationale for this observation is clear. Party hubs are probably the central components of large complexes, such as the catalytic subunit of the RNA polymerases, which are only fully active when the rest of the complex components are present. By contrast, date hubs are probably proteins that function on substrates or bind cofactors at specific times, such as cyclin-dependent kinase-2, which associates with different cyclins at different points during the cell cycle (see, for example, REF. 91).

Towards a molecular picture of an entire cell

A more challenging future role for networks will be to guide the interpretation of results from one of the most exciting areas of structural biology — electron tomography[92,93]. This technique is now delivering 40-Å-resolution images of single cells and is showing the true network-like structure of proteins in a cell. A significant problem for these initiatives is to deduce what is being seen in the tomograms. Ribosomes and elements of the cytoskeleton are readily identifiable, as are several other large complexes for which X-ray or relatively

Electron tomography
A structural technique that allows a single cell to be studied using cryo-freezing and by obtaining data using a series of tilt angles in the electron beam, such that a three-dimensional image can be reconstructed.

high-resolution electron-microscopy images are already available[94]. However, complexes cannot be assigned to most of the densities in these images. FIGURE 2 shows how some of the hubs and interconnected proteins in the *Mycoplasma pneumoniae* protein-interaction network correspond to important known complexes such as the ribosome or ATP synthase. The 3D structures of some of these complexes can be modelled on the basis of equivalent structures that have been determined for other organisms, and if they are sufficiently large, they can be identified and placed in the cell tomogram (FIG. 2).

The availability of interactomes for whole organisms will provide the complete set of complexes and interactions, but will lack structural details. The challenge is to somehow combine this information with the methods discussed above to build models for all of these complexes and to use them to identify the locations of these complexes in cell tomograms. Such approaches will reveal the real molecular organization of a cell and will allow systems biology to move from abstract representations to the physical world. At first, these models, similar to many of the currently available experimental structures, will be static and will lack the dynamic realism of live cells. However, they will provide a crucial framework for the integration of other data (for example, cellular-localization and gene-expression data) and for models that capture the specific dynamic and regulatory aspects of the cell.

Concluding remarks

Protein interactomes provide a rather abstract network of macromolecules, which can be useful for deducing the global features of a cell network. However, they have a limited relationship with physical reality. The real picture of a cell will ultimately come when complete interactomes, pathways and high-resolution tomograms can be complemented by a near complete repertoire of the 3D structures of protein complexes. This places structural biology — experimental and computational — in a crucially important position for systems biology. Structural information for interacting cellular components will produce a more and more complete whole-cell framework at atomic-level detail, which will be of immense benefit to anybody studying or modelling biological systems.

1. Levesque, M. P. & Benfey, P. N. Systems biology. *Curr. Biol.* **14**, R179–R180 (2004).
2. Auffray, C., Imbeaud, S., Roux-Rouquie, M. & Hood, L. From functional genomics to systems biology: concepts and practices. *C R Biol.* **326**, 879–892 (2003).
3. Aggarwal, K. & Lee, K. H. Functional genomics and proteomics as a foundation for systems biology. *Brief Funct. Genom. Proteom.* **2**, 175–184 (2003).
4. Kitano, H. Computational systems biology. *Nature* **420**, 206–210 (2002).
5. Rousseau, F. & Schymkowitz, J. A systems biology perspective on protein structural dynamics and signal transduction. *Curr. Opin. Struct. Biol.* **15**, 23–30 (2005).
6. Pieper, U. *et al.* MODBASE, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.* **32**, D217–D222 (2004).
7. Muirhead, H. & Perutz, M. F. Structure of haemoglobin. A three-dimensional fourier synthesis of reduced human haemoglobin at 5.5 Å resolution. *Nature* **199**, 633–638 (1963).
8. Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* **289**, 905–920 (2000).
9. Cramer, P., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* **292**, 1863–1876 (2001).
10. Berger, I., Fitzgerald, D. J. & Richmond, T. J. Baculovirus expression system for heterologous multiprotein complexes. *Nature Biotechnol.* **22**, 1583–1587 (2004).
11. Tan, S. A modular polycistronic expression system for overexpressing protein complexes in *Escherichia coli*. *Protein Expr. Purif.* **21**, 224–234 (2001).
12. Kim, K. J. *et al.* Two-promoter vector is highly efficient for overproduction of protein complexes. *Protein Sci.* **13**, 1698–1703 (2004).
13. Frank, J. Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu. Rev. Biophys. Biomol. Struct.* **31**, 303–319 (2002).
14. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
The first high-throughput application of an interaction-discovery technique: the two-hybrid system being applied to the complete genome of *Saccharomyces cerevisiae*.
15. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98**, 4569–4574 (2001).
16. Rain, J. C. *et al.* The protein–protein interaction map of *Helicobacter pylori*. *Nature* **409**, 211–215 (2001).
17. Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543 (2004).
18. Giot, L. *et al.* A protein interaction map of *Drosophila melanogaster*. *Science* **302**, 1727–1736 (2003).
19. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
20. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
21. Butland, G. *et al.* Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* **433**, 531–537 (2005).
22. Stelzl, U. *et al.* A human protein–protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
23. Rual, J. F. *et al.* Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437**, 1173–1178 (2005).
24. Aloy, P. *et al.* Structure-based assembly of protein complexes in yeast. *Science* **303**, 2026–2029 (2004).
The first attempt to model complexes on a large scale in an organism through the combined use of affinity purification, homology modelling and electron microscopy.
25. Aloy, P., Pichaud, M. & Russell, R. B. Protein complexes: structure prediction challenges for the 21st century. *Curr. Opin. Struct. Biol.* **15**, 15–22 (2005).
26. Aloy, P. & Russell, R. B. The third dimension for protein interactions and complexes. *Trends Biochem. Sci.* **27**, 633–638 (2002).
27. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417**, 399–403 (2002).
28. Dandekar, T., Snel, B., Huynen, M. & Bork, P. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.* **23**, 324–328 (1998).
29. Marcotte, E. M., Pellegrinin, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86 (1999).
An excellent summary of the use of genomic context to predict functional associations between proteins and its application in prokaryotes.
30. Enright, A. J., Iliopoulos, I. L, Kyrpides, N. C. & Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402**, 25–26 (1999).
31. Jansen, R. *et al.* A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science* **302**, 449–453 (2003).
32. von Mering, C. *et al.* STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33**, D433–D437 (2005).
33. Gavin, A. C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* 22 Jan 2006 (doi:10.1038/nature04532).
The first attempt to define a pseudo-biophysical measurement directly from functional genomics data (affinity-purification results) and its application in defining the modular organization of protein complexes in *S. cerevisiae*.
34. Jones, R. B., Gordus, A., Krall, J. A. & MacBeath, G. A quantitative protein interaction network for the ErbB receptors using protein microarrays. *Nature* **439**, 168–174 (2006).
35. Aloy, P. & Russell, R. B. Ten thousand interactions for the molecular biologist. *Nature Biotechnol.* **22**, 1317–1321 (2004).
36. Lu, L., Arakaki, A. K., Lu, H. & Skolnick, J. Multimeric threading-based prediction of protein–protein interactions on a genomic scale: application to the *Saccharomyces cerevisiae* proteome. *Genome Res.* **13**, 1146–1154 (2003).
The first application of interaction modelling on a genome scale. The authors suggest that this approach is roughly as accurate as high-throughput experimental approaches.
37. Sprinzak, E. & Margalit, H. Correlated sequence-signatures as markers of protein–protein interaction. *J. Mol. Biol.* **311**, 681–692 (2001).
The first attempt to deduce details of protein interactions by looking for 'domain signatures' — pairs of domains that are seen repeatedly in several interactions.
38. Wojcik, J. & Schachter, V. Protein–protein interaction map inference using interacting domain profile pairs. *Bioinformatics* **17** (Suppl. 1), 296–305 (2001).
39. Deng, M., Mehta, S., Sun, F. & Chen, T. Inferring domain–domain interactions from protein–protein interactions. *Genome Res.* **12**, 1540–1548 (2002).
40. Smith, G. R. & Sternberg, M. J. Prediction of protein–protein interactions by docking methods. *Curr. Opin. Struct. Biol.* **12**, 28–35 (2002).
41. Wodak, S. J. & Mendez, R. Prediction of protein–protein interactions: the CAPRI experiment, its evaluation and implications. *Curr. Opin. Struct. Biol.* **14**, 242–249 (2004).
42. Dominguez, C., Boelens, R. & Bonvin, A. M. HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).
43. Dobrodumov, A. & Gronenborn, A. M. Filtering and selection of structural models: combining docking and NMR. *Proteins* **53**, 18–32 (2003).
44. Morillas, M. *et al.* Structural model of a malonyl-CoA-binding site of carnitine octanoyltransferase and carnitine palmitoyltransferase I: mutational analysis of a malonyl-CoA affinity domain. *J. Biol. Chem.* **277**, 11473–11480 (2002).

45. Hothorn, M., Wolf, S., Aloy, P., Greiner, S. & Scheffzek, K. Structural insights into the target specificity of plant invertase and pectin methylesterase inhibitory proteins. *Plant Cell* **16**, 3437–3447 (2004).
46. Aloy, P., Ceulemans, H., Stark, A. & Russell, R. B. The relationship between sequence and interaction divergence in proteins. *J. Mol. Biol.* **332**, 989–998 (2003).
47. Park, S. Y., Beel, B. D., Simon, M. I., Bilwes, A. M. & Crane, B. R. In different organisms, the mode of interaction between two signaling proteins is not necessarily conserved. *Proc. Natl Acad. Sci. USA* **101**, 11646–11651 (2004).
48. Aloy, P. & Russell, R. B. Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA* **99**, 5896–5901 (2002).
The first method to use complexes of known 3D structure to test for putative interactions between the homologues of the proteins that are contained in a complex.
49. Lu, L., Lu, H. & Skolnick, J. MULTIPROSPECTOR: an algorithm for the prediction of protein–protein interactions by multimeric threading. *Proteins* **49**, 350–364 (2002).
50. Bracken, C., Iakoucheva, L. M., Romero, P. R. & Dunker, A. K. Combining prediction, computation and experiment for the characterization of protein disorder. *Curr. Opin. Struct. Biol.* **14**, 570–576 (2004).
51. Neduva, V. *et al.* Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol.* **3**, e405 (2005).
The first attempt to discover and validate new domain–motif interacting pairs in high-throughput interaction data.
52. Puntervoll, P. *et al.* ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.* **31**, 3625–3630 (2003).
53. Blom, N., Gammeltoft, S. & Brunak, S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.* **294**, 1351–1362 (1999).
54. Diella, F. *et al.* Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics* **5**, 79 (2004).
55. de Rinaldis, M., Ausiello, G., Cesareni, G. & Helmer-Citterich, M. Three-dimensional profiles: a new tool to identify protein surface similarities. *J. Mol. Biol.* **284**, 1211–1221 (1998).
56. Sheinerman, F. B., Al-Lazikani, B. & Honig, B. Sequence, structure and energetic determinants of phosphopeptide selectivity of SH2 domains. *J. Mol. Biol.* **334**, 823–841 (2003).
57. Zhou, H. X. Association and dissociation kinetics of colicin E3 and immunity protein 3: convergence of theory and experiment. *Protein Sci.* **12**, 2379–2382 (2003).
58. Kambach, C. *et al.* Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell* **96**, 375–387 (1999).
59. Kostyuchenko, V. A. *et al.* Three-dimensional structure of bacteriophage T4 baseplate. *Nature Struct. Biol.* **10**, 688–693 (2003).
60. Shin, D. S. *et al.* Full-length archaeal Rad51 structure and mutants: mechanisms for RAD51 assembly and control by BRCA2. *EMBO J.* **22**, 4566–4576 (2003).
61. Gao, H. *et al.* Study of the structural dynamics of the *E. coli* 70S ribosome using real-space refinement. *Cell* **113**, 789–801 (2003).
62. Holmes, K. C., Angert, I., Kull, F. J., Jahn, W. & Schroder, R. R. Electron cryo-microscopy shows how strong binding of myosin to actin releases nucleotide. *Nature* **425**, 423–427 (2003).
63. Inbar, Y., Benyamini, H., Nussinov, R. & Wolfson, H. J. Prediction of multimolecular assemblies by multiple docking. *J. Mol. Biol.* **349**, 435–447 (2005).
64. Karp, P. D. *et al.* Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **33**, 6083–6089 (2005).
65. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277–D280 (2004).
66. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res.* **33**, D428–D432 (2005).
The description of a systems-biology representation of pathway information — a qualitative framework on which quantitative data can be superimposed when they become available.
67. Plotnikov, A. N., Schlessinger, J., Hubbard, S. R. & Mohammadi, M. Structural basis for FGF receptor dimerization and activation. *Cell* **98**, 641–650 (1999).
68. Stauber, D. J., DiGabriele, A. D. & Hendrickson, W. A. Structural interactions of fibroblast growth factor receptor with its ligands. *Proc. Natl Acad. Sci. USA* **97**, 49–54 (2000).
69. Yeh, B. K. *et al.* Structural basis by which alternative splicing confers specificity in fibroblast growth factor receptors. *Proc. Natl Acad. Sci. USA* **100**, 2266–2271 (2003).
70. Olsen, S. K. *et al.* Insights into the molecular basis for fibroblast growth factor receptor autoinhibition and ligand-binding promiscuity. *Proc. Natl Acad. Sci USA* **101**, 935–940 (2004).
71. Mohammadi, M. *et al.* Structures of the tyrosine kinase domain of fibroblast growth factor receptor in complex with inhibitors. *Science* **276**, 955–960 (1997).
72. Mohammadi, M. *et al.* Identification of six novel autophosphorylation sites on fibroblast growth factor receptor 1 and elucidation of their importance in receptor activation and signal transduction. *Mol. Cell. Biol.* **16**, 977–989 (1996).
73. Dunican, D. J., Williams, E. J., Howell, F. V. & Doherty, P. Selective inhibition of fibroblast growth factor (FGF)-stimulated mitogenesis by a FGF receptor-1-derived phosphopeptide. *Cell Growth Differ.* **12**, 255–264 (2001).
74. Zhou, M. M. *et al.* Structure and ligand recognition of the phosphotyrosine binding domain of Shc. *Nature* **378**, 584–592 (1995).
75. Cussac, D., Frech, M. & Chardin, P. Binding of the Grb2 SH2 domain to phosphotyrosine motifs does not change the affinity of its SH3 domains for Sos proline-rich motifs. *EMBO J.* **13**, 4011–4021 (1994).
76. Hadari, Y. R., Kouhara, H., Lax, I. & Schlessinger, J. Binding of Shp2 tyrosine phosphatase to FRS2 is essential for fibroblast growth factor-induced PC12 cell differentiation. *Mol. Cell. Biol.* **18**, 3966–3973 (1998).
77. Farooq, A., Zeng, L., Yan, K. S., Ravichandran, K. S. & Zhou, M. M. Coupling of folding and binding in the PTB domain of the signaling protein Shc. *Structure* **11**, 905–913 (2003).
78. Nioche, P. *et al.* Crystal structures of the SH2 domain of Grb2: highlight on the binding of a new high-affinity inhibitor. *J. Mol. Biol.* **315**, 1167–1177 (2002).
79. Maignan, S. *et al.* Crystal structure of the mammalian Grb2 adaptor. *Science* **268**, 291–293 (1995).
80. Zhou, M. M. *et al.* Structural basis for IL-4 receptor phosphopeptide recognition by the IRS-1 PTB domain. *Nature Struct. Biol.* **3**, 388–393 (1996).
81. Li, N. *et al.* Guanine-nucleotide-releasing factor hSos1 binds to Grb2 and links receptor tyrosine kinases to Ras signalling. *Nature* **363**, 85–88 (1993).
82. Ghose, R., Shekhtman, A., Goger, M. J., Ji, H. & Cowburn, D. A novel, specific interaction involving the Csk SH3 domain and its natural ligand. *Nature Struct. Biol.* **8**, 998–1004 (2001).
83. Boriack-Sjodin, P. A., Margarit, S. M., Bar-Sagi, D. & Kuriyan, J. The structural basis of the activation of Ras by Sos. *Nature* **394**, 337–343 (1998).
84. Nassar, N. *et al.* The 2.2 Å crystal structure of the Ras-binding domain of the serine/threonine kinase c-Raf1 in complex with Rap1A and a GTP analogue. *Nature* **375**, 554–560 (1995).
85. Bejsovec, A. Wnt pathway activation: new relations and locations. *Cell* **120**, 11–14 (2005).
86. Haq, S. *et al.* Glycogen synthase kinase-3β is a negative regulator of cardiomyocyte hypertrophy. *J. Cell Biol.* **151**, 117–130 (2000).
87. Barabasi, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
88. Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
89. Aloy, P. & Russell, R. B. Potential artefacts in protein-interaction networks. *FEBS Lett.* **530**, 253–254 (2002).
90. Han, J. D. *et al.* Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature* **430**, 88–93 (2004).
91. Miller, M. E. & Cross, F. R. Cyclin specificity: how many wheels do you need on a unicycle? *J. Cell Sci.* **114**, 1811–1820 (2001).
92. Medalia, O. *et al.* Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. *Science* **298**, 1209–1213 (2002).
The first electron tomogram of a single cryo-frozen cell at a resolution of 4nm, which reveals much of the detail of the inside of a eukaryotic cell.
93. Nickell, S., Kofler, C., Leis, A. P. & Baumeister, W. A visual approach to proteomics. *Nature Rev. Mol. Cell. Biol.* **7**, 225–230 (2006).
94. Sali, A., Glaeser, R., Earnest, T. & Baumeister, W. From words to literature in structural proteomics. *Nature* **422**, 216–225 (2003).
95. Stoevesandt, O., Köhler, O., Fischer, R., Johnston, I. & Brock, R. One-step analysis of protein complexes in microliters of cell lysate. *Nature Methods* **2**, 833–835 (2005).
96. Marcotte, E. M. *et al.* Detecting protein function and protein–protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
97. Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng.* **14**, 609–614 (2001).
98. Xenarios, I. *et al.* DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30**, 303–305 (2002).
99. Alfarano, C. *et al.* The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.* **33**, D418–D424 (2005).
100. Tetko, I. V. *et al.* MIPS bacterial genomes functional annotation benchmark dataset. *Bioinformatics* **21**, 2520–2521 (2005).
101. Qin, J., Vinogradova, O. & Gronenborn, A. M. Protein–protein interactions probed by nuclear magnetic resonance spectroscopy. *Methods Enzymol.* **339**, 377–389 (2001).
102. Stark, H., Dube, P., Luhrmann, R. & Kastner, B. Arrangement of RNA and proteins in the spliceosomal U1 small nuclear ribonucleoprotein particle. *Nature* **409**, 539–542 (2001).
103. Aloy, P. *et al.* A complex prediction: three-dimensional model of the yeast exosome. *EMBO Rep.* **3**, 628–635 (2002).

Acknowledgements

We thank T. Gibson (European Molecular Biology Laboratory (EMBL), Heidelberg, Germany), R. Jackson (University of Leeds, UK) and A. Bonvin (University of Utrecht, The Netherlands) for helpful discussions.

Competing interests statement

The authors declare no competing financial interests.

DATABASES

The following terms in this article are linked online to:

UniProtKB: http://ca.expasy.org/sprot
FGF1 | FGF2 | FGF10 | FGFR1 | FRS2 | GRB2 | SHC1 | SHP2 | SOS1

Protein Data Bank: http://www.rcsb.org/pdb
1BKD | 1C17 | 1C1Y | 1EVT | 1FA0 | 1FFK | 1FJT | 1FMK | 1GRI | 1IRS | 1N5Z | 1PFM | 1R17

FURTHER INFORMATION

BioCarta: http://www.biocarta.com
BioCyc: http://biocyc.org
ELM (The Eukaryotic Linear Motif resource for Functional Sites in Proteins): http://elm.eu.org
HADDOCK (High Ambiguity Driven protein–protein DOCKing based on biochemical and/or biophysical information): http://www.nmr.chem.uu.nl/haddock
InterPReTS (Interaction Prediction through Tertiary Structure): http://interprets.embl.de
iSPOT (Sequence Prediction Of Target): http://cbm.bio.uniroma2.it/ispot
KEGG (Kyoto Encyclopedia of Genes and Genomes): http://www.genome.jp/kegg
NetPhos 2.0 Server: http://www.cbs.dtu.dk/services/NetPhos
PhosphoELM: http://phospho.elm.eu.org
Reactome: http://www.reactome.org
STKE (Signal Transduction Knowledge Environment): http://www.stke.org
STRING (Search Tool for the Retrieval of Interacting Genes/Proteins): http://string.embl.de
Access to this interactive links box is free online.