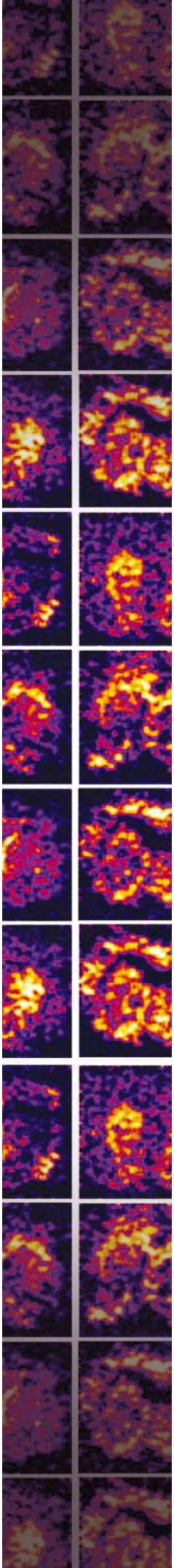


From words to literature in structural proteomics

Andrej Sali*, Robert Glaeser†, Thomas Earnest‡ & Wolfgang Baumeister§

**Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biomedical Research, University of California, San Francisco, California 94143, USA*
†Department of Molecular and Cell Biology, Stanley/Donner ASU, University of California, Berkeley, California 94720, USA
‡Berkeley Center for Structural Biology, Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA
§Department of Structural Biology, Max Planck Institute of Biochemistry, Am Klopferspitz 18 a, 82152 Martinsried, Germany
(e-mail: baumeist@biochem.mpg.de)

Technical advances on several frontiers have expanded the applicability of existing methods in structural biology and helped close the resolution gaps between them. As a result, we are now poised to integrate structural information gathered at multiple levels of the biological hierarchy — from atoms to cells — into a common framework. The goal is a comprehensive description of the multitude of interactions between molecular entities, which in turn is a prerequisite for the discovery of general structural principles that underlie all cellular processes.

The structures of individual macromolecules are often uninformative about function if taken out of context. Just as words must be assembled into sentences, paragraphs, chapters and books to make sense, vital cellular functions are performed by structured ensembles of proteins (that is, complexes), not by freely diffusing and occasionally colliding proteins[1]. Frequently, these complexes comprise ten or more subunits (Fig. 1). Recent proteomics studies with yeast, for example, have indicated that the number of complexes that exist at least transiently in a cell has been underestimated. The techniques of isolation and purification that are traditionally used in biochemistry tend to select for the most robust complexes, whereas the more weakly interacting and transient complexes escape attention and, therefore, analysis.

In recent years, two trends have emerged in structural biology: efforts to achieve a comprehensive coverage of individual protein structures (so-called structural genomics) and efforts to analyse the structures of large complexes[2,3]. Structural biology has flourished in the wake of technological innovations in fields as diverse as biochemistry, molecular biology, computational biology, computer hardware and software, nuclear magnetic resonance (NMR) magnets and optimized pulse sequences, and synchrotron radiation, as well as advances in light and electron microscopy (EM) instrumentation and in detector technology. Notwithstanding the value and importance of the individual techniques, a combination of approaches is likely to be more powerful than any single method alone. In this review we discuss some integrated strategies and tactics that can be used for characterizing molecular complexes and for describing their interactions in a cellular context.

The challenge of myriads of complexes

Given the average length of 466 residues for a yeast protein and 173 residues for a domain in the CATH database[4] (a hierarchical classification of protein domain structures), one can estimate that, on average, a protein is folded into approximately two domains. In the evolution of proteins, domains are important units that are shuffled, duplicated, and fused into larger proteins. Although the universe of distinct amino acid sequences is essentially unlimited, the number of different folding patterns for the domains is not. Extrapolation based on the existing databases of protein sequence and structure indicates that most of the natural domain sequences assume one of a few thousand folds[5], of which ~1,000 are already known[4].

In contrast to the folds, there are no satisfactory estimates of the number of different non-covalent macromolecular complexes with a unique structure and biological function. Such estimates are non-trivial to make because of the multitude of the component types (for example, proteins and nucleic acids), and the varying lifespan of the complexes (for example, transient complexes such as those involved in signalling, and stable complexes such as the ribosome). In addition, there is no self-evident definition of what is a 'complex' and whether two complexes are of different types. In an extreme view, a whole cell or even an organism may be seen as a single giant complex.

The Protein Quaternary Structure (PQS) database currently contains ~10,000 structurally defined protein assemblies of presumed biological significance, derived from a variety of organisms (http://pqs.ebi.ac.uk/pqs-doc.shtml); each assembly consists of at least two protein chains. Just like the folds, these assemblies can be organized into ~3,000 groups such that the members of the same assembly group share more than 30% sequence identity between the equivalent constituent protein chains (Fig. 1).

The most comprehensive information about both stable and transient protein complexes exists for the yeast proteome of ~6,200 proteins. But even for this model genome, uncertainties in the number, types and sizes of the complexes arise because of the difficulty in unravelling physical interactions from functional links[6], binary from multiple physical interactions, transient from stable interactions, and direct interactions from indirect physical interactions through intermediates. In addition, each method may be impacted differently by the localization of the proteins in the cellular environment and may have significantly different rates of false positives and negatives.

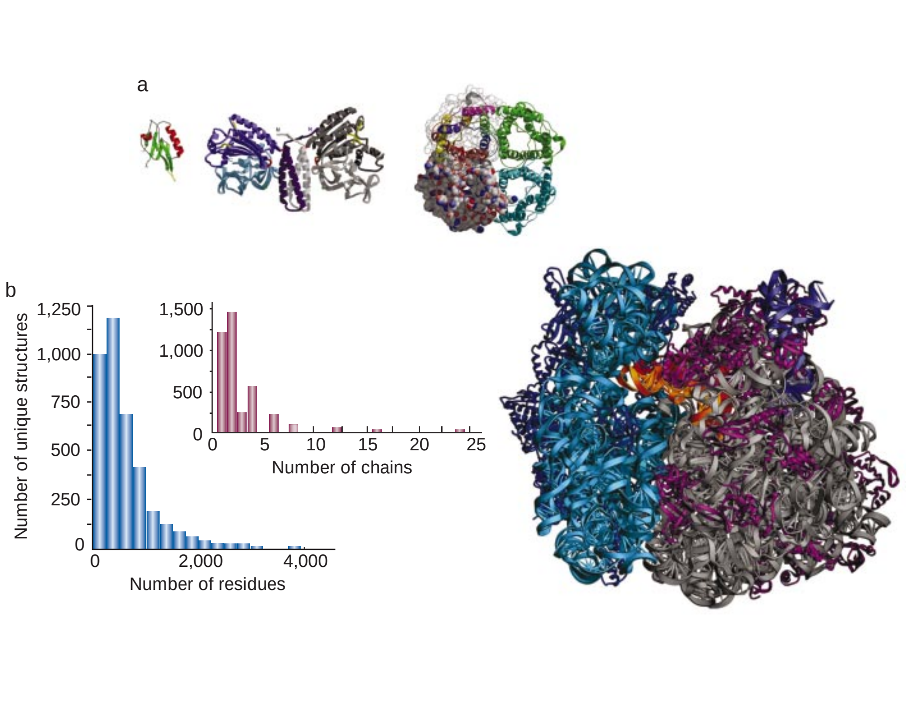


Figure 1 Illustration of the size range of biomolecular structures solved by X-ray crystallography and the size distribution of structures contained in the Protein Quaternary Structure (PQS) database (http://pqs.ebi.ac.uk). **a**, X-ray crystallography can deal with a wide range of complexity. From top left to right, structures of: the PDZ domain of dishevelled, a molecular recognition domain that leads to protein–protein interactions; CheA, a dimeric multidomain bacterial signalling molecule; aquaporin, which serves as a transmembrane water channel; and 70S ribosome, which is the molecular machine for protein biosynthesis. **b**, The main histogram shows the distribution of the size of the entries in the PQS database. The 15,190 entries with at least one protein chain of at least 30 residues, when compared with each other, produced 3,876 clusters with more than 30% sequence identity and less than 30-residue length difference among the members within the same cluster. The inset shows the distribution of the number of chains in the representative structures for each group. As expected, the structures of large complexes are under-represented, given an estimated average size of a yeast complex of 7.5 proteins.

The Munich Information Center for Protein Sequences (MIPS)[7] and Yeast Proteome Database (YPD)[8] list ~11,000 binary interactions and functional links documented by focused, small-scale experiments[9], corresponding on average to ~3.5 partners per protein. Large-scale yeast two-hybrid data[10,11] indicate 1.7 partners per protein, when artefactual interactions are removed from consideration[12]. On the other hand, the affinity purification of 1,739 yeast protein baits indicated 232 distinct complexes of an average size of 7.5 proteins, suggesting that the whole yeast proteome may contain ~900 complexes[13]. A comparison of these purified complexes against the complexes of known structure revealed that most of them are stable as opposed to transient, whereas the reverse applies to the interactions detected by the yeast two-hybrid methods[14–17]. Only one-third of the binary interactions and functional links obtained by more than one high-throughput method occur in the curated MIPS/YPD set of the ~11,000 binary interactions and links, suggesting that the lower bound on the binary protein–protein interactions and functional links in yeast is ~30,000 (refs 9,18). This number corresponds to ~9 protein partners per protein or 3.6 protein partners per domain, not necessarily all direct or at the same time.

The human proteome may have an order of magnitude more complexes than the yeast cell; and the number of different complexes across all relevant genomes may be several times larger still. Therefore, there may be thousands of biologically relevant macromolecular complexes whose structures are yet to be characterized[19].

Towards an unabridged dictionary of proteins

Currently, X-ray crystallography is the most prolific technique for the structural analysis of proteins and protein complexes, and it still is the 'gold standard' in terms of accuracy. While this technique has provided the majority of structures in the database of biomolecular structures, the fraction determined by NMR spectroscopy is also significant (currently 14%)[20]. From the earliest structures of myoglobin and haemoglobin through the recent studies of RNA polymerase[21], the ribosomal subunits[22–24], and the complete ribosome and its functional complexes[25], these structural data have contributed tremendously to our understanding of biology at the molecular level. As seen in Fig. 1, the sizes of the structures determined by X-ray crystallography range from small proteins, such as the 100-residue PDZ domain, which recognizes and binds other proteins, to the 70S ribosome, which consists of 52 proteins and 3 RNA molecules, and has a relative molecular mass of ~2,500,000 (M_r 2,500K).

Crystallography requires that milligram quantities of a pure and monodisperse protein can be prepared, and that the protein can be

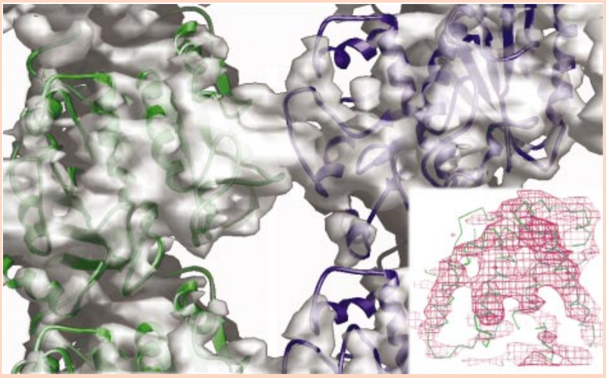


Figure 2 Docking the atomic model of tubulin into the cryo-EM density map of the assembled microtubule. The atomic model of tubulin, represented by its ribbon diagram, is shown docked into the 3D density of an intact, 13-protofilament microtubule, represented by the grey, transparent surface of the protein. The atomic model of 2D crystals of tubulin was refined at a resolution of 3.5 Å (ref. 38), and the model was docked as a rigid body into the microtubule density, which was obtained at a resolution of 8 Å by applying single-particle averaging methods to very short segments of ice-embedded microtubules[41]. The docked (hybrid) model shows which residues are responsible for forming the lateral contacts between individual protofilaments, information that could not be deduced from the structure of the protofilament alone. A short helix (upper centre) that is well ordered in the crystal structure is also shown to be disordered (lacking density) in the microtubule. The high precision with which this docking is specified by the data is shown more clearly in the insert, where the atomic model (represented by the Cα backbone) is embedded within its corresponding portion of the 3D density (represented by the wire basket volume).

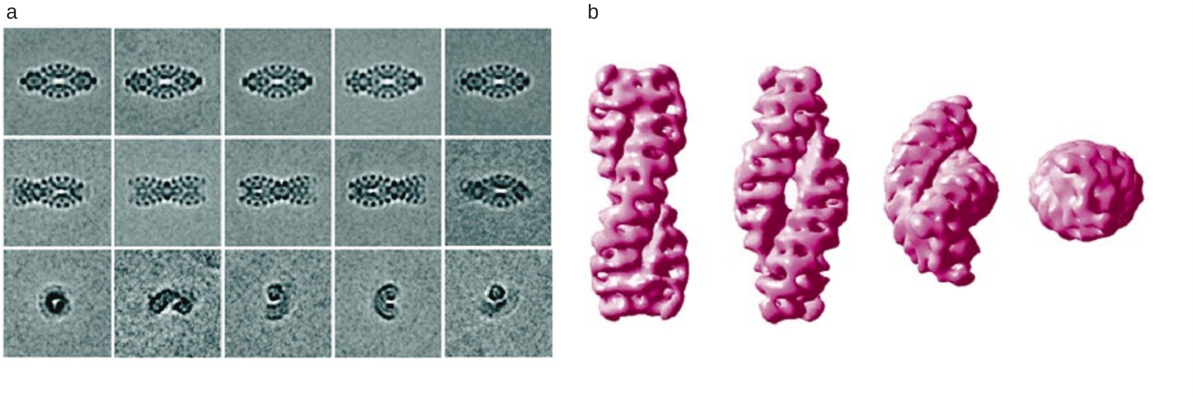


Figure 3 Representative example that illustrates the type of 3D reconstructions that can be obtained with large macromolecular complexes by single-particle cryo-EM. In this example, the specimen is a giant assembly of *Drosophila melanogaster* tripeptidyl peptidase II (TPP II). The protein monomers have a relative molecular mass of 150,000 (M_r 150K) and the intact assembly has a particle mass of ~6,600K (ref. 42). **a**, Some of the distinct views that are obtained by averaging many equivalent projections of individual particles randomly oriented within a thin film of vitreous ice. A wide variety of side views can be distinguished, corresponding to rotation of the particle around its long axis. In addition, other projections, shown in the bottom row, correspond to particles that are viewed directly on axis or at a small tilt relative to the axial view. **b**, 3D surface representation of the TPP II complex at 3.3-nm resolution, in which the particle is first rotated about its long axis and then it is tilted to bring the long axis perpendicular to the page.

induced to form three-dimensional (3D) periodic arrays (that is, crystals). Therefore, almost all proteins used for structural studies are expressed in heterologous expression systems. Bacterial expression systems are simple and rapid, in addition to being amenable to incorporation of selenium as an anomalous scatterer for determining phases. However, overexpression in bacteria may not produce large amounts of the correctly folded protein, or the protein may lack appropriate post-translational modification. To overcome such limitations, there are a number of strategies that involve using genes from different species, altering constructs, screening for solubility, and utilizing different cellular or cell-free expression systems. The constructs can be altered in numerous ways, such as by the addition of tags, separation of proteins into domains, or the use of gene shuffling methods.

Once the proteins are expressed and purified, it is necessary to form crystals of sufficient quality to collect high-resolution (at least 2.5 Å) data for structure determination. Because crystallization conditions cannot be pre-determined, it is necessary to screen a wide range of conditions (such as pH, salt, protein concentration and cofactors). Over the past few years, this area has benefited enormously from automation and technologies allowing the use of small sample volumes[26]. Particularly for proteins and protein complexes with low yields, the ability to screen more conditions at the required protein concentration is critical.

Currently, most biological crystallography experiments are done at synchrotrons, where the brightness (high flux of well-collimated X-rays) and tunability expand the capabilities and throughput enormously. The increase in the amount and diversity of structural data that have been obtained in the past five to ten years has been greatly enhanced by the availability of beamlines and detectors of increasing performance. As the systems have evolved from primitive to 'user-friendly', robotic crystal mounting and alignment systems have also been implemented at beamlines[27] to increase the throughput and productivity of these expensive and oversubscribed resources. Once data are obtained, usually in one to several hours on modern third-generation synchrotrons, the analysis of the primary data can also be completed in several hours.

Box 1

Are crystals necessary in electron crystallography?

The fundamental role of crystals within crystallography is that they make it easy to merge the data that are generated by vastly more scattering events than could be tolerated by a single molecule. But because the alignment of high-resolution images of single particles can be done *in silico*, one has to seriously ask whether or not crystals are really needed. Indeed, it now is easier to use single particles than crystals to obtain 3D reconstructions at a resolution of 1–2 nm, as long as the particle has a relative molecular mass of at least 250K–500K.

On the other hand, because the best electron micrographs rarely provide data whose quality is as good as 10% of what physics would allow it to be, it is thought that computational alignment at atomic resolution will require that particles be larger than ~2,000K–4,000K (ref. 43). Furthermore, the number of molecular images that must be merged is approximately 100 times more than would be required if the image quality were nearly perfect. The task of merging data from images of as many as one million individual particles, the number currently used to obtain high-resolution structures of specimens prepared as 2D crystals, is estimated to require at least 10^{17} floating-point operations[63]. This task would require a full day of dedicated use of a teraflop computer. Improvements in affordable clusters will soon bring this much computing power into a well-equipped cryo-EM facility, and thus it is likely that computational capacity will keep pace with the projected improvement in speed of data collection. The limiting factor for 2D crystals, on the other hand, is the low rate at which high-resolution images are obtained with highly tilted samples. The steep decline in success of recording images at high tilt angle is the result of some form of specimen charging or beam-induced movement that is still not fully understood.

Although work at atomic resolution with 2D crystals remains an attractive approach for high-profile and otherwise intractable specimens such as tubulin[35], a solution to the problem of beam-induced movement must still be found before 2D crystals can be used for work at a pace comparable to that of X-ray crystallography. It is likely that cryo-EM images of single-particle specimens would approach atomic resolution, if we could overcome the same problem that now limits the image quality in highly tilted 2D crystals. Should this type of improvement in performance be realized, accurate alignment of nearly any macromolecular complex could be done *in silico*, and crystals would indeed no longer be needed for crystallography.

Box 2

Experimental methods for structural characterization of assemblies

A variety of methods are available for the experimental determination of macromolecular assembly structure (see Fig. 4)

X-ray crystallography is the most powerful method for structure determination because it is capable of providing an atomic structure of the whole assembly[22,64]. When suitable crystals and high-resolution crystallographic data are obtained, there is little need for other methods of structure characterization.

Nuclear magnetic resonance (NMR) spectroscopy allows determination of atomic structures of increasingly large subunits and even their complexes[65–69]. Although NMR analysis is generally not as applicable as X-ray crystallography to protein structures with more than 300 amino acid residues, it can be applied to molecules in solution and is more suitable than X-ray crystallography to study their dynamics and interactions in solution.

Electron crystallography (two-dimensional electron microscopy or 2D EM) and single-particle analysis can reveal the shape and symmetry of an assembly, sometimes at near-atomic resolution, but more frequently at an intermediate resolution[70]. Segmentation of the electron density may lead to an approximate configuration of subunits in a complex[71]. Proteins whose structures are already known can then be fitted into these density maps with an accuracy approaching one-tenth the resolution of the EM reconstruction[47–50].

Electron tomography is based upon multiple tilted views of the same object[54]. Although it can be used to study the structure of isolated macromolecular assemblies at relatively low resolution, its true potential lies in visualizing the assemblies in an unperturbed cellular context.

Immuno-electron microscopy can be used to determine an approximate position of a protein in the context of an assembly[72]. This task is achieved by using a construct of the protein of interest that binds to a gold-labelled antibody. The relative position of the gold particles is then identified by EM.

Chemical crosslinking with mass spectroscopy can be used to identify binary and higher-order protein contacts[73]. The approach relies on bi- and tri-functional crosslinking reagents that covalently link proteins interacting with each other. Proteolytic digestion and subsequent mass spectroscopic identification of the crosslinked species reveal their composition. In addition, chemical crosslinking of specific residue types has recently been used to obtain intramolecular distance restraints[74].

Affinity purification with mass spectroscopy combines purification of protein complexes with identification of their individual components by mass spectroscopy (see reviews in this issue by Aebersold and Mann, page 198, and Fields and co-workers, page 208). During cell lysis, the whole assembly is partially broken into smaller complexes that are then isolated by a variety of methods, such as those relying on fusion proteins or antibodies as baits for affinity purification. Subunits in these smaller complexes are usually identified by a combination of gel electrophoresis and mass spectroscopy. Examples include the U1 subunit of the yeast and human spliceosome[75,76], identification of proteins that interact with the GroEL complex[77], the sampling of protein interactions in the yeast nuclear-pore complex[72], and a high-throughput identification of the hundreds of distinct protein complexes in budding yeast[13,78].

Fluorescence resonance energy transfer (FRET) occurs when a higher-energy fluorophore stimulates emission by a lower-energy fluorophore that is within ~60 Å of its inducer. It can be applied to monitor protein interactions if one protein is fused to a fluorescence donor and its potential partner to a fluorescence acceptor (see accompanying review by Fields and co-workers). Fluorescence donors and acceptors are usually spectral derivatives of the green fluorescence protein.

Site-directed mutagenesis and a variety of biochemical experiments (for example, footprinting) can reveal which subunits in a complex interact with each other and sometimes what face is involved in the interaction[79–81].

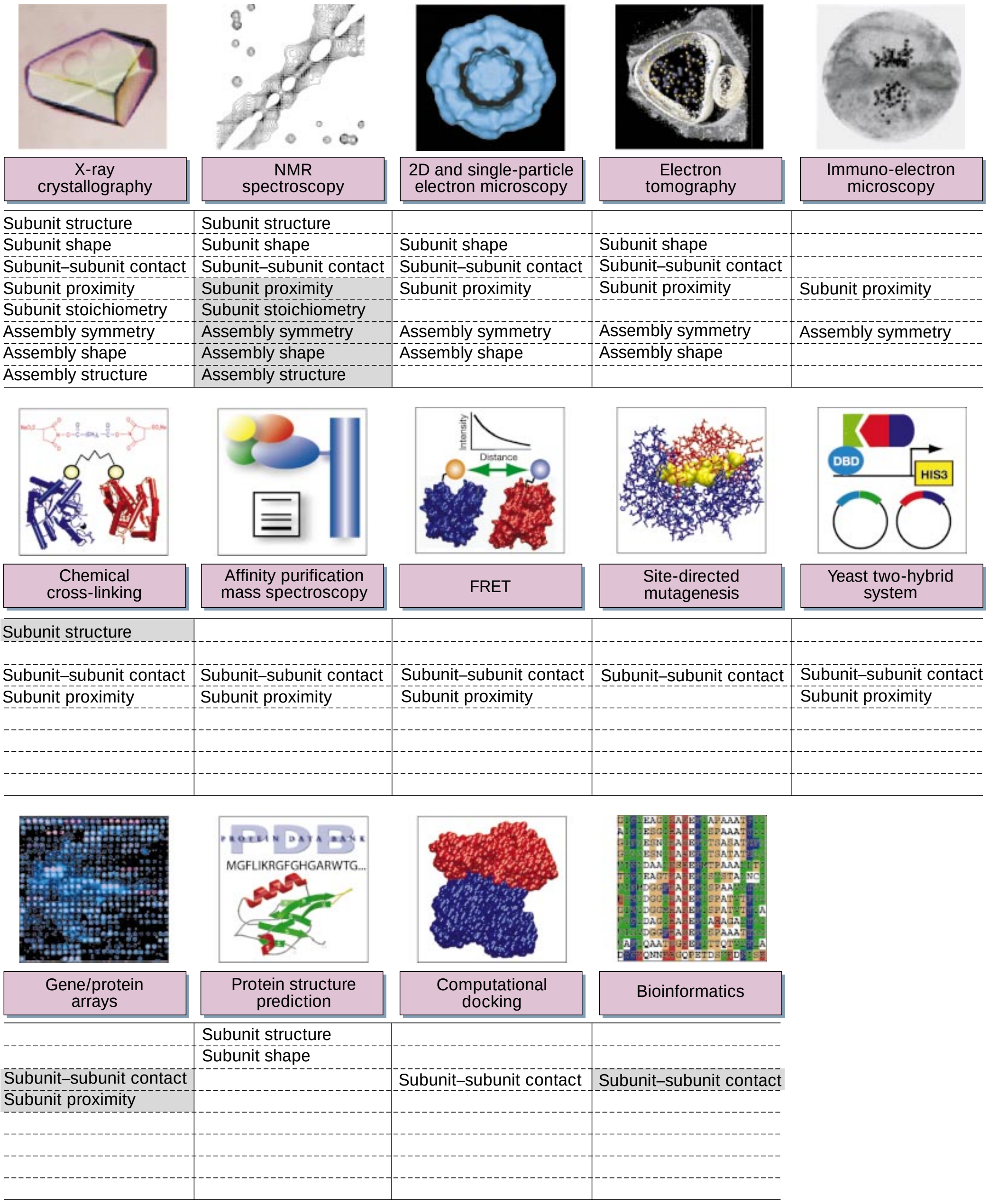
Yeast two-hybrid system detects binary protein interactions by activating expression of a reporter gene upon direct binding between the two tested proteins (see review by Fields and co-workers). The approach is based on the modularity of transcription factors that consist of a DNA-binding and an activation domain, each of them fused to two different genes encoding for the proteins whose interaction is tested. If the two expressed fusion proteins are in contact with each other, the two modules of the transcription factor are united, thereby inducing transcription of a set of reporter genes. Expression of reporter genes, in turn, is easily detected by a variety of tests, such as yeast colony colour and ability to grow in deficient media. The method is suitable for high-throughput applications (ref. 11; and see review by Fields and co-workers).

Protein arrays immobilize a variety of 'bait' proteins, such as antibodies and glutathione *S*-transferase, into an array on a specially treated surface; the array is then probed with sample proteins, resulting in a detection of binary interactions (see review by Fields and co-workers). Messenger RNA expression arrays immobilize stretches of mRNA and are used to measure the concentration of mRNA species in a sample as a function of tissue type, cell cycle and other environmental conditions[82,83]. Such data sets have been used to detect functionally linked proteins, which include proteins whose expression is co-regulated because they are members of the same assembly, are encoded on the same operon, or belong to the same biochemical pathway[6].

Increasingly, therefore, structures are solved within hours after data collection begins, although most structures still need a great deal more time for the screening of crystals, full data collection, and the processing and analysis that leads to an accurate high-resolution structure. Nevertheless, as the beamlines become more automated and as higher-level control and processing software is further developed, it is becoming feasible to integrate the data collection, processing and analysis steps — from crystal mounting through structure refinement — to form a 'pipeline' of information for structure determination. The technological advances, such as third-generation synchrotrons and charge-coupled device (CCD)-based detectors, have also been critical for the success of structure determinations of several large complexes and viruses. Crystals from such samples typically have very large unit-cell dimensions and diffract even more weakly than 'ordinary' biomolecular crystals.

Recently, several international efforts have been initiated to determine the structures of at least one member from each domain family, such that the structures of the remaining protein sequences can be characterized based on their similarity to the known structures[28,29]. Structural genomics aims to construct a taxonomy of protein structures that will serve as a 'dictionary' for the interpretation of the genomic data. In the United States, the Protein Structure Initiative of the National Institute of General Medical Sciences (NIGMS) has funded nine pilot centres to develop high-throughput pipelines for structure determination[30]. The NIGMS initiative is paralleled by similar efforts in Europe and Japan. Following the success of the genome sequencing programmes, where the use of automation has been important in the increase of productivity, these structural genomics programmes are currently implementing automation of protein production, crystallization, data collection and analysis.

Although it is legitimate to ask how successful structural genomics will be in terms of structures solved versus targets chosen for cloning, a fair assessment at this point in time is difficult. In the



X-ray crystallography	NMR spectroscopy	2D and single-particle electron microscopy	Electron tomography	Immuno-electron microscopy
Subunit structure	Subunit structure			
Subunit shape	Subunit shape	Subunit shape	Subunit shape	
Subunit–subunit contact	Subunit–subunit contact	Subunit–subunit contact	Subunit–subunit contact	
Subunit proximity	Subunit proximity	Subunit proximity	Subunit proximity	Subunit proximity
Subunit stoichiometry	Subunit stoichiometry			
Assembly symmetry	Assembly symmetry	Assembly symmetry	Assembly symmetry	Assembly symmetry
Assembly shape	Assembly shape	Assembly shape	Assembly shape	
Assembly structure	Assembly structure			

Chemical cross-linking	Affinity purification mass spectroscopy	FRET	Site-directed mutagenesis	Yeast two-hybrid system
Subunit structure				
Subunit–subunit contact	Subunit–subunit contact	Subunit–subunit contact	Subunit–subunit contact	Subunit–subunit contact
Subunit proximity	Subunit proximity	Subunit proximity		Subunit proximity

Gene/protein arrays	Protein structure prediction	Computational docking	Bioinformatics
	Subunit structure		
	Subunit shape		
Subunit–subunit contact		Subunit–subunit contact	Subunit–subunit contact
Subunit proximity			

Figure 4 Experimental and theoretical methods that can provide information about a macromolecular assembly structure. The annotations below each of the panels list the aspects of an assembly that might be obtained by the corresponding method. Subunit and assembly structure indicate an atomic or near-atomic resolution at 3 Å or better. Subunit and assembly shape indicate the density or surface envelope at a low resolution of worse than 3 Å. Subunit–subunit contact indicates knowledge about protein pairs that are in contact with each other, and in some cases about the face that is involved in the contact. Subunit proximity indicates whether two proteins are close to each other relative to the size of the assembly, but not necessarily in direct contact. Subunit stoichiometry indicates the number of subunits of a given type that occur in the assembly. Assembly symmetry indicates the symmetry of the arrangement of the subunits in the assembly. Grey boxes indicate extreme difficulty in obtaining the corresponding information by a given method.

early years, it is first necessary to establish the appropriate infrastructure, and it will take time until this investment pays dividends. Success also depends on the choice of targets; there are easy proteins and families, as well as more difficult ones, such as membrane proteins. Whereas success rates of 1–10% per attempted protein are often quoted, this estimate may be misleadingly pessimistic. Many target families have >10 members, a large number of which are usually attempted in parallel. Therefore, the likelihood that at least one of the targeted family members yields a structure is higher than 10%. Whatever the timeframe may be, there is no doubt that structural genomics will make a major contribution to the proteomics dictionary of words and phrases. But words or even phrases alone do not make literature.

Using EM images to produce three-dimensional structures

A powerful advantage of EM is the fact that it is possible to treat images of single molecules in the same way as crystalline arrays. The ability to use non-crystalline particles means, in turn, that it is possible to work

Box 3

Theoretical methods for structural characterization of assemblies

Non-experimental methods used to provide information about macromolecular assembly structure include protein structure prediction, computational docking and a variety of bioinformatics techniques (see Fig. 4).

Protein structure prediction can be used to characterize sequences whose structures have not been obtained experimentally[84]. There are two types of methods corresponding to the two distinct sets of principles that guide the behaviour of proteins on vastly different timescales: the laws of physics and the rules of evolution.

The first approach, *de novo* or *ab initio* methods, predicts the structure from sequence alone, without relying on similarity at the fold level between the modelled sequence and any of the known structures[85]. The *de novo* methods assume that the native structure corresponds to the global free-energy minimum accessible during the lifespan of the protein and attempt to find this minimum by an exploration of many conceivable protein conformations. The two key components of *de novo* methods are the procedure for efficiently carrying out the conformational search, and the free-energy function used for evaluating possible conformations.

De novo prediction of protein structure directly from its sequence is becoming increasingly more successful. For roughly 35% of proteins shorter than 150 amino acids that have been examined, one of the five most commonly recurring models generated has sufficient global similarity to the true structure to recognize it in a search of the protein structure database[86]. But the accuracy of even the 'correct' models tends to be only ~4 Å root-mean-square deviation (RMSD) over ~80 residues, too low for problems requiring high-resolution structure information.

The second class of methods of protein structure prediction, including threading and comparative or homology modelling, rely on detectable similarity spanning most of the modelled sequence and at least one known structure[87]. Modelling of a sequence based on known structures consists of four steps: finding known structures related to the sequence to be modelled, aligning the sequence with the related structures, building a model, and assessing the model. The templates for modelling may be found by sequence comparison methods or by sequence–structure threading methods that can sometimes reveal more distant relationships than purely sequence-based methods[88]. In the latter case, fold assignment and alignment are achieved by threading the sequence through each of the structures in a library of all known folds; each sequence–structure alignment is assessed by the energy of a corresponding coarse model, not by sequence similarity as in sequence comparison methods. Next, given a sequence–structure alignment, comparative model building produces an all-atom model of the sequence.

High-accuracy comparative models are based on more than 50% sequence identity to their templates. They tend to have approximately 1Å RMS error for the main-chain atoms, which is comparable to the accuracy of a medium-resolution nuclear magnetic resonance structure or a low-resolution X-ray structure. Low-accuracy comparative models are based on less than 30% sequence identity, and tend to contain less than 70% of residues within 3.5 Å of their correct positions. It is currently possible to model domains in ~60% of all known protein sequences[89]. Although the current number of modelled proteins may look impressive given the early stage of structural genomics, usually only one domain per protein is modelled (on the average, proteins have slightly more than two domains) and two-thirds of the models are based on less than 30% sequence identity to the closest template.

Computational docking is based on maximizing the shape and chemical complementarities between a given pair of interacting proteins[90]. Although these methods are generally not yet sufficiently accurate to predict whether two proteins actually interact with each other, they can sometimes correctly identify the interacting surfaces between two known or modelled subunits[91].

Bioinformatics analysis of genomic sequences, multiple sequence alignments and protein structures may indicate the presence and location of protein interaction interfaces. For example, a pair of proteins in a given genome that appear as a fused multidomain protein in another genome indicates a binary interaction between the two proteins in the first genome[6,92]. Likewise, co-occurrence of two proteins in the same genomic neighbourhood indicates a functional link, especially in prokaryotes[93]. Similarity between the phylogenetic trees for two families of orthologues also indicates an interaction[6,94,95]. Correlated mutations resulting in co-variation between alignment positions in two families of proteins are a weak signal that members of the two families may interact with each other[96]. Analyses of multiple sequence alignments and known protein structures, such as the evolutionary trace method[97], may help in identification of a binding site on a given protein structure. And finally, interactions may be inferred from considerations of protein sequence and structure homology[98,99].

with very small quantities of material, the purity need not be at the standard required for crystallization, and specimen tilting (a bottleneck discussed below) is not needed to collect data for a 3D reconstruction. The electron microscope produces images that represent only 2D projections of the specimen, in which all information about the third dimension of the object has been lost. Nevertheless, the full 3D structure of the object can be reconstructed again if one is able to start with many such projections, each showing the object from a different angle[31]. As a result, the unique contributions that can be made by EM include studies of large, complex assemblies without any requirement for crystallization, and, as will be discussed later, even their visualization within whole cells by electron tomography.

Unfortunately, the electrons in a microscope also represent a beam of ionizing radiation that damages the sample while the image is being formed. As a result, it is necessary to limit the electron exposure to a value that is so low that the images have extremely high levels of 'shot noise' (statistical variation in the number of electrons recorded at each point in the image). Equivalent images of separate molecules must therefore be averaged to reduce the statistical noise that is present in each such image.

If the specimen is one molecule thick with all molecules in the same orientation (as in a 2D crystal), the necessary spatial averaging of images is easy. In fact, 3D reconstructions that have been obtained at a high enough resolution to trace the polypeptide chain have all been produced with the use of 2D crystals[32–38]. Although only ~100 images of highly tilted crystals are needed to produce such a reconstruction, collection of this amount of experimental data is nevertheless slow because the yield of good images drops to 1% or less of that obtained with untilted specimens. As a result, structural studies with 2D crystals have only seldom been taken to a high enough resolution to allow building an atomic model directly into the 3D reconstruction.

Other specimens may exist in the form of long helices or other particles with very high symmetry (for example, icosahedra). These high-symmetry particles usually do not need to be tilted, as the individual particles are naturally rotated by a random amount relative to one another. The number of protein monomers within one such particle remains relatively small, and thus data from many equivalent particles may still have to be averaged to obtain a reconstruction. In practice, such reconstructions have rarely extended

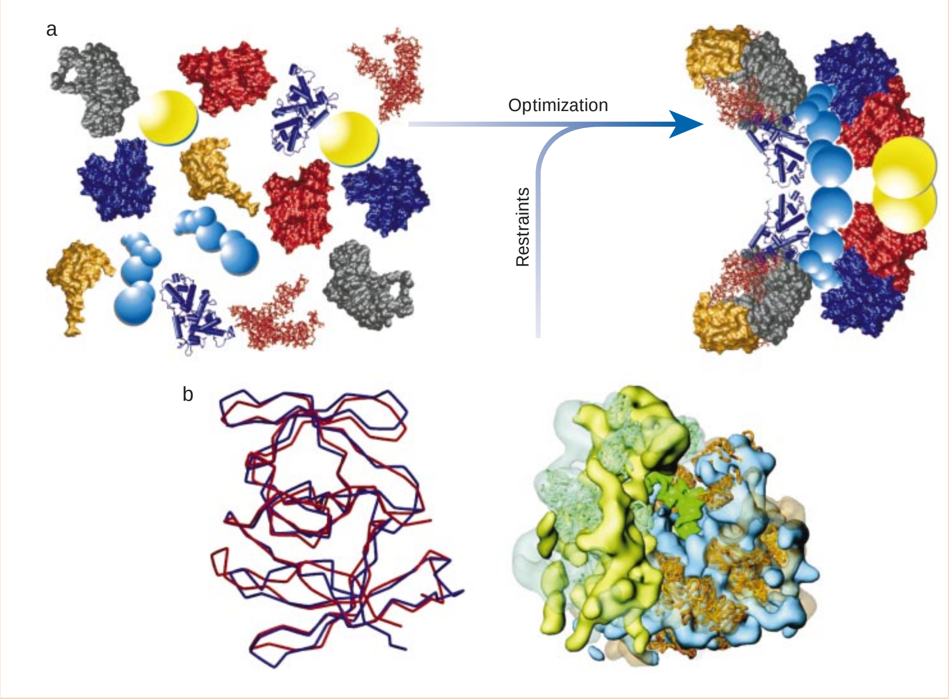


Figure 5 Hybrid approaches to structure determination of macromolecular complexes. **a**, Scheme illustrating the integration of a diverse set of structures varying in reliability and resolution into a hypothetical hybrid assembly structure. **b**, Hybrid assembly of the 80S ribosome from yeast[34]. Superposition of a comparative protein structure model for a domain in protein L2 from *Bacillus stearothermophilus* with the actual structure (1RL2) (left). A partial molecular model of the whole yeast ribosome (right) was calculated by fitting atomic rRNA (not shown) and comparative protein structure models (ribbon representation) into the electron density of the 80S ribosomal particle.

beyond about 7-8-Å resolution[39,40]. Even so, the ability to visualize elements of secondary structure at this resolution makes it easy to fit a previously determined atomic model of protein monomers into the density. The recent docking of the atomic structure of tubulin into the EM density map of a complete microtubule[41] illustrates just how precise this docking can be. This type of docking can then provide accurate images of the protein–protein contacts that lead to the assembly of larger macromolecular machines (Fig. 2).

Because the electron microscope produces images, and not only diffraction intensities, it is possible to determine the positions and relative orientations of randomly distributed, asymmetric macromolecules. The individual images must then be sorted into a large number of distinct classes of views before they can be averaged. This step in the process is illustrated in Fig. 3a, which shows a gallery of 12 different class averages obtained from ice-embedded specimens of *Drosophila melanogaster* tripeptidyl peptidase II (TPP II)[42]. Once a large set of views is in hand, the 3D reconstruction is computed in much the same way as if the average projections had been computed from images of tilted, 2D crystals. As in the example of TPP II that is shown in Fig. 3b, the resulting 3D reconstruction immediately shows how a large, multi-protein complex is assembled from its individual parts. These single particles must be large in size, however, to provide sufficient signal for the alignment at high resolution[43]. In addition, structure determination by single-particle cryo-EM involves far greater amounts of computation than does structure determination based on 2D crystals or particles with very high internal symmetry (Box 1).

Although the capabilities of single-particle cryo-EM are powerful, the method still remains slow compared to other structure-determination technologies, such as X-ray crystallography or NMR spectroscopy. Completion of a structure at the modest resolution of ~2 nm currently may require a month or more for data collection and perhaps another month for data processing. If the goal is to obtain a density map in which features of secondary structure are clearly visible, data collection may extend over several months. A further drawback of cryo-EM is the fact that data collection remains a specialist craft that requires many months, even years of training, before one is able to take full advantage of the high performance of modern electron microscopes.

But it is not necessary for data collection to take as long as it currently does, or to be so dependent on the scientist having a high level of acquired technical expertise. Instead, recording a large number of particle images that are invisible to the human eye on the viewing screen involves blindly following a prescribed sequence of repetitive operations. In principle, such a task is better suited for a computer than a human operator. Indeed, automated implementations of single-particle data-collection operations have recently been published[44,45]. The next frontier where work has already begun includes automation of the steps in which images of individual particles are selected within digitized micrographs and the data are merged into a 3D reconstruction of the particle. In one recent demonstration, for example, data were collected and a 3D reconstruction was obtained for the tobacco mosaic virus particle at a resolution of ~1 nm in a period of less than 24 hours[46]. Further development of automated data collection and analysis promises to reduce the turnaround time for producing 3D density maps of large, macromolecular particles from months or years to days or weeks.

The 3D reconstructions obtained by cryo-EM are likely to be used primarily for docking (that is, assembling) atomic-resolution models of component macromolecules into the 3D densities of intact complexes. When the resolution of the density map is high enough to see helices and regions of β-sheet, the docking can be done precisely and with little ambiguity. At lower resolution, however, the docking must be performed with caution, and researchers continue to develop quantitative criteria that can guide the operation[47–50]. It is therefore fortunate that the throughput of cryo-EM should soon become well matched to the combined throughput of X-ray crystallography and NMR spectroscopy, which are the primary sources of the structures of the individual components. In turn, atomic models of the various assembled components can then be used to interpret each of the recognizable densities that are visualized within whole-cell tomograms.

Hybrid approaches to structure determination

X-ray crystallography may provide high-resolution structures of large complexes, if they can be purified in sufficient quantities and crystallized. Single-particle EM can provide medium-resolution structures (~1 nm) of complexes even if only small amounts of material are available and can tolerate some sample heterogeneity. Even so, these 'direct' methods are surely not capable of characterizing the myriads of stable complexes that exist in a cell. In addition, most of the transient complexes cannot be addressed at all with these approaches. Therefore, there is a great need for hybrid methods where both high throughput and highest possible resolution are achieved by integrating information from different sources. This integration should be performed in an objective manner, such that it is reproducible by any expert.

The hybrid assembly of a complex needs to reflect spatial restraints of varying accuracy and resolution that originate from vastly different experiments and theoretical considerations (Fig. 4, and Boxes 2 and 3). To this end, it is useful to express structure determination as an optimization problem. In this view, 3D models that are consistent with the input information are calculated by optimizing a scoring function. The three components of this approach are: representation of an assembly; a scoring function consisting of individual spatial restraints; and optimization of the scoring function to obtain the models. Figure 5a illustrates how the subunits of a hypothetical complex (left) can be assembled through optimization with respect to restraints from a variety of methods to obtain the final assembly model (right). Each subunit in an assembly can be represented by a set of points that depend on what is known about the subunit. If an experimentally determined structure of a protein is available or a comparative protein structure model can be calculated, each atom can be represented by its own point. If protein domains can be assigned based on biochemical characterization or bioinformatics analysis (for example, by scanning against a sequence database of domains or by prediction of transmembrane spanning domains), a single point represents each domain. Otherwise, a single point can represent the whole subunit.

The most important aspect of the calculation is to accurately capture all of the existing experimental and theoretical information about the structure of a modelled assembly. For example, the shape, density and symmetry of a complex may be derived from EM; upper distance bounds on residues from different subunits may be obtained from X-ray crystallography or NMR spectroscopy and chemical crosslinking; and protein–protein contact restraints may be obtained from immuno-purification with mass spectroscopy and bioinformatics analysis of an alignment of homologous sequences. An 'ensemble' of models that minimize violations of the input restraints can be obtained by optimizing the scoring function, relying on an optimization method such as simulated annealing with molecular dynamics applied in Cartesian space. Because the optimization is likely to be stochastic, a large number of models need to be calculated and assessed. Examples of predicting assembly structures through satisfaction of varied spatial restraints include the *Escherichia coli* 30S ribosomal subunit[51] and the yeast exosome[52].

A sample study that illustrates some of the points made above is the hybrid assembly of the 80S ribosome (Fig. 5b). A partial molecular model of the whole yeast ribosome was calculated by fitting atomic ribosomal RNA and comparative protein structure models into the electron density of the 80S ribosomal particle, obtained by EM at 15-Å resolution[53]. Most of the models for 40 out of the 75 ribosomal proteins were based on approximately 30% sequence identity to their template structures. Typical accuracy of a comparative model in this range of sequence similarity is indicated by a comparison of a model for a domain in protein L2 from *Bacillus stearothermophilus* with the actual structure. The fitting of the subunits into the electron density was made possible by the atomic structures of the whole small and large ribosomal subunits from archaea.

Visualizing complexes using electron tomography

Electron tomography is by no means a new imaging technology, but it has only recently gathered momentum[54,55] (Fig. 6). With the advent of computer-controlled electron microscopes and the automation of elaborate image acquisition procedures, it became possible to obtain molecular-resolution tomograms of structures as large and complex as whole prokaryotic cells or thin eukaryotic cells embedded in amorphous ice[56]. Non-invasive imaging of whole, vitrified cells is where electron tomography can make a unique contribution and will probably have the greatest impact. The emerging picture of the cell is one of a giant supra-molecular assembly; but on the nanoscale, the cytoplasm is mostly an uncharted territory. Just as high-resolution 3D structures of macromolecules provide valuable insights into their working, a better understanding of cellular functions will arise from the ability to visualize macromolecules in an unperturbed cellular context.

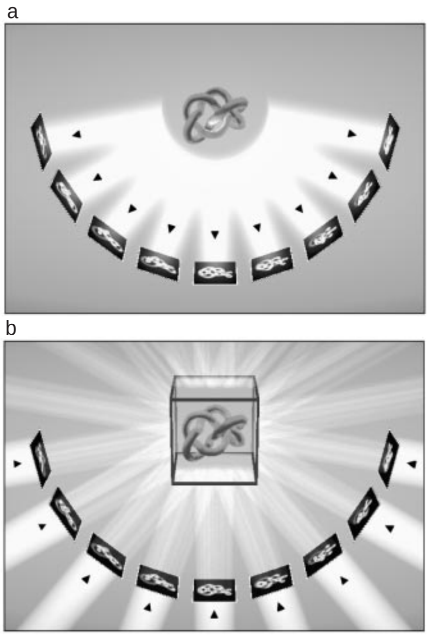


Figure 6 Principle of electron tomography. **a**, Schematic representation of data acquisition. A flexible rope knot represents the object, emphasizing that electron tomography can retrieve 3D information from structures with individual topologies. A set of projection images is recorded on a charge-coupled device camera while the object is tilted incrementally around an axis perpendicular to the electron beam. Owing to the limited accuracy ('eucentricity') of the tilting device, the specimen has to be recentred and refocused at each tilt angle. Automated procedures have been developed to perform this task with negligible exposure of the object to the electron beam[62]. **b**, The back-projection method explains the principle of the 3D reconstruction in an intuitive manner. For each projection, a back-projection body is calculated, and the sum of all projection bodies yields the density distribution of the original object — the tomogram. To compensate for the fact that high-resolution features change more rapidly with tilt angle than do low-resolution features, an appropriate weighting function has to be applied to the data in the 2D images before calculating the reconstruction. The quality of a tomogram depends critically on covering as wide a tilt range as possible (typically ± 70°), with tilt increments as small as possible (1 to 3°). However, each additional exposure to the beam increases the amount of radiation damage and the cumulative dose must not exceed a tolerable limit.

Tomograms of cells at molecular resolution are essentially 3D images of the cell's entire proteome. They reveal information about the spatial relationships of macromolecules in the cytoplasm, the 'interactome'. But exploitation of this information is confronted with two problems. Cryo-tomograms are contaminated by substantial residual noise and distorted by missing data resulting from the restricted tilt range. Moreover, the cytoplasm is very densely

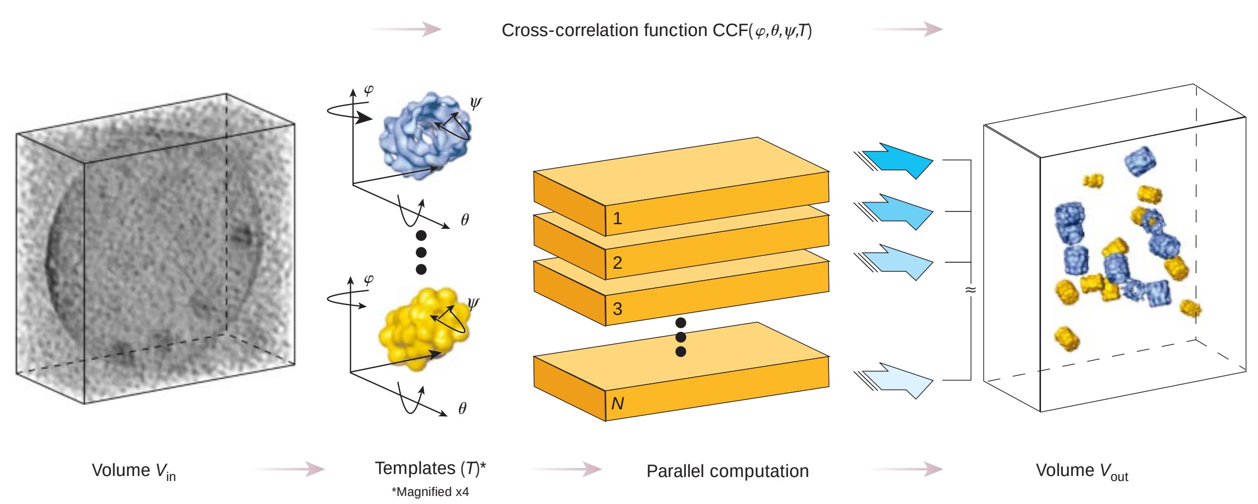


Figure 7 Mapping the spatial distribution of complexes and their interactions within cells. A molecular-resolution tomogram of a cell (V_{in}) is essentially a 3D image of the cell's entire proteome. Residual noise and molecular crowding hamper visualization of the information that is present in this tomogram. As a result, 3D pattern recognition must be used to 'mine' this information. One approach that has been demonstrated to be effective is template matching. Templates of the macromolecular complexes (or even parts thereof) that are of interest must first be obtained by techniques such as cryo-EM or hybrid X-ray/NMR and cryo-EM reconstruction. These templates T (magnified × 4 in this figure) are used to search for matching structures by cross-correlation, and the result is refined by multivariate statistical analysis. Because both the positions of the complexes and their orientations are initially unknown, V_{in} must be scanned for all possible orientations of each of the templates. The result (V_{out}) shows the positions and orientations of the complexes in the cell. In principle, it should be possible to chart the cellular 'interactome' — the spatial relationships of all major complexes of a cell — by this approach.

populated ('molecular crowding'), with molecules literally touching each other[57]. Under these conditions, segmentation and feature extraction based on visual inspection is usually impossible, except for some easily recognizable features, such as membranes and the cytoskeleton.

Nevertheless, pattern-recognition techniques can be used, in one guise or another, to detect and identify specific molecules[58]. Provided that a high- or medium-resolution structure of the molecule of interest is available, it can be used as a template to perform a systematic search of the reconstructed volume for matching structures (Fig. 7). Such a molecular signature-based approach, while computationally demanding, can be efficiently parallelized. Once the spatial coordinates of a complex in a cell have been determined, sub-tomograms that encompass the complex and its neighbourhood can be extracted for further analysis and averaging. Multivariate statistical analysis of such sub-tomograms can be used to explore variations in their functional environment[59].

The feasibility of template matching has been demonstrated with 'phantom cells' (lipid vesicles filled with macromolecules), which provide a realistic experimental scenario and facilitate an assessment of the fidelity of the approach. With the current (non-isotropic) resolution of 4–5 nm, one can address only larger ($M_r > 400$K) complexes in a cellular context. To widen the scope of cellular tomography, it will be necessary to improve the resolution. Theoretical considerations[60] and ongoing instrumental improvements (such as liquid helium versus liquid nitrogen temperature, improved detectors and dual-axis tilting) make a resolution near 2 nm a realistic goal[61].

Perspectives

The possibility seems now assured of assembling a structural picture that can be 'zoomed' continuously from the details of atomic models all the way up to the full complexity of an intact cell. Structural genomics will bring us closer to a comprehensive dictionary of proteins in the foreseeable future, while EM techniques and hybrid approaches will allow us to assemble proteins as words into meaningful sentences. A comprehensive description of large complexes will generally require the use of a number of experimental models (Box 2), underpinned by a variety of theoretical approaches (Box 3) to maximize efficiency, completeness, accuracy and resolution of the experimental determination of assembly composition and structure. In conjunction with the non-invasive 3D imaging of whole cells, these approaches might ultimately enable us to read the molecular book of the cell. □

doi:10.1038/nature01513

1. Alberts, B. The cell as a collection of protein machines — preparing the next generation of molecular biologists. *Cell* **92,** 291–294 (1998).
2. Baumeister, W. & Steven, A. C. Macromolecular electron microscopy in the era of structural genomics. *Trends Biochem. Sci.* **25,** 624–631 (2000).
3. Sali, A. & Kuriyan, J. Challenges at the frontiers of structural biology. *Trends Biochem. Sci.* **24,** M20–M24 (1999).
4. Orengo, C. A. *et al.* The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics* **2,** 11–21 (2002).
5. Govindarajan, S., Recabarren, R. & Goldstein, R. A. Estimating the total number of protein folds. *Proteins* **35,** 408–414 (1999).
6. Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O. & Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402,** 83–86 (1999).
7. Mewes, H. W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30,** 31–34 (2002).
8. Costanzo, M. C. *et al.* YPD, PombePD and WormPD: model org anism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res.* **29,** 75–79 (2001).
9. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417,** 399–403 (2002).
10. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98,** 4569–4574 (2001).
11. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403,** 623–627 (2000).
12. Aloy, P. & Russell, R. B. Potential artefacts in protein-interaction networks. *FEBS Lett.* **530,** 253–254 (2002).
13. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415,** 141–147 (2002).
14. Aloy, P. & Russell, R. B. The third dimension for protein interactions and complexes. *Trends Biochem. Sci.* **27,** 633–638 (2002).
15. Jansen, R., Greenbaum, D. & Gerstein, M. Relating whole-genome expression data with protein-protein interactions. *Genome Res.* **2,** 37–46 (2002).
16. Ge, H., Liu, Z., Church, G. M. & Vidal, M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genet.* **4,** 482–486 (2001).
17. Edwards, A. M. *et al.* Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.* **10,** 529–536 (2002).

18. Kumar, A. & Snyder, M. Protein complexes take the bait. *Nature* **415,** 123–124 (2002).
19. Abbott, A. The society of proteins. *Nature* **417,** 894–896 (2002).
20. Westbrook, J. *et al.* The Protein Data Bank: unifying the archive. *Nucleic Acids Res.* **30,** 245–248 (2002).
21. Cramer, P., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: RNA polymerase II at 2.8 Ångstrom resolution. *Science* **292,** 1863–1876 (2001).
22. 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).
23. Harms, J. *et al.* High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. *Cell* **107,** 679–688 (2001).
24. Wimberly, B. T. *et al.* Structure of the 30S ribosomal subunit. *Nature* **407,** 327–339 (2000).
25. Yusupov, M. M. *et al.* Crystal structure of the ribosome at 5.5 Å resolution. *Science* **292,** 883–896 (2001).
26. Abola, E., Kuhn, P., Earnest, T. & Stevens, R. C. Automation of X-ray crystallography. *Nature Struct. Biol.* **7,** 973–977 (2000).
27. Snell, G. *et al.* Automatic sample mounting and alignment system for biological crystallography. *J. Synchrotron Radiat.* (in the press).
28. Burley, S. K. *et al.* Structural genomics: beyond the Human Genome Project. *Nature Genet.* **23,** 151–157 (1999).
29. Vitkup, D., Melamud, E., Moult, J. & Sander, C. Completeness in structural genomics. *Nature Struct. Biol.* **8,** 559–566 (2001).
30. Structural genomics. *Nature Struct. Biol.* **7**(Suppl.), 927–994 (2000).
31. Frank, J. *Three-dimensional Electron Microscopy of Macromolecular Assemblies* (Academic, London, 1996).
32. Henderson, R., Baldwin, J. M. & Ceska, T. A. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.* **213,** 899–929 (1990).
33. Kuhlbrandt, W., Wang, D. N. & Fujiyoshi, Y. Atomic model of plant light-harvesting complex by electron crystallography. *Nature* **367,** 614–621 (1994).
34. Grigorieff, N., Ceska, T. A., Downing, K. H., Baldwin, J. M. & Henderson, R. Electron-crystallographic refinement of the structure of bacteriorhodopsin. *J. Mol. Biol.* **259,** 393–421 (1996).
35. Nogales, E., Wolf, S. G. & Downing, K. H. Structure of the αβ tubulin dimer by electron crystallography. *Nature* **391,** 199–203 (1998).
36. Mitsuoka, K. *et al.* The structure of bacteriorhodopsin at 3.0 Å resolution based on electron crystallography: implication of the charge distribution. *J. Mol. Biol.* **286,** 861–882 (1999).
37. Murata, K. *et al.* Structural determinants of water permeation through aquaporin-1. *Nature* **407,** 599–605 (2000).
38. Lowe, J., Li, H., Downing, K. H. & Nogales, E. Refined structure of αβ-tubulin at 3.5 Å resolution. *J. Mol. Biol.* **313,** 1045–1057 (2001).
39. Conway, J. F. *et al.* Visualization of a 4-helix bundle in the hepatitis B virus capsid by cryo-electron microscopy. *Nature* **386,** 91–94 (1997).
40. Bottcher, B., Wynne, S. A. & Crowther, R. A. Determination of the fold of the core protein of hepatitis B virus by cryo-electron microscopy. *Nature* **386,** 88–91 (1997).
41. Li, H. L., DeRosier, D. J., Nicholson, W. V., Nogales, E. & Downing, K. H. Microtubule structure at 8 Å resolution. *Structure* **10,** 1317–1328 (2002).
42. Rockel, B., Peters, J., Kuhlmorgen, B., Glaeser, R. M. & Baumeister, W. A giant protease with a twist: the TPP II complex from *Drosophila* studied by electron microscopy. *EMBO J.* **21,** 5979–5984 (2002).
43. Henderson, R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Q. Rev. Biophys.* **28,** 171–193 (1995).
44. Carragher, B. *et al.* Leginon: an automated system for acquisition of images from vitreous ice specimens. *J. Struct. Biol* **132,** 33–45 (2000).
45. Zhang, P. J., Beatty, A., Milne, J. L. S. & Subramaniam, S. Automated data collection with a Tecnai 12 electron microscope: applications for molecular imaging by cryomicroscopy. *J. Struct. Biol.* **135,** 251–261 (2001).
46. Zhu, Y. X., Carragher, B., Kriegman, D. J., Milligan, R. A. & Potter, C. S. Automated identification of filaments in cryoelectron microscopy images. *J. Struct. Biol.* **135,** 302–312 (2001).
47. Rossmann, M. G., Bernal, R. & Pletnev, S. V. Combining electron microscopic with X-ray crystallographic structures. *J. Struct. Biol.* **136,** 190–200 (2001).
48. Wriggers, W. & Birmanns, S. Using *Situs* for flexible and rigid-body fitting of multiresolution single-molecule data. *J. Struct. Biol.* **133,** 193–202 (2001).
49. Volkmann, N. & Hanein, D. Quantitative fitting of atomic models into observed densities derived by electron microscopy. *J. Struct. Biol.* **125,** 176–184 (1999).
50. Chacon, P. & Wriggers, W. Multi-resolution contour-based fitting of macromolecular structures. *J. Mol. Biol.* **317,** 375–384 (2002).
51. Malhotra, A., Tan, R. K. & Harvey, S. C. Prediction of the three-dimensional structure of *Escherichia coli* 30S ribosomal subunit: a molecular mechanics approach. *Proc. Natl Acad. Sci. USA* **87,** 1950–1954 (1990).
52. Aloy, P. *et al.* A complex prediction: three-dimensional model of the yeast exosome. *EMBO Rep.* **3,** 628–635 (2002).
53. Spahn, C. M. *et al.* Structure of the 80S ribosome from *Saccharomyces cerevisiae*–tRNA-ribosome and subunit-subunit interactions. *Cell* **107,** 373–386 (2001).
54. Baumeister, W. Electron tomography: towards visualizing the molecular organization of the cytoplasm. *Curr. Opin. Struct. Biol.* **12,** 679–684 (2002).
55. Baumeister, W., Grimm, R. & Walz, J. Electron tomography of molecules and cells. *Trends Cell Biol.* **9,** 81–85 (1999).
56. Medalia, O. *et al.* Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. *Science* **298,** 1209–1213 (2002).
57. Grunewald, K., Medalia, O., Gross, A., Steven, A. & Baumeister, W. Prospects of electron cryotomography to visualize macromolecular complexes inside cellular compartments: implications of crowding. *Biophys. Chem.* (in press).
58. Bohm, J. *et al.* Toward detecting and identifying macromolecules in a cellular context: template matching applied to electron tomograms. *Proc. Natl Acad. Sci. USA* **97,** 14245–14250 (2000).
59. Frangakis, A. S. *et al.* Identification of macromolecular complexes in electron cryotomograms of phantom cells. *Proc. Natl Acad. Sci. USA* **99,** 14153–14158 (2002).
60. Grimm, R. *et al.* Electron tomography of ice-embedded prokaryotic cells. *Biophys. J.* **74,** 1031–1042 (1998).
61. Plitzko, J. *et al. In vivo veritas*: electron cryotomography of cells. *Trends Biotechnol.* **20,** S40–S44 (2002).
62. Koster, A. J. *et al.* Perspectives of molecular and cellular electron tomography. *J. Struct. Biol.* **120,** 276–308 (1997).
63. Glaeser, R. M. Electron crystallography: present excitement, a nod to the past, anticipating the future. *J. Struct. Biol.* **128,** 3–14 (1999).
64. Zhang, G. Y. *et al.* Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution. *Cell* **98,** 811–824 (1999).
65. Fiaux, J., Bertelsen, E. B., Horwich, A. L. & Wuthrich, K. NMR analysis of a 900K GroEL–GroES complex. *Nature* **418,** 207–211 (2002).
66. Yee, A. *et al.* An NMR approach to structural proteomics. *Proc. Natl Acad. Sci. USA* **99,** 1825–1830 (2002).
67. Fushman, D., Xu, R. & Cowburn, D. Direct determination of changes of interdomain orientation on ligation: use of the orientational dependence of ^{15}N NMR relaxation in Abl SH(32). *Biochemistry* **38,** 10225–10230 (1999).
68. Nakanishi, T. *et al.* Determination of the interface of a large protein complex by transferred cross-saturation measurements. *J. Mol. Biol.* **318,** 245–249 (2002).
69. Pellecchia, M., Sem, D. S. & Wuthrich, K. NMR in drug discovery. *Nature Rev. Drug Discov.* **1,** 211–219 (2002).
70. Frank, J. Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu. Rev. Biophys. Biomol. Struct.* **31,** 303–319 (2002).
71. Volkmann, N. A novel three-dimensional variant of the watershed transform for segmentation of electron density maps. *J. Struct. Biol.* **138,** 123–129 (2002).
72. Rout, M. P. *et al.* The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* **148,** 635–651 (2000).
73. Rappsilber, J., Siniossoglou, S., Hurt, E. C. & Mann, M. A generic strategy to analyze the spatial organization of multi-protein complexes by cross-linking and mass spectrometry. *Anal. Chem.* **72,** 267–275 (2000).
74. Young, M. M. *et al.* High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. *Proc. Natl Acad. Sci. USA* **97,** 5802–5806 (2000).
75. Neubauer, G. *et al.* Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc. Natl Acad. Sci. USA* **94,** 385–390 (1997).
76. Neubauer, G. *et al.* Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nature Genet.* **20,** 46–50 (1998).
77. Houry, W. A., Frishman, D., Eckerskorn, C., Lottspeich, F. & Hartl, F. U. Identification of *in vivo* substrates of the chaperonin GroEL. *Nature* **402,** 147–154 (1999).
78. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415,** 180–183 (2002).
79. Miras, I., Schaeffer, F., Beguin, P. & Alzari, P. M. Mapping by site-directed mutagenesis of the region responsible for cohesin-dockerin interaction on the surface of the seventh cohesin domain of *Clostridium thermocellum* CipA. *Biochemistry* **41,** 2115–2119 (2002).
80. Wells, J. A. Systematic mutational analyses of protein-protein interfaces. *Methods Enzymol.* **202,** 390–411 (1991).
81. Jin, L., Cohen, F. E. & Wells, J. A. Structure from function: screening structural models with functional data. *Proc. Natl Acad. Sci. USA* **91,** 113–117 (1994).
82. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470 (1995).
83. Lockhart, D. J. & Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **405,** 827–836 (2000).
84. Baker, D. & Sali, A. Protein structure prediction and structural genomics. *Science* **294,** 93–96 (2001).
85. Bonneau, R. & Baker, D. *Ab initio* protein structure prediction: progress and prospects. *Annu. Rev. Biophys. Biomol. Struct.* **30,** 173–189 (2001).
86. Bonneau, R. *et al. De novo* prediction of three-dimensional structures for major protein families. *J. Mol. Biol.* **322,** 65–78 (2002).
87. Marti-Renom, M. A. *et al.* Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 291–325 (2000).
88. Domingues, F. S., Lackner, P., Andreeva, A. & Sippl, M. J. Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J. Mol. Biol* **297,** 1003–1013 (2000).
89. Pieper, U., Eswar, N., Stuart, A. C., Ilyin, V. A. & Sali, A. MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.* **30,** 255–259 (2002).
90. Smith, G. R. & Sternberg, M. J. E. Prediction of protein-protein interactions by docking methods. *Curr. Opin. Struct. Biol.* **12,** 28–35 (2002).
91. Strynadka, N. C. J. *et al.* Molecular docking programs successfully predict the binding of a β-lactamase inhibitory protein to TEM-1 β-lactamase. *Nature Struct. Biol.* **3,** 233–239 (1996).
92. Enright, A. J., Iliopoulos, I., Kyrpides, N. C. & Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402,** 86–90 (1999).
93. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA* **96,** 2896–2901 (1999).
94. Goh, C. S., Bogan, A. A., Joachimiak, M., Walther, D. & Cohen, F. E. Co-evolution of proteins with their interaction partners. *J. Mol. Biol.* **299,** 283–293 (2000).
95. Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.* **14,** 609–614 (2001).
96. Pazos, F. & Valencia, A. *In silico* two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **47,** 219–227 (2002).
97. Lichtarge, O., Bourne, H. R. & Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257,** 342–358 (1996).
98. Lappe, M., Park, J., Niggemann, O. & Holm, L. Generating protein interaction maps from incomplete data: application to fold assignment. *Bioinformatics* **17,** S149–S156 (2001).
99. Aloy, P. & Russell, R. B. Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA* **99,** 5896–5901 (2002).

Acknowledgements We thank N. Eswar for preparing the histograms in Fig. 1, M. Simon, B. Jap and H. Noller for permission to use the structural images in Fig. 1, K. H. Downing for Fig. 2, B. Rockel for Fig. 3, and F. Alber for Figs 4 and 5a. We are also grateful to P. Bjorkman and H. Moss for commenting on the manuscript. This work has been supported in part by NIH grants (to A.S., R.M.G. and T.E.), the Agouron Institute (T.E.) and a Max-Planck Research Award (W.B.).