Integrating Diverse Data for Structure Determination of Macromolecular Assemblies

Frank Alber,[1,3] Friedrich Förster,[1] Dmitry Korkin,[1,4] Maya Topf,[2] and Andrej Sali[1]

[1]Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, University of California at San Francisco, California 94158-2330; email: frido@salilab.org; sali@salilab.org

[2]School of Crystallography, Birkbeck College, University of London, London WC1E 7HX, United Kingdom; email: m.topf@mail.cryst.bbk.ac.uk

[3]Present addresses: Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, Los Angeles, California 90089-2910; email: alber@usc.edu

[4]Informatics Institute and Department of Computer Science, University of Missouri at Columbia, Missouri 65211; email: korkin@korkinlab.org

Annu. Rev. Biochem. 2008. 77:443–77

First published online as a Review in Advance on March 4, 2008

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

This article's doi: 10.1146/annurev.biochem.77.060407.135530

0066-4154/08/0707-0443$20.00

Key Words

architecture, assembly, complex, configuration, hybrid methods, restraints

Abstract

To understand the cell, we need to determine the macromolecular assembly structures, which may consist of tens to hundreds of components. First, we review the varied experimental data that characterize the assemblies at several levels of resolution. We then describe computational methods for generating the structures using these data. To maximize completeness, resolution, accuracy, precision, and efficiency of the structure determination, a computational approach is required that uses spatial information from a variety of experimental methods. We propose such an approach, defined by its three main components: a hierarchical representation of the assembly, a scoring function consisting of spatial restraints derived from experimental data, and an optimization method that generates structures consistent with the data. This approach is illustrated by determining the configuration of the 456 proteins in the nuclear pore complex (NPC) from baker's yeast. With these tools, we are poised to integrate structural information gathered at multiple levels of the biological hierarchy—from atoms to cells—into a common framework.

Contents

NPC: nuclear pore complex

Configuration: component positions and the presence of interactions

Resolution: for a density map, resolution is the minimum distance between two points at which they can be distinguished

INTRODUCTION

Assemblies as Functional Modules of the Cell

Macromolecular assemblies consist of noncovalently interacting macromolecular components, such as proteins and nucleic acids. They vary widely in size and play crucial roles in most cellular processes (1). Many assemblies are composed of tens and even hundreds of individual components. For example, the nuclear pore complex (NPC) of ~456 proteins regulates macromolecular transport across the nuclear envelope (NE); the ribosome consists of ~80 proteins and ~15 RNA molecules and is responsible for protein biosynthesis.

Need for assembly structures. A comprehensive characterization of the structures and dynamics of biological assemblies is essential for a mechanistic understanding of the cell (2–5). Even a coarse characterization of the configuration of macromolecular components in a complex (**Figure 1**) helps to elucidate the principles that underlie cellular processes, in addition to providing a necessary starting point for a higher-resolution description.

Scope. Complete lists of the macromolecular components of biological systems are becoming available (6). However, the identification of complexes between these components is a nontrivial task. This difficulty arises partly from the multitude of component types and the varying life spans of the complexes (7). The most comprehensive information about binary protein interactions is available for the *Saccharomyces cerevisiae* proteome, consisting of ~6200 proteins. This data has been generated by methods such as the yeast two-hybrid system (8, 9) and affinity purifications coupled with mass spectrometry (10–12). The lower bound on binary protein interactions in yeast has been estimated to be ~30,000 (7), corresponding to the average of ~9 protein partners per protein, although not necessarily all at the same time. The number of higher-order complexes in yeast is estimated to be ~800 on the basis of affinity purification experiments (10–13). 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 in a few hundred key cellular processes whose stable structures and

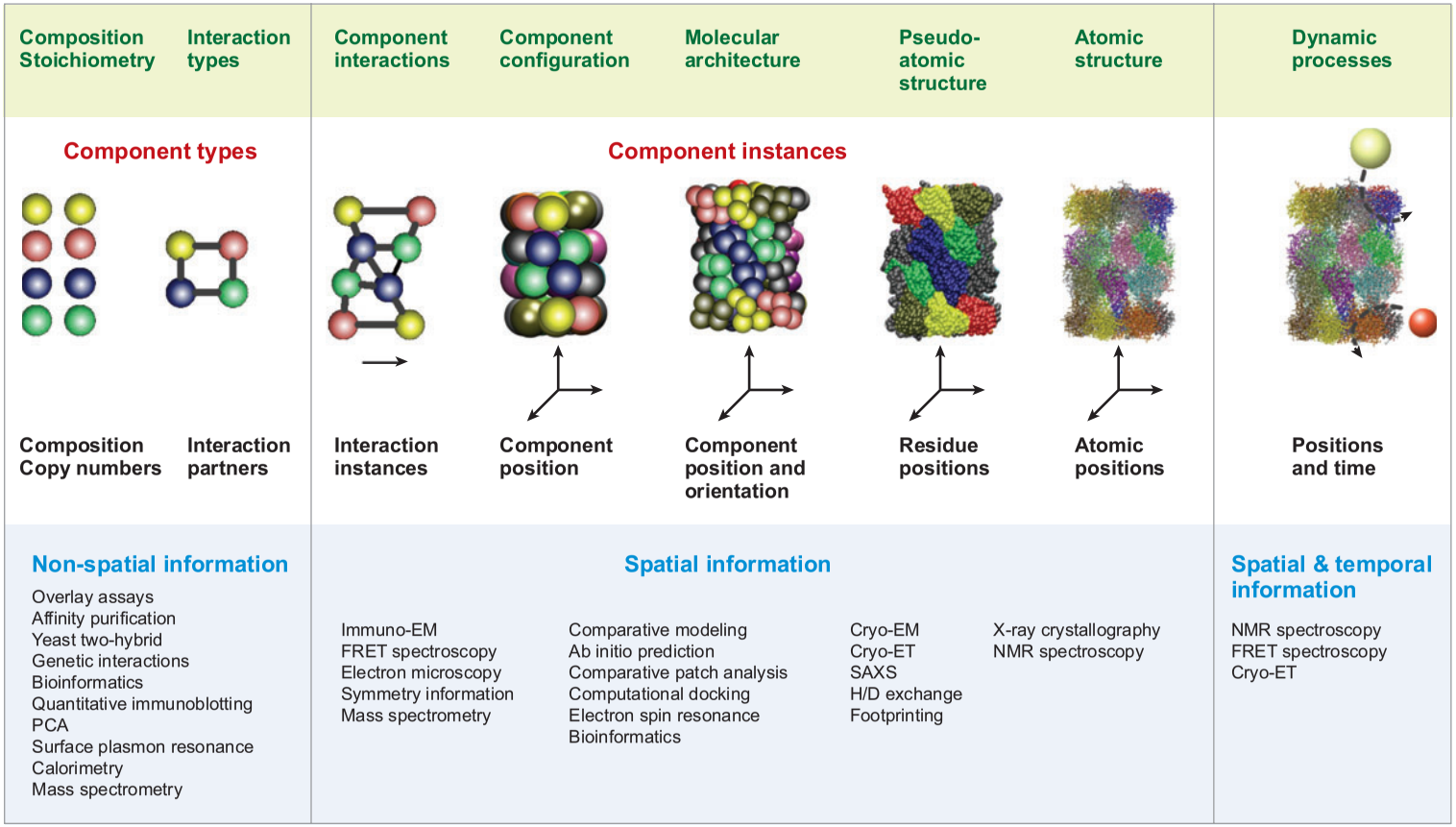


Figure 1

Structural information about an assembly. Varied experimental methods can determine the copy numbers (stoichiometry) and types (composition) of the components (whether or not components interact with each other), positions of the components, and their relative orientations. Importantly, some methods identify only component types and do not distinguish between different instances of a component of the same type when more than one copy of it is present in the assembly. Other methods do identify specific instances of a component. Integration of data from varied methods generally increases the accuracy, efficiency, and coverage of structure determination. Abbreviations: Cryo-ET, cryo-electron tomography; EM, electron microscopy; FRET, fluorescence resonance energy transfer; H/D, hydrogen/deuterium; PCA, protein-fragment complementation assay; SAXS, small-angle X-ray scattering.

transient interactions are yet to be characterized (1, 14).

Difficulties. Compared to structure determination of the individual components, however, structural characterization of macromolecular assemblies is usually more difficult and represents a major challenge in structural biology (2, 3). For example, X-ray crystallography is limited by the difficulties of growing suitable crystals and building molecular models into large unit cells; nuclear magnetic resonance (NMR) spectroscopy is limited by size; electron microscopy (EM), affinity purification, yeast two-hybrid system, calorimetry, footprinting, chemical cross-linking, small-angle X-ray scattering (SAXS), and fluorescence resonance energy transfer (FRET) spectroscopy are limited by low resolution of the corresponding structural information; and computational protein structure modeling and docking are limited by low accuracy.

Integrative approach. These shortcomings can be minimized by simultaneous consideration of all available information about a given assembly (**Figure 1**) (3, 15–17). This information may vary greatly in terms of its accuracy and precision and includes data from both experimental methods and theoretical considerations, such as those listed above. The integration of structural information about an assembly from various sources can only be achieved by computational means. In this review, we focus on the computational aspects of data integration.

Chapter outline. We begin by reviewing the types of spatial information generated by experimental and computational methods that have allowed structural biology to shift its focus from individual proteins to large assemblies. Such data include atomic and residue positions from X-ray crystallography and NMR spectroscopy, shape and density for an assembly from EM and SAXS, as well as component proximities and interactions from proteomics methods, mass spectrometry, and labeling techniques.

Next, we review computational methods that generate models of assembly structures on the basis of given information. In particular, we focus on advances in computational methods for comparative modeling of complexes and for docking atomic structures of proteins to other proteins. We also review methods for the fitting of component structures into density maps, typically determined by cryo-EM and cryo-electron tomography (cryo-ET). In addition, we outline computational methods that use data from SAXS experiments in solution.

Finally, we offer a perspective on generating macromolecular assemblies that are consistent with all available information from experimental methods, physical theories, and statistical preferences extracted from biological databases. Such an integrative system, in principle, achieves higher completeness, resolution, accuracy, precision, and efficiency than a structure characterization using any of the individual types of data alone (3, 18). We illustrate this approach by its application to the determination of the configuration of 456 proteins in the yeast NPC (18, 19).

SOURCES OF SPATIAL INFORMATION

Various experimental methods produce different types of structural information (**Figure 1**). This information differs in terms of the spatial features it restrains as well as in resolution, accuracy, and quantity. The stoichiometry and composition of protein components in an assembly can be determined by methods such as quantitative immunoblotting and mass spectrometry. The positions of the components can be elucidated by cryo-EM and labeling techniques. Whether or not components interact with each other can be measured by the yeast two-hybrid system and affinity purification. Relative orientations of components and information about interacting residues can be inferred from cryo-EM, hydrogen/deuterium

Electron microscopy (EM): includes single-particle EM, electron tomography, and electron crystallography and provides spatial maps of assemblies at pseudoatomic or lower resolutions

SAXS: small-angle X-ray scattering

FRET: fluorescence resonance energy transfer

Accuracy: difference between the determined structure and the actual native structure

Precision: variability among structural solutions consistent with the data

ET: electron tomography

(H/D) exchange, OH radical footprinting, and chemical cross-linking. At the highest resolution, information about the atomic structures of components and their interactions can be determined by X-ray crystallography and NMR spectroscopy. Importantly, some methods do not distinguish between different instances of a component of the same type, resulting in ambiguity when more than one copy of the component is present in the assembly (e.g., proteomics methods, including the yeast two-hybrid system and affinity purification). Structures can be described at different levels of resolution, including the component configuration (specifying component positions and the presence of interactions), the molecular architecture (specifying the components' configuration and relative orientations), pseudoatomic models (specifying atomic positions with errors larger than the size of an atom), and atomic structures (specifying atomic positions with precision smaller than the size of an atom).

Molecular architecture: components' configuration and relative orientations

Pseudoatomic models: atomic positions with errors larger than the size of an atom

X-ray Crystallography and NMR Spectroscopy

X-ray crystallography has been the most prolific technique for the structural analysis of proteins and protein complexes and is still the "gold standard" in terms of accuracy and resolution. X-ray crystallography measures the structure factor amplitudes and approximate phases for a crystal sample. Together with a molecular mechanics force field, this information is used in an optimization process that can result in an atomic structure of the assembly (20, 21).

NMR spectroscopy allows determination of atomic structures of increasingly large subunits and even complexes in solution under near-native conditions (22, 23). Data from NMR spectroscopy include upper distance bounds between pairs of atoms and dihedral angle values of certain groups of atoms. In combination with a molecular mechanics force field, this information can result in an atomic structure of the protein through an optimization process (24). NMR spectroscopy is also increasingly used to determine the interacting surfaces of protein components in complexes from chemical shift perturbations (25) and residue dipolar coupling (26). Such information can be combined with computational docking to obtain approximate structures of protein complexes (see Protein-Protein Docking below). It is particularly useful that NMR spectroscopy methods can be applied to weak and transient protein complexes, which are difficult to study by other structural methods (27, 28). For instance, transient encounter complexes in protein-protein associations can be visualized (29). Moreover, in-cell NMR spectroscopy provides a means of analyzing the structural changes that accompany protein interactions in vivo and at atomic resolution (30).

Number of structures. The number of structures of macromolecular assemblies solved by X-ray crystallography or NMR spectroscopy is still relatively small. In the Protein Data Bank (PDB) (31), there are approximately 5000 binary interfaces with less than 30% sequence identity to each other. It will likely be many years before we have a complete repertoire of high-resolution structures for the hundreds of binary and higher-order complexes in a typical cell. This discrepancy is due mainly to the difficult production of sufficient quantities of the sample and its crystallization.

Utility of atomic structures. Atomic structures of protein complexes provide templates for the comparative modeling (32) of protein complexes (see Comparative modeling, below). They are also used to derive statistical potentials of mean force (33, 34) that are useful for the generation and assessment of protein complexes by computational docking (see Protein-Protein Docking, below). Several attempts have been made to classify protein complexes (35) and protein-protein interfaces (33) on the basis of their atomic structures, providing a basis for analysis of the physical

principles, function, and evolution of interactions between proteins.

Atomic models: atomic positions with precision smaller than the size of an atom

Electron Microscopy

The different variants of EM are electron crystallography (36, 37), single-particle EM (38, 39), and ET (40–42).

Electron crystallography. Electron crystallography requires macromolecules to be arranged in two-dimensional crystals (typically for membrane proteins) (36, 37) or helical fibers (often for proteins involved in filaments) (43). The resolution obtained in electron crystallography is frequently sufficient to trace the protein backbone (<4.5 Å) or at the very least to obtain pseudoatomic models by fitting component atomic structures into the map (~5–10 Å) (44). However, the technique is not used often owing to difficulties in obtaining periodic arrays and to high technical demands.

Single-particle EM. Single-particle EM can be applied to a dried and heavy-metal-stained sample (negative stain EM) or to a hydrated and frozen sample (cryo-EM). Although negative stain EM can only determine the envelope of an assembly, cryo-EM also determines the whole electron-optical density distribution (38). Imaging by single-particle EM requires neither large quantities of the sample nor the sample in a crystalline form (38, 39, 45). Single-particle EM is a powerful tool for the investigation of macromolecular assembly structures that exist in different conformational states (46) or for those whose X-ray structure determination is difficult. An assembly typically needs to weigh above ~250 kDa. Typical resolutions are currently in the intermediate range (5–15 Å) (47). The cryo-EM density maps are particularly useful when combined with atomic-resolution structures of the components, as reviewed in the section Structure Characterization from Density Maps, below (48). For example, fitting of atomic structures and models of proteins and nucleic acids into cryo-EM maps has resulted in quasi-atomic models of viral subunit assemblies (48, 49), ribosomes and ribosome-interacting proteins (46), and various other assemblies (50). The number of single-particle reconstructions deposited in the Electron Microscopy Database is 446 (January 3, 2008) (51), indicating that single-particle EM is increasingly becoming a standard method in structural biology.

Cryo-ET. Cryo-ET can be used to obtain three-dimensional structures of pleomorphic objects such as whole cells (40–42). The tremendous potential of cryo-ET lies in visualizing assemblies in an unperturbed cellular context (52). Prokaryotic and thin eukaryotic cells can be imaged in toto, and recent advances in sectioning of vitrified samples make it possible to gain insights into thicker cells, such as tissue cells (53). The current achievable resolution is in the range of 5–10 nm (54). For some macromolecules, higher resolutions can be obtained by averaging putatively identical particles. Cryo-ET is particularly attractive for studying membrane-bound complexes (55). Images of retroviral envelope protein complexes in situ were obtained at approximately 30-Å resolution, where rigid-body fitting is applicable, albeit with low accuracy (56). Lower resolution, but invaluable, insights have been obtained into important assemblies such as the NPC (57–59).

Cryo-ET can potentially characterize transient interactions by imaging them in an unperturbed environment. Prerequisite to an analysis of proximities of macromolecules are methods to systematically determine atlases of stable assemblies. The problem of introducing electron-dense labels noninvasively favors the identification of assemblies on the basis of known structural signatures (i.e., using template matching, see below) (60). Studies using phantom cells (i.e., liposomes with known content) have indicated that this approach to locating large assemblies (M_W > 1 MDa) is feasible (61); recently, the first ribosomal atlas

of the whole *Spiroplasma melliferum* cell was determined (62).

Small-Angle X-Ray Scattering

Small-angle scattering of X-rays (SAXS) and neutrons is another biophysical method that can provide low-resolution information about the shape of an assembly (63). These techniques study purified proteins and complexes in solution. In SAXS, the molecule's rotationally averaged scattering pattern is measured as a function of spatial frequency, typically to 1–3-nm resolution. This spectrum can be readily transformed into a radial distribution function, which is essentially a histogram of all pairwise distances of the atoms in an assembly weighted by their respective atomic numbers. Because of rotational averaging, the information content of a SAXS spectrum is dramatically reduced compared to a diffraction pattern in X-ray crystallography or even a density map from EM. A conservative estimate of the information content is given by the Whittaker-Shannon sampling criterion, which specifies the number of independent sampling points, N, as a function of resolution, r, of the dataset and the diameter, D, of the macromolecule under scrutiny: $N = 2Dr$. For a particle with diameter of 100 Å and a resolution of 20 Å, this criterion yields N equal to 10. In comparison, if we apply this criterion to an EM map at the same resolution, we obtain N_{EM} equal to 4190. Nevertheless, SAXS can provide important shape information about proteins and assemblies in the size range of 50–250 kDa, which are not amenable to cryo-EM and NMR spectroscopy. In addition, the ease of altering solution conditions makes SAXS ideal for studying differences between varied conformational states of the same assembly (64). Examples where atomic quaternary structure models could be obtained using SAXS in conjunction with atomic structures of fragments include the Ras activator son of sevenless (65) and the different nucleotide-bound conformations of the ATPase GspE (66).

Proteomics Methods and Mass Spectrometry

A variety of proteomics methods produce spatial information at relatively low resolution whose use for structure determination is generally not straightforward.

Spatial information from proteomics. Information about the presence of a binary protein interaction can be translated into an upper bound on the distance between the two corresponding components in higher-order complexes. Therefore, even when the details about their binding interface are not available, the distance between the component centroids can still be restrained. Moreover, the protein composition of copurified complexes can reveal the proximity of a group of proteins without revealing the underlying binary interaction network. Nevertheless, such data implies the presence of a minimum number of protein interactions, so that all components are connected to each other. Any predicted assembly structure must be consistent with such connectivity data. It is, in fact, possible to impose an appropriate spatial restraint that enforces such a connectivity condition during an optimization process (see below) (19, 67).

Ambiguity. A distinctive feature of proteomics-based data is that protein interactions cannot be unambiguously assigned to distinct pairs of protein instances if multiple copies of one or both proteins are present in the assembly. This multiplicity always applies if the assembly is built from identical symmetry units. Moreover, it is, in principle, also possible that not all in vitro detected interactions are present in a given complex at the same time. It is therefore not always easy to define what constitutes a unique biologically active complex on the basis of binary protein interaction data alone. To incorporate spatial information from proteomics methods, these spatial ambiguities have to be considered (see below) (19, 67).

Binary interactions from varied methods. Methods for detecting binary protein interactions include the yeast two-hybrid system (68, 69), protein fragment complementation (70), a combination of phage display with other techniques (71), protein arrays (72), calorimetry (73), and solid-phase detection by surface plasmon resonance (74, 75). Physical protein interactions have also been inferred from genetic interactions through reduced activity or lethality of double-knockout mutant yeast strains (76). Because of the relatively low resolution of some of these biochemical characterizations and their relatively high false-positive rates, care is needed in their interpretation. For example, assessing the biochemically derived interaction sets against known structures of complexes identified potential sources of systematic errors in interaction discovery, such as indirect interactions in yeast two-hybrid systems, obstruction of interfaces by molecular labels, and artificial promiscuity in the detected interactions (77, 78).

Higher-order interactions from affinity purification. An affinity purification experiment combines the purification of protein complexes with the identification of their individual components by mass spectrometry (79). The proximity between the identified components is established because they are directly or indirectly associated with the tagged bait protein (10–12). The method can also be applied to characterize protein-DNA (80) and protein-RNA complexes (81). Several different complex purification strategies exist (79).

A particularly powerful variant uses a single fluorescent affinity tag that allows visualization of the target protein in live cells, followed by rapid extraction and detection of interacting macromolecular partners (82). This method determined the localization and specific interactions of viral proteins with host-cell interaction partners at different stages during infection (83). With the advance of fast purification strategies, weak and transient interactions in complexes will also be studied.

A powerful complement of affinity purification is electrospray ionization mass spectrometry (84). This technique separates unique intact complexes and their fragments by their charge-to-mass ratios, which allows the determination of their composition and stoichiometry.

Although each affinity purification experiment on its own is relatively uninformative about the structure of the parent assembly, considerable synergy exists when a set of independent affinity purifications represents assembly fragments that vary in size and overlap in their composition. In fact, a relatively modest set of affinity purifications that partially overlap in their composition can be sufficient for the identification of the configuration of the protein components in an assembly, especially if combined with other independent data from proteomic and labeling methods (67). For instance, the combination of a large set of affinity purification data has contributed to the identification of the protein configuration and interaction map of the NPC (see below) (18, 19). In other studies, the generation of fragment complexes by mass spectrometry has revealed the subunit architectures of the 19S proteasome (85) and other complexes, such as the yeast exosome and the tryptophan RNA-binding attenuation protein (TRAP) complex (84).

Labeling Techniques

Several labeling techniques can be used to determine the approximate positions of protein components in an assembly (86). The idea is to tag the protein component of interest with a probe, which can then be detected by EM or other methods.

Antibody labels in immuno-EM. In immuno-EM, the recognized label is an antibody, which is typically conjugated to nanometer-sized gold beads to enhance the visibility in EM images (86). The gold-labeled antibody binds either to a primary antibody

directed against an epitope in the protein of interest or to a Protein A tag that is fused with the protein (87). Usually, many images of an assembly with the labeled protein are superposed to obtain a distribution of the gold particles for a more accurate positioning of the tagged protein. The localization of proteins using gold-tagged beads is usually limited by the relatively large variability of gold particle positions, typically because of errors in EM image alignment and linker protein flexibility. Nevertheless, the experiment is informative about the assembly structure if the corresponding error bars are still smaller than the dimension of the assembly. For instance, immunogold labeling in combination with transmission EM determined the organization of the NPC components along two principal axes of the nuclear pore (87) as well as the distribution of proteins in the p97-Ufd1-Npl4 complex (88).

Other labels. The choice of labels is not limited to antibodies. For instance, histidine tags can be detected using NiNTA-conjugated gold particles (88), and proteins can also be identified by interacting proteins that are covalently bound to beads of gold (59).

Labeling in single-particle EM. In single-particle EM analysis, the precision is often sufficiently high to detect the label without the use of gold beads. The label is located by comparison of the labeled and unlabeled densities. This approach was used to determine the molecular architectures of numerous complexes, including the CCT chaperonin (89). It also allowed the identification of conformational variations among splicesosomal complexes (90).

Biochemical and Biophysical Methods

There are a variety of biochemical and biophysical methods that can be used to derive information about the relative position as well as the relative orientation of the components in a larger complex. For example, site-directed mutagenesis can identify residues mediating the interaction (91). Various forms of chemical footprinting (92, 93) and hydrogen-deuterium exchange (94) can identify surfaces buried upon complex formation. Proximities of labeled groups on interacting proteins can be detected by chemical cross-linking (95–97) and FRET spectroscopy (98–100); for instance, the protein organization of the yeast spindle pole body was established to a large extent with distances from FRET experiments (98). A method for obtaining long-range distance restraints in protein complexes is pulsed dipolar spin resonance spectroscopy that provides the separation distance of two specifically placed spins within a protein complex. It has been used for the rigid-body refinement of the protein components in complexes (101), such as the *Escherichia coli* chemosome (102).

COMPUTATIONAL METHODS FOR ASSEMBLY STRUCTURE DETERMINATION

The experimental data about a structure, described above, must be converted to an explicit structural model through computation. We now describe such computational methods. We focus on the type of information they use to calculate the assembly structures, rather than on how they calculate them. The methods reviewed in this section use one or two dominant types of information and do not aim to combine explicitly many different types of information.

Template-Based Modeling

Template-based modeling methods rely on known structures of homologous complexes. Such methods include comparative modeling and threading.

Comparative modeling. It may be possible to model a protein assembly using standard comparative modeling techniques (32,

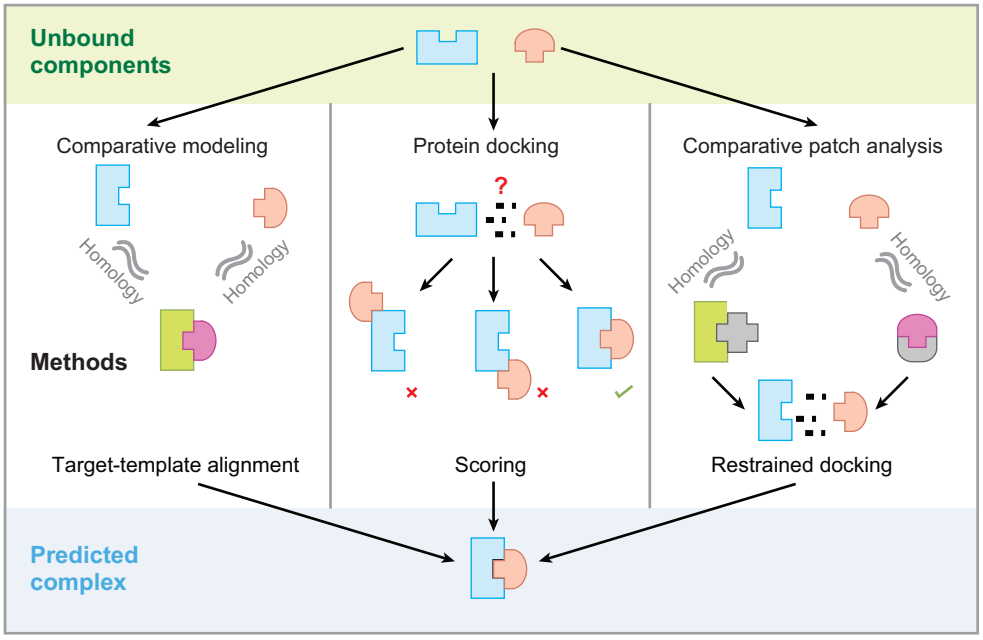


Figure 2

Three computational methods for modeling structures of protein complexes. Comparative modeling builds a model of a complex by using a known structure of a similar complex as a template. Protein docking can be applied when no structure of a similar assembly is known because it relies on searching through possible complex configurations and assessing them by geometrical and physicochemical complementarities. Comparative patch analysis is a hybrid of protein docking and comparative modeling; it restrains docking to only refined interaction modes suggested by structurally defined interactions between each of the complex components, or their homologs, with any other protein (138).

33, 103–105) (**Figure 2**). The requirement is the availability of structural templates that can be reliably aligned to the sequences of the target assembly. Such templates may cover the entire assembly or a sufficiently overlapping set of its fragments. A comparative model of an assembly can be assessed with a variety of different energy and scoring functions, including empirical statistical potentials that are designed to score component interactions and are derived from binary interfaces of known structures (33, 77). Comparative modeling assumes that homologous subunits constituting the target and template assemblies form equivalent interactions (33, 103, 104). Indeed, interaction modes between proteins of the same fold tend to be structurally similar [interaction root-mean-square deviation (iRMSD) $\leq$ 10 Å] when the sequence identity is above ~30% (106). Below this cutoff, the structures of protein interfaces may be different.

Threading. Binary interfaces that are distantly related to template structures can be modeled with MULTIPROSPECTOR, which uses threading of individual protein sequences onto a library of structurally defined interactions (107). The individual sequences are then scored on the basis of how well they fit the proposed folds as well as on the interface between them (107).

Both types of approaches have been applied to study large collections of sequences and interactions (103, 104, 108). The applicability of template-based modeling is limited to protein assemblies whose homologs are currently in the PDB (31) and Protein Quaternary Structure (109) databases. In a recent

study, 3387 binary and 1234 higher-order protein complexes could be predicted for *S. cerevisiae* (33).

Protein-Protein Docking

The structure of a binary protein complex can be predicted by computational protein docking if atomic structures of its components are available from experiments or modeling (110) (**Figure 2**). Unlike template-based modeling, protein docking can also be applied when structures of homologous assemblies are unknown and can thus predict novel binding modes. Protein docking relies on a global search of a large set of possible assembly configurations, maximizing geometrical and physicochemical complementarities between the pair of constituting components (111–115). Although the vast majority of protein docking methods are applied to protein assemblies of two components, an approach of docking multiple components was recently proposed (116). Protein docking methods vary in component representation, scoring of configurations, and optimization protocols.

Rigid and flexible docking. Most protein docking methods treat components as rigid bodies (111, 112, 115, 117). Other methods incorporate side-chain and backbone flexibilities of the component residues (113, 114). However, these methods are usually computationally expensive because the search space of possible assembly configurations is significantly increased.

CAPRI. Docking methods are systematically assessed every two years through blind trials in the Critical Assessment of PRediction of Interactions (CAPRI) (118). At the meeting in 2005, 2 out of the 30 participating groups predicted 8 out of 9 assemblies with acceptable accuracy (118). One group was able to predict four target assemblies with high accuracy (at least 50% of native contacts, ligand backbone RMSD $\leq$ 5 Å, interface backbone RMSD $\leq$1 Å). Even if the docking methods are not sufficiently accurate to predict whether or not two proteins actually interact with each other, they can, in many cases, correctly identify the interacting surfaces between two structurally defined components.

Challenges. The low accuracy of computational protein docking is usually due to the (*a*) conformational differences in the bound and unbound states of assembly subunits, (*b*) limitations in the sampling of relevant configurations, or (*c*) difficulty of discriminating the native-like configurations from the large number of nonnative alternatives (119, 120). As a result, a typical docking method produces an ensemble of candidate solutions, and it is often difficult to select the native-like mode.

Restrained docking. Varied experimental information about component interactions in an assembly can be used to increase the accuracy of protein docking (27). These methods incorporate the additional data either after model building as a filter or during model building to bias the search. The experimental data can provide information at the atomic level, e.g., chemical shift perturbation in NMR spectroscopy (121, 122); residue level, e.g., hydrogen/deuterium exchange (94, 123); site-directed mutagenesis (91); lower-resolution levels, e.g., chemical cross-linking (124, 125); and residue dipolar coupling in NMR spectroscopy (26, 126).

Many protein docking methods use experimental data at the assessment stage to exclude some of the candidate solutions that are inconsistent with the data. Such programs include Hex (127), GRAMM (128), and RosettaDOCK (91) for use in combination with mutagenesis data (91, 127); DOT (94) for use together with hydrogen/deuterium exchange data; BIGGER (121), AUTODOCK (129), and FTDOCK (130), which employ NMR chemical shift perturbation; and others (26, 124).

Another group of methods incorporates the experimental data directly into the scoring function to be optimized or the

optimization protocol. The additional pseudoenergy term penalizes violations of the experimental data (131, 132). For instance, the program HADDOCK can use ambiguous interaction restraints implied by chemical shift perturbations from NMR experiments or mutagenesis data (27, 133); weighted geometric docking uses experimental data to favor certain areas of the subunit surfaces during the rotation-translation scan (134, 135). Another method, TREEDOCK, limits the configurational search using pairs of anchor atoms, which require contact between them (136).

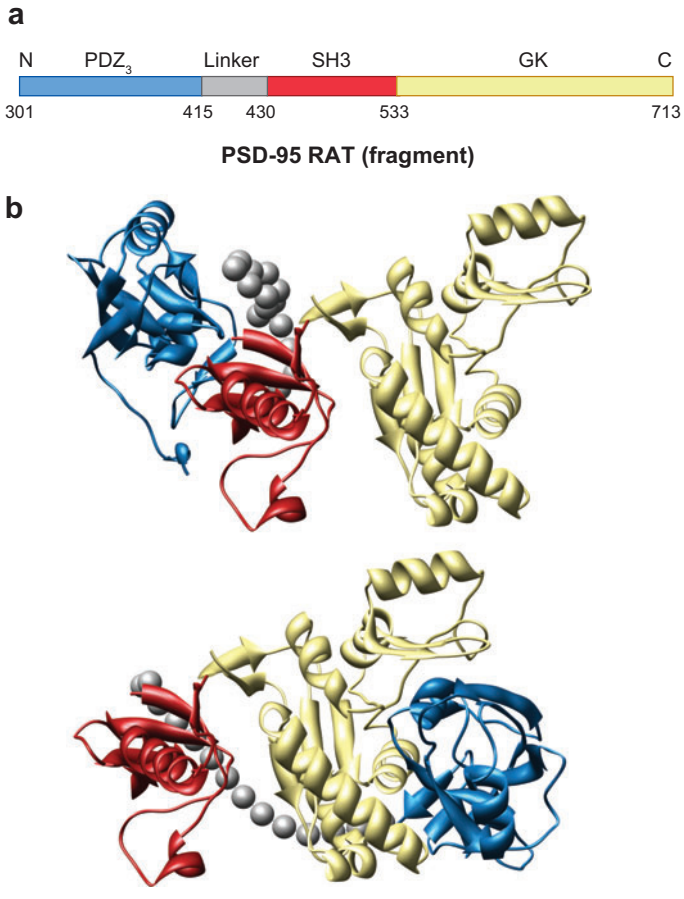


Figure 3

Two predicted binding modes of the core fragment of rat PSD-95 (138). PSD-95 consists of three domains, the PDZ_3 domain (*blue*), the SH3 domain (*red*), and the GK domain (*yellow*). The gray spheres mimic the residues of the interdomain linker between PDZ_3 and SH3. (*a*) The domain architecture of the PSD-95 core fragment. (*b*) The two predicted configurations of PSD-95 (138).

Comparative Patch Analysis

The locations of protein-binding sites on proteins are often conserved in evolution, irrespective of the folds of their binding partners (**Figure 2**) (137). This feature is exploited in comparative patch analysis, which is a hybrid of protein docking and comparative modeling. The method restrains computational docking to binding sites that are conserved within families of homologous domains (**Figure 2**) (138). To determine the conserved binding sites, the following strategy is applied. First, for each subunit in a binary complex, a set of protein-binding sites of its homologs represented in the PIBASE database of structurally defined interfaces is identified (139). Second, these binding sites are mapped onto the partner subunit surface using structure-based alignments between the subunit and each of its homologs. Third, all pairs of the mapped binding sites are used as starting points for restrained docking to obtain candidate models of the binary complex. This ensemble of models is then ranked using a measure of geometric complementarity and a statistical potential score. Comparative patch analysis has a greater applicability than comparative modeling and a higher accuracy than protein docking (138).

Application. Comparative patch analysis was used to model the tertiary structure of the core fragment of rat PSD-95 (**Figure 3**) (138). PSD-95 is a key protein in the postsynaptic density that serves as a structural scaffold for other signaling proteins. Although the structures of its five individual domains have been solved, the complete structure of PSD-95 has not been determined. In addition, structures of neither the PDZ-SH3 nor the PDZ-GK homologs are available, rendering comparative modeling inapplicable. Moreover, computational protein docking

results are ambiguous, generating an ensemble of complexes without any predominant binding modes. By contrast, each of the subunit families is known to repeatedly utilize a small number of binding sites for different protein interactions, indicating that comparative patch analysis may be useful. Comparative patch analysis of the PSD-95 core fragment suggests two alternate configurations, which potentially correspond to the different functional forms of PSD-95 (**Figure 3**). Thus, this finding provides a possible structural explanation for the experimentally observed cooperative folding transitions in PSD-95 and its homologs.

Comparative modeling, protein docking, and comparative patch analysis can benefit from including not only local information on protein-protein interfaces, but also global information about the overall shape of an assembly, as discussed next.

Structure Characterization from Density Maps

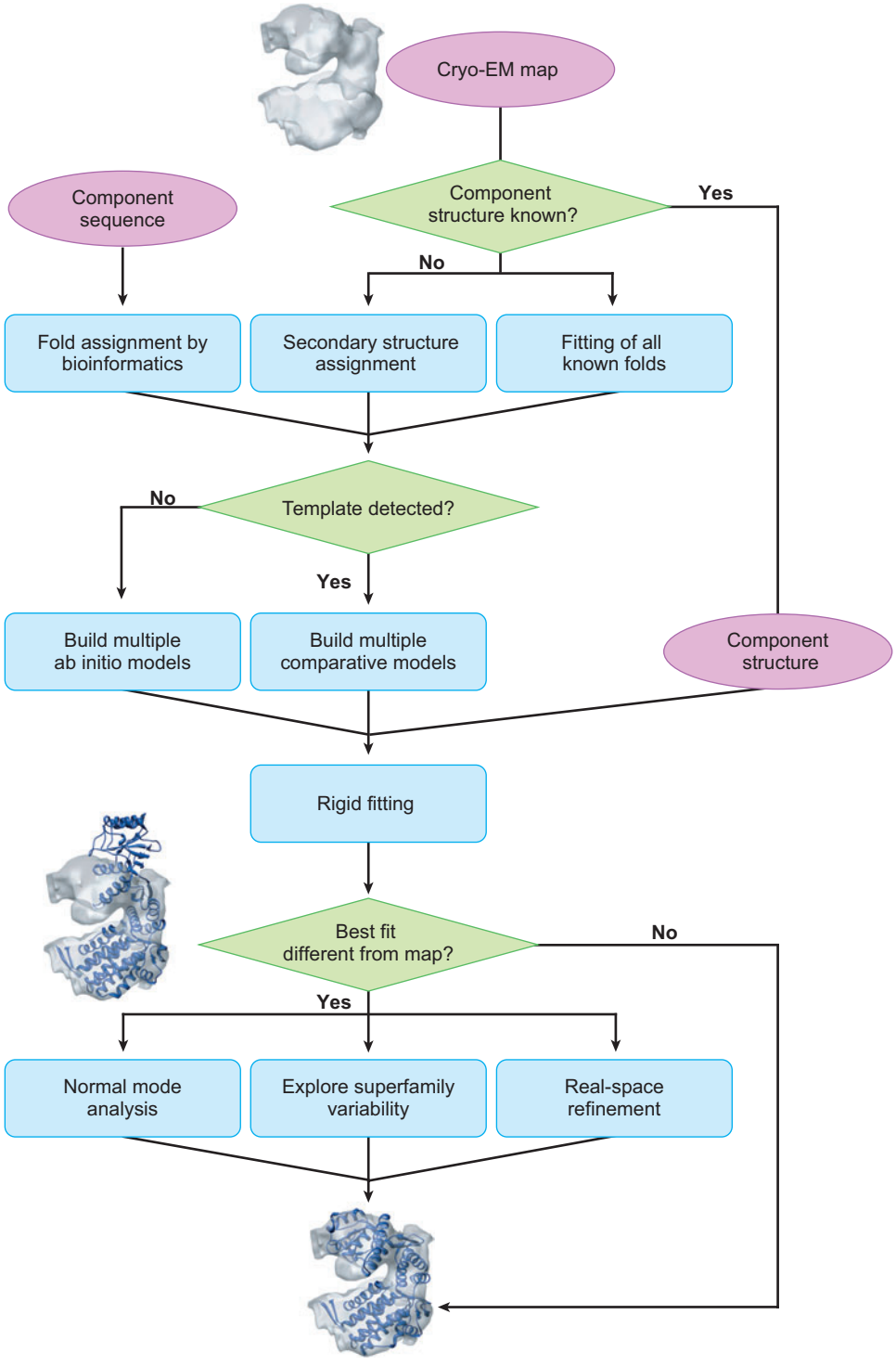
Structures of macromolecular assemblies, organelles, and even whole cells can be characterized by density maps derived from single-particle EM, electron crystallography, and cryo-ET. If structural models of components are available at a resolution higher than that of the map, it is usually helpful to fit these models into the map (**Figure 4**). Here, we focus on various fitting and segmentation methods, first for single-particle EM and electron crystallography and then for cryo-ET.

Segmentation. Interpretation of an EM map usually begins by identifying different structural units (e.g., secondary structure elements, domains, nucleic acids, proteins) in the density map by means of segmentation techniques (47). The size of the segmented units depends on the resolution of the map. For example, at 5–12-Å resolution, secondary structure segments can be seen (47). The segmentation is usually performed manually, with the aid of visualization tools such as Amira (**http://www.tgs.com/**) and Chimera (140). In addition, automated segmentation methods have been developed recently for both cryo-EM and cryo-ET maps (141, 142). Methods for assigning structural units in a given density map include skeletonization as well as identification of α-helices and β-sheets (47). If the component folds are known, the identified units can be useful in detecting the positions of the components in the map. Otherwise, when the folds are unknown, the identified units can help in predicting the component folds (143).

Density fitting: fitting of a component into a density map to maximize the overlap between the component and the map

Pseudoatomic structures. In many cases, atomic-resolution models of the components are often available from experiment or prediction (144). By fitting these models into the corresponding density map at better than approximately 20-Å resolution (145), a pseudoatomic interpretation of the map can be obtained (48, 146). The utility of fitting component structures into a cryo-EM map is demonstrated by a detailed pseudoatomic model of the mammalian 80S ribosome at 8.7-Å resolution (186) (**Figure 5**).

Rigid fitting. In some cases, the fitting of a component into a given map can be performed manually using interactive visualization tools such as Chimera (140). However, automated computational methods decrease the level of subjectivity as well as increase the accuracy and efficiency (145). Fitting of component structures is usually designed to optimize a similarity score between the component and the density map (e.g., the cross-correlation coefficient) as a function of its translation and rotation relative to the map (rigid fitting). Many methods for cryo-EM density fitting have been developed, including SITUS (147), COAN (148), EMFIT (149), DOCKEM (150), FOLDHUNTER (151), URO (152), CHARMM (153), ModEM (146), and ADP-EM (153a). To improve the positioning of components, one



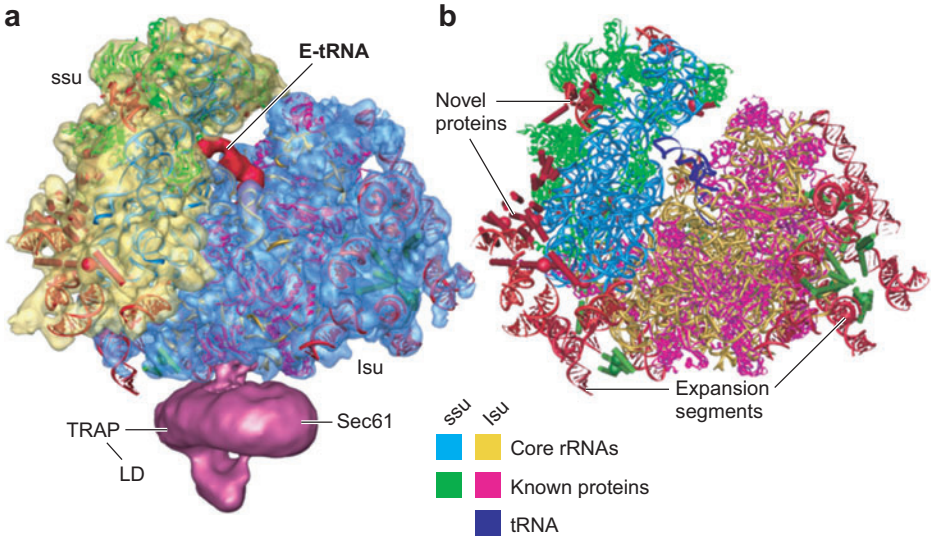


Figure 5

A model of the cytoplasmic 80S ribosome on the basis of an integrated protocol of comparative protein structure modeling, ab initio RNA modeling, density fitting, and density map analysis (186). To construct the model, thousands of comparative models of mammalian proteins were calculated and fitted into the cryo-EM map at 8.7-Å resolution using Mod-EM (146). The models with the highest combination of cross-correlation and statistical potential (34) scores were selected (161). Conserved mammalian core rRNA and ab initio models of expansion segments were fitted and refined in the density using RSRef (159). In addition, the resolution of the map in combination with the final model enabled the identification of the approximate positions of 20 novel proteins (without a homolog in bacteria), many containing rod-like features corresponding to α-helices (**Figure 4**). As a result, it was possible to identify unique interactions between mammalian proteins and expansion segments and obtain insights into conformational changes during translation. The final model is shown in a front view within the density map (*a*) and on its own (*b*). The E-site tRNA is shown in red between the small (ssu) and large subunits (lsu) of the ribosome. This specimen contained a native ER channel (*magenta*) comprised of Sec61 and the tryptophan RNA-binding attenuation protein (TRAP), with a prominent lumenal domain (LD). The subunit rRNAs and conserved proteins are color coded. The novel proteins (*spheres* and *rods*) and expansion segments (*red helices*) are also included.

can use additional experimental (e.g., labeling by gold) and computational information, e.g., statistical potentials (34) and geometric complementarity between domains (154; K. Lasker, M. Topf, A. Sali, & H. Wolfson, unpublished information).

←

Figure 4

Analysis of a density map from single-particle EM. If the atomic structure of the component is not known, bioinformatics methods for fold recognition can be applied to its sequence. These methods can also be combined with analysis of its density to identify structural features, such as secondary structure segments (47) or the complete fold (156). If a template fold has been detected, comparative modeling can be used to obtain a structural model; if the fold is not known, ab initio modeling can be used instead (155). Next, the component structure, experimentally determined or modeled, is rigidly fitted in the density map. For models, multiple models can be fitted, and the one that fits the density best is selected (146, 158). Finally, if there are differences between the fitted model and the map, flexible fitting can be applied to improve the fit by modifying the conformation of the component structure (145, 157, 185, 187).

Conformational variations. A common problem in fitting is that the isolated component structure may be in a different conformational state than in the intact assembly structure. These conformational differences can originate from the varied conditions under which the isolated component and assembly structures were determined as well as from errors in the experimental methods (such as crystal packing and noise). Common conformational differences are shear and hinge movements of domains and secondary structure elements, as well as loop distortions and movements. Furthermore, when an experimentally determined structure of the component is unavailable, the use of structure prediction methods to obtain component models (155) can introduce additional errors, such as misassignment of secondary structure elements to incorrect sequence regions and their shifts in space caused by target-template misalignment in comparative modeling (144).

Fitting multiple conformations. The simplest approach to consider conformational variations of component structures in fitting is to generate a set of different conformations, fit each of them into the density map, and select the top ranking conformation. This approach relies on a high correlation between the accuracy of a model and the cross-correlation score between the model and the corresponding density map (146). There are several different approaches to generating such models. First, candidate conformations can be calculated by exploring the structural variability within the fold superfamily of a component (**Figure 4**); for example, a number of alternative comparative models, which are based on different templates, can be fitted into the map, and the best fitting one is selected (MODELLER/Mod-EM) (146). Second, if the component's fold is unknown, candidate models for fitting can correspond to representatives of all known domain folds (Spi-EM) (156). Third, varied models can be created through rigid-body transformations of secondary structure segments guided by a principal component analysis of structurally aligned protein domains in the target's superfamily (S-flexfit) (157). Finally, a large number of models can also be produced by ab initio prediction on the basis of protein sequence alone (ROSETTA/FOLDHUNTER) (158).

Flexible fitting. The efficiency of conformational search is increased by considering the fit between the component and the map during the sampling. This goal can be achieved by optimizing the conformation of the component simultaneously with its position and orientation in the cryo-EM map while ideally maintaining correct stereochemistry. Such flexible fitting methods are similar to crystallographic refinement programs, except that they generally refine rigid bodies consisting of a number of atoms (e.g., secondary structure segments and domains) instead of individual atoms.

Several flexible fitting methods have been developed, utilizing different sampling and scoring schemes. For example, the real-space refinement programs RSRef (145, 159) and Flex-EM (187) rely on standard optimization methods, including conjugate gradients, molecular dynamics with simulated annealing, and Monte Carlo sampling. Another approach to flexible fitting attempts to improve the efficiency of conformational sampling is by a normal mode analysis of the component structure (160).

Certain large conformational changes can be efficiently sampled by Moulder-EM (161), a genetic algorithm protocol that generates comparative models through iterative sequence alignment, model building, model fitting, and model assessment. Conformational sampling arises from the iterative changes in the alignment on which the model is based. The fitness function of this genetic algorithm combines the cross-correlation score between the model and the map with an atomic distance-dependent statistical potential for model assessment (34).

Cellular atlas. Compared to electron crystallography and single-particle EM densities, cryo-ET maps are of lower resolution. However, cryo-ET can be used to reconstruct density maps of large cellular volumes, even whole cells. Such tomograms can be used to identify locations of assemblies in the cell (162). This task, also known as template matching, is challenging owing to a low signal-to-noise ratio, varying contrast throughout tomograms, and missing structure factors because of the limited tilt range ("missing wedge effect") (141, 162). The most common and general molecular detection algorithm is a locally normalized, matched filter, introduced for rigid-body fitting (150, 163). It was modified to account for the missing-wedge effect and applied to tomograms (61, 62, 164). These studies demonstrated that it is feasible to identify large macromolecular complexes (>500 kDa) within tomograms with high fidelity (61). Detection will be greatly facilitated by future instrumental advances in cryo-ET that will improve the resolution of the tomograms to the expected limit of approximately 2 nm (45).

Structure Characterization from Small-Angle X-ray Scattering

SAXS provides an approximate radial distribution function of a macromolecule in solution. For structure determination, additional information is needed because the radial distribution function alone is relatively uninformative about the details of molecular structure. We summarize different methods for integrating SAXS data into computational modeling of macromolecules.

SAXS data as a filter. Similarly to other types of experimental information, SAXS data can be used as a filter for a set of models generated independently by other methods. At the protein domain level, simulations have indicated that SAXS spectra can be used to choose close-to-native models from different comparative models (165). In quaternary structure determination, experimental SAXS spectra have been employed to choose one of a number of quaternary structure arrangements that resulted from computational docking of two assembly components to each other (65). Furthermore, SAXS spectra have been used to verify the accuracy of coarse-grained simulations of lipoproteins (166).

SAXS data in optimization. SAXS data can also be a term in a scoring function that is optimized to obtain a model consistent with the data. The first approaches to optimize models on the basis of SAXS data relied on representing macromolecular surfaces using spherical harmonics (167). However, this representation has a relatively low resolution and led to the development of alternative methods. Owing to the sparseness of SAXS data, virtually all subsequently developed methods aimed to integrate additional information into structure determination.

Coarse-grained approaches represent the macromolecule as an assembly of beads on a grid (168–170). This representation enforces an overall mass by using a required number of beads and potential geometrical symmetry by symmetric sampling. In addition, compactness of the models is ensured by restricting the sampling to the vicinity of a compact initial model (168, 169) or by including appropriate terms into the scoring function (170, 171). Higher-resolution modeling approaches represent a protein as a chain of beads rather than a grid (171).

If high-resolution structural information is available for some parts of the protein, the conformational sampling can be focused only on the undefined segments. One approach has been developed to approximate the structure of missing loops in structures derived by X-ray crystallography using coarse-grained beads (172). Another approach was developed to determine the spatial arrangement of domains of known structure and the structures of their connecting linkers (173). Here, the elements of a known structure are kept rigid, and their translations and rotations are optimized using a simulated annealing protocol.

If flexible linkers are present, they are represented as beads that tether the domains.

Completeness of structural coverage: the fraction of the studied assembly that is represented in its model

Heterogeneous samples. All SAXS modeling methods described above assume that the protein is present in only one conformation under the conditions examined. If multiple conformations are present, the SAXS spectrum of the sample is a weighted average of the SAXS spectra of each conformation. Recently, methods have been proposed to fit an ensemble of models to a given spectrum (174).

Integration of SAXS data with other information. The recent renaissance of SAXS is to a large extent the result of efforts on integrating SAXS with other structural information from additional complementary sources (64). This integration is necessary because the information content of SAXS data is low, given the number of degrees of freedom that typically need to be determined and even more so for heterogeneous samples. For example, the SAXS data of proteins or smaller complexes can be considered simultaneously with corresponding cryo-EM maps (175). Another approach, which included SAXS data into molecular dynamics simulations, has shown promising results for simulated examples (176). Recently, SAXS spectra have been incorporated into a protocol for structure determination by NMR spectroscopy (177). The SAXS data contain global information on the protein that is complementary to the short-range restraints from NMR spectroscopy and, hence, significantly increase the accuracy of models for multidomain proteins compared to models based on NMR spectra only.

To generalize further, we have incorporated the use of SAXS data into our program, the Integrated Modeling Platform (IMP), for the modeling of proteins and their assemblies by satisfaction of spatial restraints (**http://salilab.org/imp;** 167a; F. Förster, B. Webb, K.A. Krukenberg, H. Tsuruta, D.A. Agard, & A. Sali, unpublished information). We have used the SAXS-based restraint in conjunction with restraints on the rigidity of domains, steric clashes, stereochemistry, and an atomic distance-based statistical potential (34) (**Figure 6**). By thorough sampling of the configurational space and subsequent clustering of low scoring solutions, we aim to enumerate models that are consistent with the given SAXS data. The integration of additional information reduces the ambiguity of such models.

COMPREHENSIVE DATA INTEGRATION BY SATISFACTION OF SPATIAL RESTRAINTS

Detailed structural characterization of assemblies is often difficult by any single existing experimental or computational method. We suggest that this barrier can be overcome by hybrid approaches that integrate data from diverse biochemical and biophysical experiments as well as computational methods. This information may vary greatly in terms of its resolution, accuracy, and quantity. Here, we outline an approach for generating structures of macromolecular assemblies that are consistent with all available information from experimental methods, physical theories, and statistical preferences extracted from biological databases. Such an integrative system will help maximize efficiency, resolution, accuracy, precision, and completeness of the structural coverage of macromolecular assemblies.

Theory and Method

We begin this section by describing the underlying theory and methods of our hybrid approach to characterizing macromolecular assembly structures. Then, we highlight an example, the structure determination of the NPC (2, 17–19).

Formalization of the problem. The complete process of structure determination can be seen as a potentially iterative series of four steps, including data generation by experiments, data translation into spatial restraints,

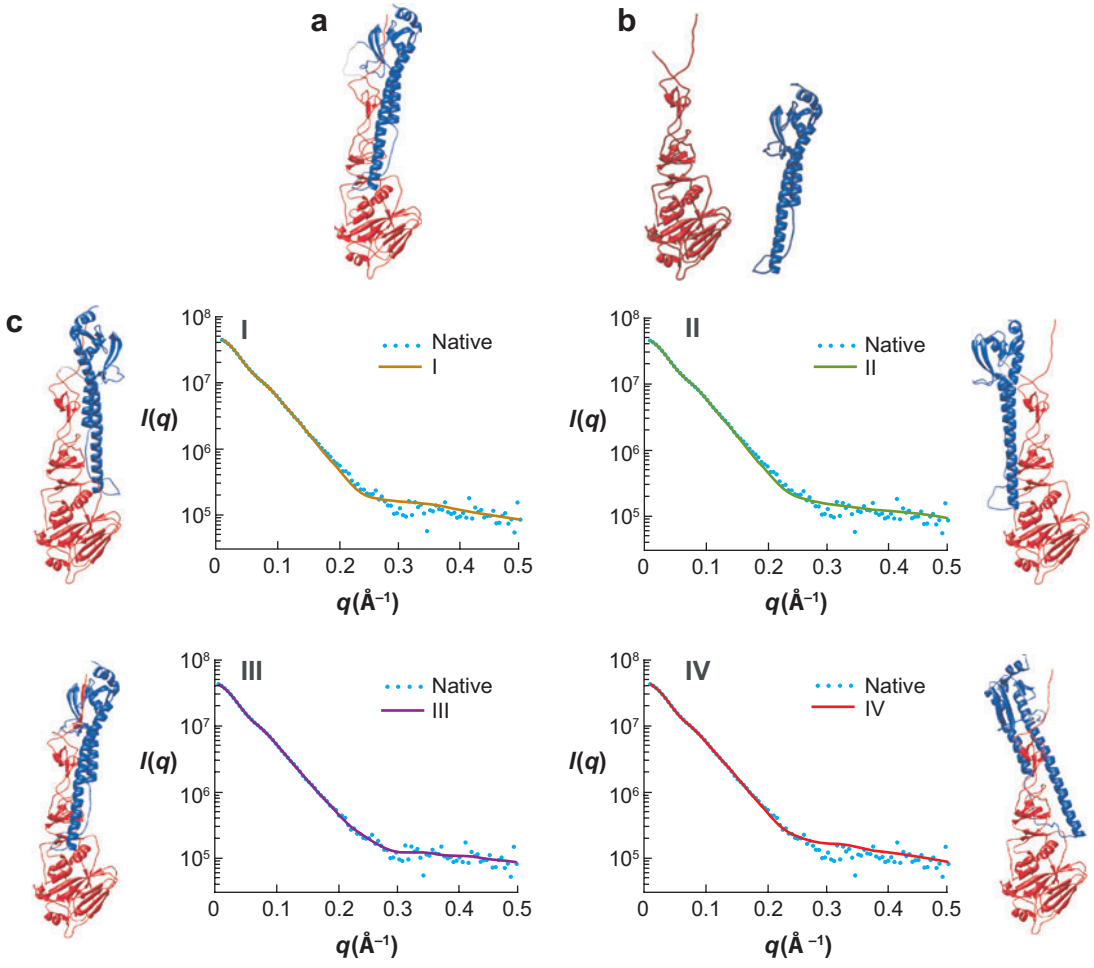
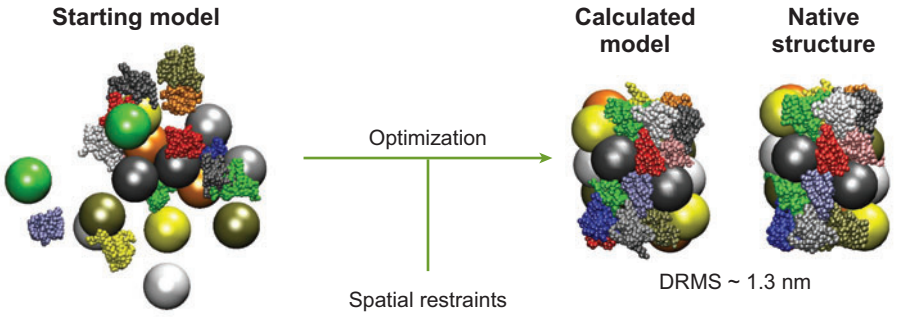


Figure 6

Modeling influenza hemagglutinin using SAXS. (*a*) Native hemagglutinin consists of two domains (*blue* and *red*) (PDB code 1a0d). The SAXS spectrum of this structure was simulated with the addition of some white noise, which typically occurs in experimental data. (*b*) We approximated the hemagglutinin domains by their structures in the postcleavage forms (2viu). (*c*) Models were obtained by optimization from 600 different initial "seeds." The optimized models were subsequently clustered into four major groups. The models with the lowest SAXS penalty (χ^2 of the experimental data and the SAXS spectrum of the model) from each cluster (*top*) and the corresponding SAXS spectra (*bottom*) are shown. The model from cluster III has the lowest SAXS score ($\chi^2 = 0.84$ compared to 5.13, 3.86, and 3.68 for the models from clusters I, II, and IV, respectively) and is closest to the native state in terms of its Cα RMSD (2.7 Å compared to 13.6, 16.4, and 14.6 Å). However, the differences between the cluster scores are small, demonstrating the problem of ambiguity in modeling an assembly structure on the basis of its SAXS spectrum. Abbreviations: $I(q)$, scattering intensity; native, simulated spectrum, which was used as the input for modeling; q[Å^{-1}], spatial frequency in Å^{-1}.

calculation of an ensemble of structures by satisfaction of spatial restraints, and an analysis of the ensemble. The structural characterization part of the process can be expressed as an optimization problem (**Figure 7**). In this view, models that are consistent with the input information are calculated by optimizing a scoring function. The three components of



Immuno-EM	EM	EM	Cross-linking	Affinity purifications	Ultra-centrifugation
14 positions	Assembly shape	Assembly symmetry	36 cross-links	28 affinity purifications	7 S-values
14 position restraints	Shape restraints	Symmetry restraints	36 distance restraints	28 connectivity restraints	7 protein shapes

Figure 7

Characterization of an assembly configuration on the basis of data simulated from a known native structure (67). The simulated data include protein positions (e.g., from immuno-EM), assembly shape (e.g., from EM), relative proximity of components (e.g., from cross-linking and affinity purification). The data is translated into spatial restraints that are then summed to obtain a scoring function. A random starting structure is optimized by a combination of conjugate gradients and molecular dynamics with simulated annealing to minimize violations of all restraints. The listed data were sufficient to identify the coarse relative position of each protein (i.e., the protein configuration). To illustrate the possibility of using different representations for different proteins, a protein is represented either by an X-ray structure or by a single sphere that best reproduces its hydrodynamic properties determined by ultracentrifugation. Abbreviations: DRMS, distance root-mean-square difference between the protein centroids in the determined model and the native structure; EM, electron microscopy.

this approach are (*a*) a representation of the modeled assembly, (*b*) a scoring function consisting of the individual spatial restraints, and (*c*) optimization of the scoring function to obtain all possible models that satisfy the input restraints.

Representation. The modeled structure is represented by a hierarchy of particles, defined by their positions and other properties (**Figure 7**). For a protein assembly, the hierarchy can include atoms, atomic groups, amino acid residues, secondary structure segments, domains, proteins, protein subcomplexes, symmetry units, and the whole assembly. The coordinates and properties of particles at any level are calculated from those at the highest resolution level. Different parts of the assembly can be represented at different resolutions to reflect the input information about the structure (**Figure 7**). Moreover, different representations can also apply to the same part of the system. For example, affinity purification may indicate proximity between two proteins, and cross-linking may indicate which specific residues are involved in the interaction.

Scoring function. The most important aspect of structure characterization is to accurately capture all experimental, physical, and statistical information about the modeled

structure. This objective is achieved by expressing knowledge of any kind as a scoring function whose global optimum corresponds to the native assembly structure (34). One such function is a joint probability density function (pdf) of the Cartesian coordinates of all assembly proteins, given the available information, I, about the system, $p(C|I)$, where $C = (c_1, c_2, \ldots, c_n)$ is the list of the Cartesian coordinates, c_i, of the n component proteins in the assembly. The joint pdf, p, gives the probability density that a component, i, of the native configuration is positioned very close to c_i, given the information, I, we wish to consider in the calculation. In general, I may include any structural information from experiments, physical theories, and statistical preferences. For example, when I reflects only the sequence and the laws of physics under the conditions of the canonical ensemble, the joint pdf corresponds to the Boltzmann distribution. If I also includes a crystallographic dataset sufficient to define the native structure precisely, the joint pdf is a Dirac delta function centered on the native atomic coordinates.

The complete joint pdf is generally unknown but can be approximated as a product of pdfs, p_f, that describe individual assembly features (e.g., distances, angles, interactions, or relative orientations of proteins):

$$p(C|I) = \prod_f p_f(C|I_f).$$

The scoring function $F(C)$ is then defined as the logarithm of the joint pdf:

$$F(C) = -\ln \prod_f p_f(C|I_f) = \sum_f r_f(C).$$

For convenience, we refer to the logarithm of a feature pdf as a restraint, r_f, and the scoring function is therefore a sum of the individual restraints.

Restraints. A restraint, r_f, can in principle have any functional form. However, it is convenient if ideal solutions consistent with the data correspond to values of 0 and values larger than 0 correspond to a violated restraint; for example, a restraint is frequently a harmonic function of the restrained feature.

Restrained features. The restrained features, in principle, include any structural aspect of an assembly, such as contacts, proximity, distances, angles, chirality, surface, volume, excluded volume, shape, symmetry, and localization of particles and sets of particles.

Translating data into restraints. A key challenge is to accurately express the input data and their uncertainties in terms of the individual spatial restraints. An interpretation of the data in terms of a spatial restraint generally involves identifying the restrained components (i.e., structural interpretation) and the possible values of the restrained feature implied by the data. For instance, the shape, density, and symmetry of a complex or its subunits may be derived from X-ray crystallography and EM (38); upper distance bounds on residues from different proteins may be obtained from NMR spectroscopy (22) and chemical cross-linking (95); protein-protein interactions may be discovered by the yeast two-hybrid system (178) and calorimetry (73); two proteins can be assigned to be in proximity if they are part of an isolated subcomplex identified by affinity purification in combination with mass spectrometry (79). Increasingly, important restraints will be derived from pairwise molecular docking (118), statistical preferences observed in structurally defined protein-protein interactions (139), and analysis of multiple sequence alignments (179).

Conditional restraints. If structural interpretation of the data is ambiguous (i.e., the data cannot be uniquely assigned to specific components), only "conditional restraints" can be defined. For example, when there is more than one copy of a protein per assembly, a yeast two-hybrid system indicates only which protein types but not which instances interact with each other. Such ambiguous information must be translated into

a conditional restraint that considers all alternative structural interpretations of the data (**Figure 8**). The selection of the best alternative interpretation is then achieved as part of the structure optimization process.

Optimization methods. Structures can be generated by simultaneously minimizing the violations of all restraints, resulting in configurations that minimize the scoring function *F*. It is crucial to have access to multiple optimization methods to choose one that works best with a specific scoring function and representation. Optimization methods implemented in IMP currently include conjugate gradients, quasi-Newton minimization, and molecular dynamics, as well as more sophisticated schemes, such as self-guided Langevin dynamics, the replica exchange method, and exact inference (belief propagation) (K. Lasker, M. Topf, A. Sali, & H. Wolfson, unpublished information); all of these methods can refine positions of the individual particles as well as treat subsets of particles as rigid bodies.

Outcomes. There are three possible outcomes of the calculation. First, if only a single model satisfies all input information, there is probably sufficient data for prediction of the unique native state. Second, if different models are consistent with the input information, the data are insufficient to define the single native state, or there are multiple native structures. If the number of distinct models is small, the structural differences between the models may suggest additional experiments to narrow down the possible solutions. Third, if no models satisfy all input information, the data or their interpretation in terms of the restraints are incorrect.

Analysis. In general, a number of different configurations may be consistent with the input restraints. The aim is to obtain as many structures as possible that satisfy all input restraints. To comprehensively sample such structural solutions consistent with the data, independent optimizations of randomly generated initial configurations need to be performed until an ensemble of structures satisfying the input restraints is obtained. The ensemble can then be analyzed in terms of assembly features, such as the protein positions, contacts, and configuration. These features can generally vary among the individual models in the ensemble. To analyze this variability, a probability distribution of each feature can be calculated from the ensemble. Of particular interest are the features that are present in most configurations in the ensemble and have a single maximum in their probability distribution. The spread around the maximum describes how precisely the feature was determined by the input restraints. When multiple maxima are present in the feature distribution at the precision of interest, the input restraints are insufficient to define the single native state of the corresponding feature (or there are multiple native states).

Predicting accuracy. Assessing the accuracy of a structure is important and difficult. The accuracy of a model is defined as the difference between the model and the native structure. Therefore, it is impossible to know with certainty the accuracy of the proposed structure without knowing the real native structure. Nevertheless, our confidence can be modulated by five considerations: (*a*) self-consistency of independent experimental data; (*b*) structural similarity among all configurations in the ensemble that satisfy the input restraints; (*c*) simulations where a native structure is assumed, corresponding restraints are simulated from it, and the resulting calculated structure is compared with the assumed native structure; (*d*) confirmatory spatial data that were not used in the calculation of the structure (e.g., criterion similar to the crystallographic free R-factor (180) can be used to assess both the model accuracy and the harmony among the input restraints); and (*e*) patterns emerging from a mapping of independent and unused data on the structure that are unlikely to occur by chance (18, 19).

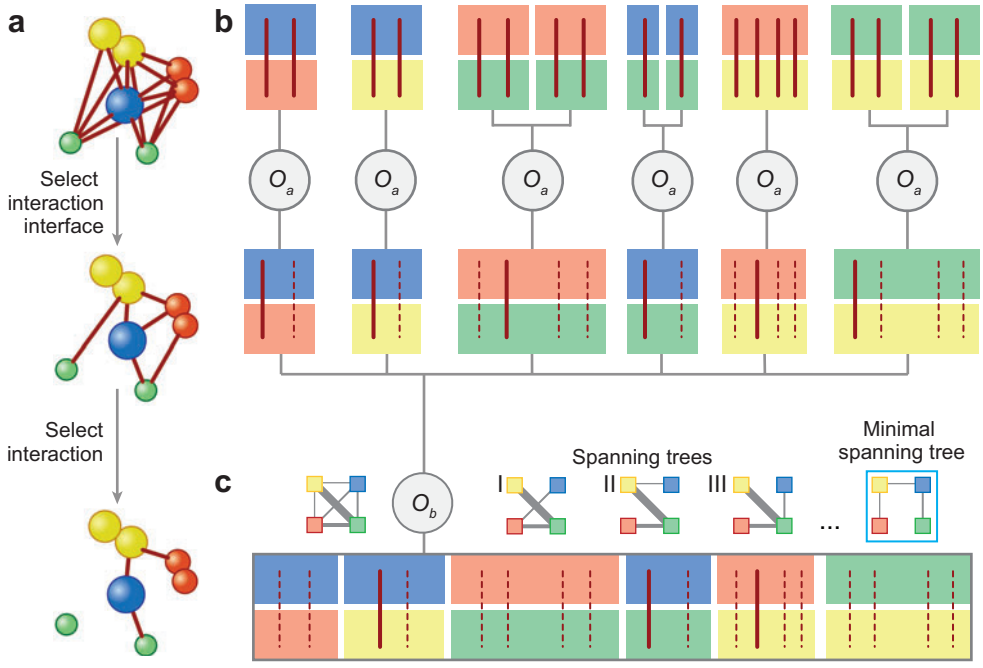


Figure 8

Conditional restraint (19). As an example, shown is a conditional restraint on protein contacts derived from a single affinity purification experiment that identified 4 protein types (*yellow*, *blue*, *red*, *green*), obtained from an assembly containing a single copy of the yellow, blue, and red protein and two copies of the green protein (18, 19, 67). (*a*) A single protein is represented by either one bead (*blue* and *green* proteins) or two beads (*yellow* and *red* proteins); alternative interactions between proteins are indicated by different edges. (*b*) Protein contacts are selected in a decision tree-like evaluation process by operator functions O_a and O_b. Red vertical lines indicate restraints that encode a protein contact; thick vertical lines are a subset of restraints that are selected for contribution to the final value of the conditional restraint, whereas dotted vertical lines indicate restraints that are not selected. Also shown are spanning trees of a "composite graph." (*c*) The composite graph is a fully connected graph that consists of nodes for all identified protein types (*square nodes*) and edges for all pairwise interactions between protein types (*left* of the O_b operator); edge weights correspond to violations of interaction restraints and quantify how consistent the corresponding interaction is with the current assembly structure. A "spanning tree" is a graph with the smallest possible number of edges that connect all nodes; a subset of 4 out of 16 spanning trees is indicated to the right of the O_b operator. The "minimal spanning tree" is the spanning tree with the minimal sum of edge weights (i.e., restraints violations). The sample affinity purification implies that at least three of the following six possible types of interactions must occur: blue-red, blue-yellow, blue-green, red-green, red-yellow, and yellow-green. In addition, (*i*) the three selected interactions must form a spanning tree of the composite graph; (*ii*) each type of interaction can involve either copy of the green protein; and (*iii*) each protein can interact through any of its beads. These considerations can be encoded through a tree-like evaluation of the conditional restraint. At the top level, all possible bead-bead interactions between all protein copies are clustered by protein types. Each alternative bead interaction can be restrained by a restraint corresponding to a harmonic upper bound on the distance between the beads; these are termed "optional restraints" because only a subset is selected for contribution to the final value of the conditional restraint. Next, an operator function (O_a) selects only the least violated optional restraint from each interaction type, resulting in six restraints (*thick red lines*) at the middle level of the tree. Finally, a minimal spanning tree operator (O_b) finds the minimal spanning tree corresponding to the combination of three restraints that are most consistent with the affinity purification (*thick red lines*). The whole restraint evaluation process is executed at each optimization step on the basis of the current configuration, thus resulting in possibly different subsets of selected optional restraints at each step (19).

Advantages. The integrative approach to structure determination has several advantages. It benefits from the synergy among the input data, minimizing the drawback of incomplete, inaccurate, and/or imprecise data sets (although each individual restraint may contain little structural information, the concurrent satisfaction of all restraints derived from independent experiments may drastically reduce the degeneracy of structural solutions). It can potentially produce all structures that are consistent with the data, not just one. The variation among the structures, consistent with the data, allows an assessment of the sufficiency of the data and the precision of the representative structure. Finally, this approach makes the process of structure determination more efficient by indicating what measurements would be most informative.

Structural Characterization of the Nuclear Pore Complex

Using the approach outlined above, we determined the native configuration of proteins in the yeast NPC (18, 19). These NPCs are large (~50 MDa) proteinaceous assemblies spanning the NE, where they function as the sole mediators of bidirectional macromolecular exchange between the nucleoplasmic and cytoplasmic compartments in all eukaryotes (181). EM images of the yeast NPC at ~200-Å resolution revealed that the nuclear pore forms a channel by stacking two similar rings, each one consisting of eight radially arranged "half-spoke" units (182). The yeast NPC is built from multiple copies of 30 different proteins, totaling ~456 proteins (called nucleoporins or nups).

Although low-resolution EM has provided valuable insights into the overall shape of the NPC, the spatial configuration of its component proteins and the detailed interaction network between them were unknown. A description of the NPC's structure was needed to understand its function and assembly as well as to provide clues to its evolutionary origins. Owing to its size and flexibility, detailed structural characterization of the complete NPC assembly has proven to be extraordinarily challenging. Further compounding the problem, atomic structures have only been solved for domains covering ~5% of the protein sequence (183).

To determine the protein configuration of the NPC, we collected a large and diverse set of biophysical and biochemical data. The data were derived from six experimental sources (**Figure 9**).

1. Quantitative immunoblotting experiments determined the stoichiometry of all 30 nups in the NPC.
2. Hydrodynamics experiments provided information about the approximate excluded volume and the coarse shape of each nup.
3. Immuno-EM provided a coarse localization for each nup along two principal axes of the NPC.
4. An exhaustive set of affinity purification experiments determined the composition of 77 NPC complexes.
5. Overlay experiments determined five direct binary nup interactions.
6. Symmetry considerations and the dimensions of the NE were extracted from cryo-EM. Moreover, bioinformatics analysis provided information about the position of transmembrane helices for the three integral membrane nups. These data were translated into spatial restraints on the NPC (**Figure 9**).

The relative positions and proximities of the NPC's constituent proteins were then produced by satisfying these spatial restraints, using the approach described above and illustrated in **Figure 10**. Optimization relies on conjugate gradients and molecular dynamics with simulated annealing. It starts with a random configuration of proteins and then iteratively moves these proteins so as to minimize violations of the restraints (**Figure 10**). To comprehensively sample all

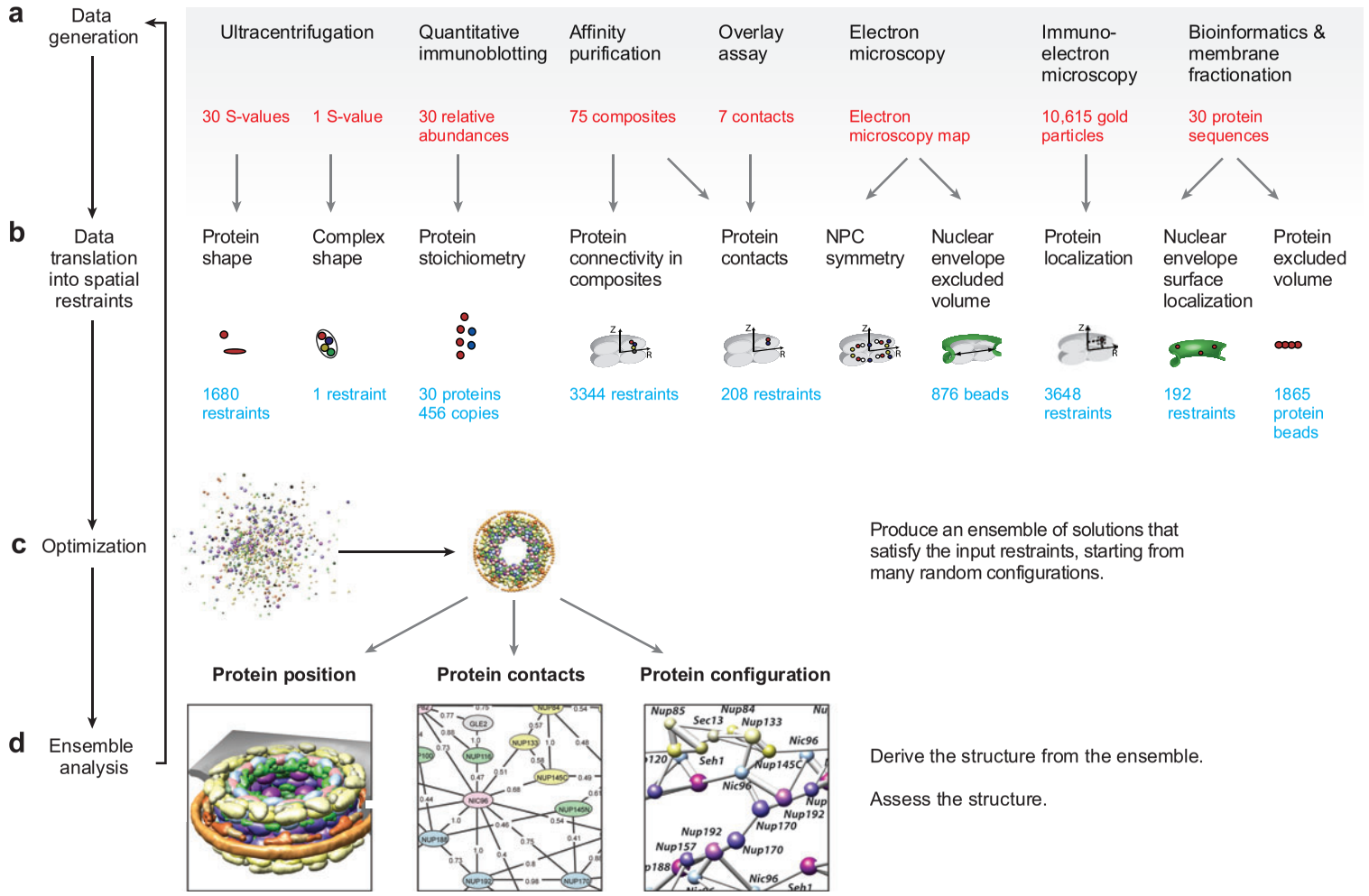


Figure 9

Determining the architecture of the NPC by integrating spatial restraints from proteomic data (19). First, (*a*) various experiments (*black*) generate structural data (*red*). Second (*b*), the data and theoretical considerations are expressed as spatial restraints (*blue*). Third (*c*), an ensemble of structural solutions that satisfy the data is obtained by minimizing the violations of the spatial restraints, starting from many different random configurations. Fourth (*d*), the ensemble is clustered into sets of distinct solutions as well as analyzed in terms of protein positions, contacts, and configuration.

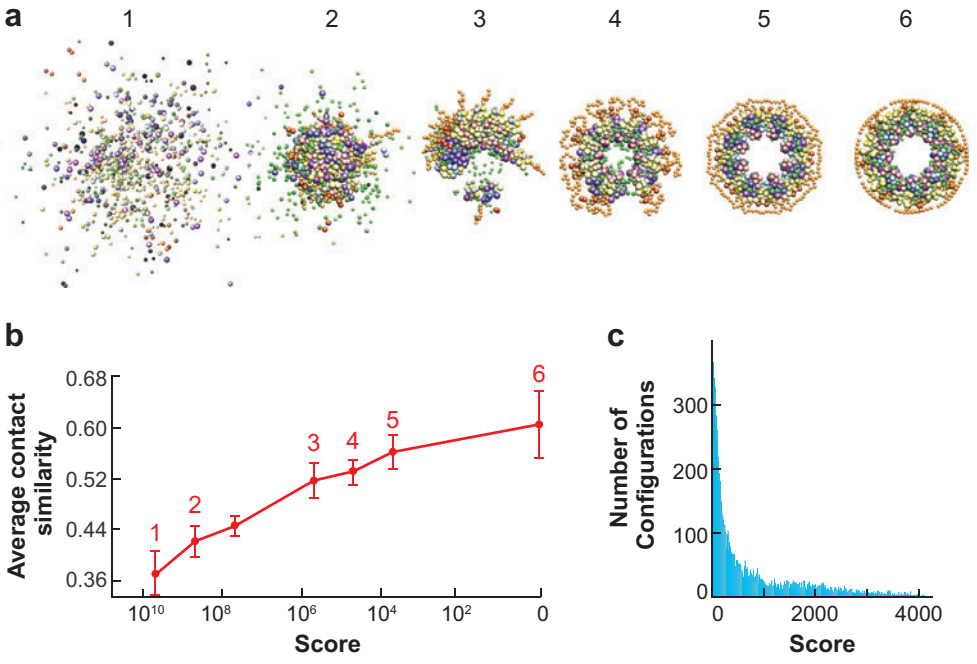


Figure 10

Calculation of the NPC bead structure by satisfaction of spatial restraints. (*a*) Representation of the optimization process as it progresses from an initial random configuration to an optimal solution. (*b*) The graph shows the relationship between the score (a measure of the consistency between the configuration and the input data) and the average contact similarity. The contact similarity quantifies how similar two configurations are in terms of the number and types of their protein contacts; two proteins are considered to be in contact when they are sufficiently close to one another given their size and shape. The average contact similarity at a given score is determined from the contact similarities between the lowest scoring configuration and a sample of 100 configurations with that given score. Error bars indicate standard deviation. Representative configurations at various stages of the optimization process from left (very large scores) to right (with a score of 0) are shown above the graph; a score of 0 indicates that all input restraints have been satisfied. As the score approaches zero, the contact similarity increases, showing that there is only a single cluster of closely related configurations that satisfies the input data. (*c*) Distribution of configuration scores demonstrates that our sampling procedure finds configurations consistent with the input data. These configurations satisfy all the input restraints within the experimental error (19).

possible structural solutions that are consistent with the data, we obtained an ensemble of 1000 independently calculated structures that satisfied the input restraints (**Figure 10*c***). After superposition of these structures, the ensemble was converted into the probability of finding a given protein at any point in space (i.e., the localization probability). The resulting localization probabilities yielded a single pronounced maximum for almost every protein, demonstrating that the input restraints define a single NPC architecture (**Figure 11**). The average standard deviation for the separation between neighboring protein centroids is 5 nm. Given that this level of precision is less than the diameter of many proteins, our map is sufficient to determine the relative position of proteins in the NPC. Although each individual restraint may contain little structural information, the concurrent satisfaction of all restraints derived from independent experiments drastically reduces the degeneracy of the structural solutions (**Figure 12**).

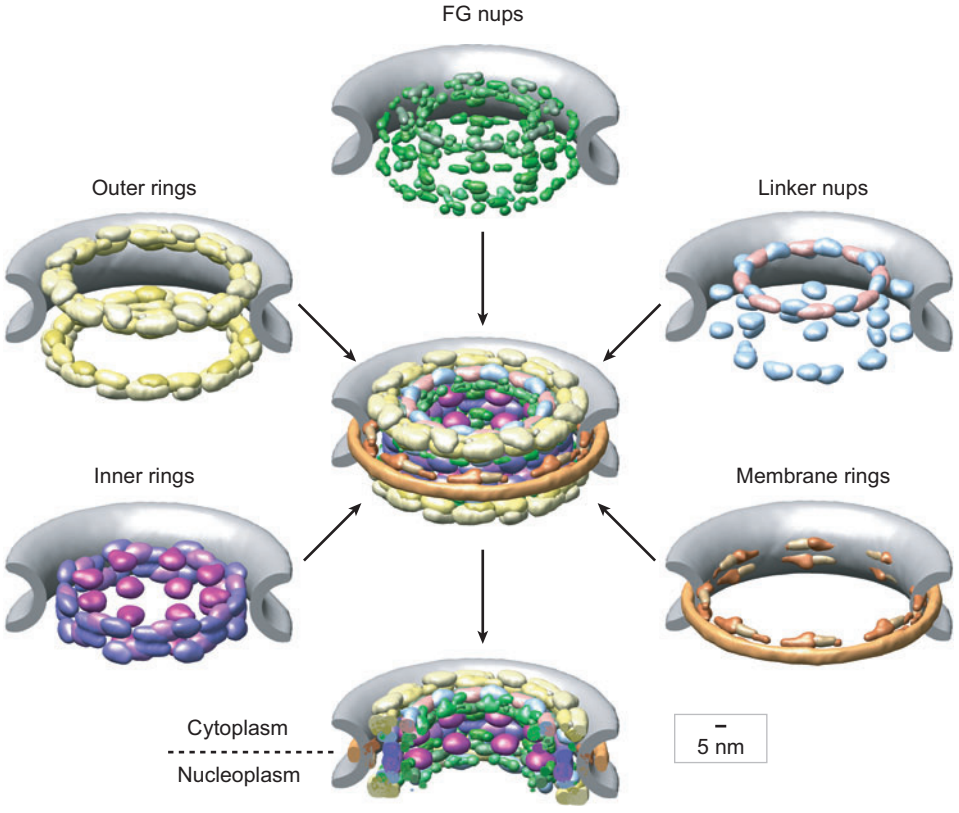


Figure 11

Localization of major substructures and their component proteins in the NPC. The proteins are represented by their localization volumes and have been colored according to their classification into five distinct substructures on the basis of their location and functional properties: the outer rings in yellow, the inner rings in purple, the membrane rings in brown, the linker nups in blue and pink, and the FG nups (for which only the structured domains are shown) in green. The pore membrane is shown in gray (18).

Our structure (**Figure 11**) reveals that half of the NPC is made of a core scaffold, which is structurally analogous to vesicle coating complexes. This scaffold forms an interlaced network that coats the entire curved surface of the NE within which the NPC is embedded. The selective barrier for transport is formed by large numbers of proteins with disordered regions that line the inner face of the scaffold. The NPC consists of only a few structural modules. These modules resemble each other in terms of the configuration of their homologous constituents. The architecture of the NPC thus appears to be based on the hierarchical repetition of the modules that likely evolved through a series of gene duplications and divergences. Thus, the determination of the NPC configuration in combination with the fold prediction (183, 184) of its constituent proteins can provide clues to the ancient evolutionary origins of the NPC.

In the future, we envision combining cryo-ET, proteomics, cross-linking, cryo-EM of subcomplexes, and experimentally determined or modeled atomic structures of the individual subunits to obtain a pseudoatomic model of the whole NPC assembly in action.

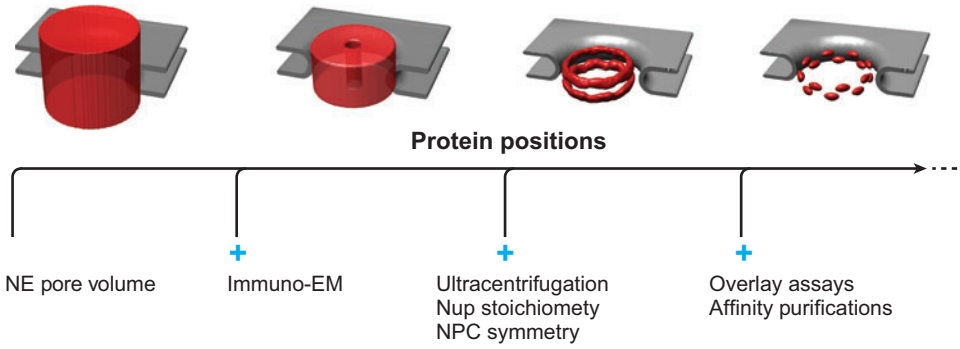


Figure 12

Synergy between varied datasets results in increased precision of structure determination. The proteins are increasingly localized by the addition of different types of synergistic experimental information. As an example, each panel illustrates the localization of 16 copies of Nup192 in the ensemble of nuclear pore complex (NPC) structures generated, using the datasets indicated below. The smaller the volume (*red*), the better localized is the protein. The NPC structure is therefore essentially "molded" into shape by the large amount of experimental data (19). Abbreviations: NE, nuclear envelope; Nup, nucleoporin protein.

CONCLUSIONS

There is a wide spectrum of experimental and computational methods for identification and structural characterization of macromolecular complexes. The data from these methods need to be combined through integrative computational approaches to achieve higher resolution, accuracy, precision, completeness, and efficiency than any of the individual methods. New methods must be capable of generating possible alternative models consistent with information from various sources, such as stoichiometry, interaction data, similarity to known structures, docking results, and low-resolution images.

Structural biology is a great unifying discipline of biology. Thus, structural characterization of many protein complexes will bridge the gaps between genome sequencing, functional genomics, proteomics, and systems biology. The goal seems daunting, but the prize will be commensurate with the effort invested, given the importance of molecular machines and functional networks in biology and medicine.

SUMMARY POINTS

1. To understand the cell, we need to determine the structures of macromolecular assemblies, many of which consist of tens and even hundreds of components.
2. A variety of experimental methods exists that generates structural information about assemblies, from atomic-resolution data to coarse descriptions of the component arrangement in the complex.
3. To maximize the completeness, accuracy, and resolution of the structural determination, a computational approach is needed that can use spatial information from a variety of experimental methods.
4. The complete process of structure determination can be seen as a potentially iterative series of four steps, including data generation by experiments, data interpretation in terms of spatial restraints, calculation of an ensemble of structures by satisfaction of spatial restraints, and an analysis of the ensemble.

5. The structure calculation part of this process is conveniently expressed as an optimization problem, and the solution requires three main components: the representation of an assembly, a scoring function, and optimization.
6. The power of the integrative approach is illustrated by its use of the proteomic data to define the configuration of proteins in large assemblies, such as the NPC.

FUTURE ISSUES

1. Information from experimental and theoretical sources in terms of individual spatial restraints needs to be quantified.
2. Individual restraints should be combined into an accurate scoring function.
3. Thorough sampling schemes for finding good solutions to a scoring function are needed.
4. Methods for a comprehensive analysis of the ensemble of models consistent with the data can be developed.
5. Accurate methods for predicting the likely accuracy of the input data and the corresponding structure are needed.
6. Robust, efficient, user friendly, and generally applicable computer software for calculating assembly structures on the basis of varied datasets should be developed.
7. Descriptions of the structures as well as dynamics of both stable and transient complexes are needed.

DISCLOSURE STATEMENT

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

ACKNOWLEDGMENTS

We are grateful to Michael P. Rout, Brian T. Chait, John Aitchison, Chris Akey, Wah Chiu, David Agard, Wolfgang Baumeister, Joachim Frank, Fred Davis, M.S. Madhusudhan, Min-yi Shen, Michael Kim, Keren Lasker, Daniel Russell, Javier Velazquez-Muriel, Bret Peterson, and Ben Webb for many discussions about structure characterization by satisfaction of spatial restraints. We are also thankful to Svetlana Dokudovskaya, Liesbeth Veenhoff, Whenzu Zhang, Julia Kipper, Damien Devos, Adisetyantari Suprapto, Orit Karni-Schmidt, and Rosemary Williams for their contribution to the determination of the NPC structure. F.F. has been funded by a Human Frontier Science Organization long-term fellowship. M.T. is grateful for a MRC career development award. We also acknowledge support from the Sandler Family Supporting Foundation, NIH/NCRR U54 RR022220, NIH R01 GM54762, Human Frontier Science Program, NSF IIS 0705196, and NSF EIA-0324645. And we are grateful for computer hardware gifts from Ron Conway, Mike Homer, Intel, Hewlett-Packard, IBM, and Netapp.

LITERATURE CITED

1. Alberts B. 1998. *Cell* 92:291–94
2. Sali A, Kuriyan J. 1999. *Trends Cell Biol.* 9:M20–24
3. Sali A, Glaeser R, Earnest T, Baumeister W. 2003. *Nature* 422:216–25
4. Sali A. 2003. *Structure* 11:1043–47
5. Robinson C, Sali A, Baumeister W. 2007. *Nature* 450:973–82
6. Aebersold R, Mann M. 2003. *Nature* 422:198–207
7. Russell RB, Alber F, Aloy P, Davis FP, Korkin D, et al. 2004. *Curr. Opin. Struct. Biol.* 14:313–24
8. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. 2000. *Nature* 403:623–27
9. Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, et al. 2000. *Proc. Natl. Acad. Sci. USA* 97:1143–47
10. Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, et al. 2007. *Mol. Cell Proteomics* 6:439–50
11. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. 2006. *Nature* 440:637–43
12. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, et al. 2006. *Nature* 440:631–36
13. Devos D, Russell RB. 2007. *Curr. Opin. Struct. Biol.* 17:370–77
14. Abbott A. 2002. *Nature* 417:894–96
15. Harris ME, Nolan JM, Malhotra A, Brown JW, Harvey SC, Pace NR. 1994. *EMBO J.* 13:3953–63
16. Malhotra A, Harvey SC. 1994. *J. Mol. Biol.* 240:308–40
17. Alber F, Eswar N, Sali A. 2004. In *Practical Bioinformatics*, ed. JM Bujnicki, pp. 73–96. Germany: Springer-Verlag
18. Alber F, Dokudovskaya S, Veenhoff L, Zhang W, Kipper J, et al. 2007. *Nature* 450:695–701
19. Alber F, Dokudovskaya S, Veenhoff L, Zhang W, Kipper J, et al. 2007. *Nature* 450:683–94
20. Adams PD, Pannu NS, Read RJ, Brunger AT. 1999. *Acta Crystallogr. D Biol. Crystallogr.* 55(Part 1):181–90
21. Cramer P, Bushnell DA, Fu J, Gnatt AL, Maier-Davis B, et al. 2000. *Science* 288:640–49
22. Fiaux J, Bertelsen EB, Horwich AL, Wuthrich K. 2002. *Nature* 418:207–11
23. Bonvin AM, Boelens R, Kaptein R. 2005. *Curr. Opin. Chem. Biol.* 9:501–8
24. Rieping W, Habeck M, Nilges M. 2005. *Science* 309:303–6
25. Zuiderweg ER. January 2002. *Biochemistry* 41:1–7
26. McCoy MA, Wyss DF. 2002. *J. Am. Chem. Soc.* 124:2104–5
27. van Dijk AD, Boelens R, Bonvin AM. 2005. *FEBS J.* 272:293–312
28. Vaynberg J, Qin J. 2006. *Trends Biotechnol.* 24:22–27
29. Tang C, Iwahara J, Clore GM. 2006. *Nature* 444:383–86
30. Burz DS, Dutta K, Cowburn D, Shekhtman A. 2006. *Nat. Methods* 3:91–93
31. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, et al. 2002. *Acta Crystallogr. D Biol. Crystallogr.* 58:899–907
32. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. 2000. *Annu. Rev. Biophys. Biomol. Struct.* 29:291–325
33. Davis FP, Braberg H, Shen MY, Pieper U, Sali A, Madhusudhan MS. 2006. *Nucleic Acids Res.* 34:2943–52
34. Shen MY, Sali A. 2006. *Protein Sci.* 15:2507–24
35. Levy ED, Pereira-Leal JB, Chothia C, Teichmann SA. 2006. *PLoS Comput. Biol.* 2:e155
36. Kuhlbrandt W, Williams KA. 1999. *Curr. Opin. Chem. Biol.* 3:537–43
37. Fujiyoshi Y. 1998. *Adv. Biophys.* 35:25–80

38. Frank J. 2006. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Oxford: Oxford Univ. Press
39. van Heel M, Gowen B, Matadeen R, Orlova EV, Finn R, et al. 2000. *Q. Rev. Biophys.* 33:307–69
40. Frank J, ed. 2006. *Electron Tomography: Methods for Three-dimensional Visualization of Structures in the Cell*. New York: Springer. 455 pp.
41. Lucic V, Förster F, Baumeister W. 2005. *Annu. Rev. Biochem.* 74:833–65
42. McIntosh R, Nicastro D, Mastronarde D. 2005. *Trends Cell Biol.* 15:43–51
43. Yonekura K, Maki-Yonekura S, Namba K. 2003. *Nature* 424:643–50
44. Fleishman SJ, Ben-Tal N. 2006. *Curr. Opin. Struct. Biol.* 16:496–504
45. Henderson R. 2004. *Q. Rev. Biophys.* 37:3–13
46. Mitra K, Frank J. 2006. *Annu. Rev. Biophys. Biomol. Struct.* 35:299–317
47. Chiu W, Baker ML, Jiang W, Dougherty M, Schmid MF. 2005. *Structure* 13:363–72
48. Rossmann MG, Morais MC, Leiman PG, Zhang W. 2005. *Structure* 13:355–62
49. Johnson JE, Chiu W. 2007. *Curr. Opin. Struct. Biol.* 17:237–43
50. Orlova EV, Saibil HR. 2004. *Curr. Opin. Struct. Biol.* 14:584–90
51. Tagari M, Newman R, Chagoyen M, Carazo JM, Henrick K. 2002. *Trends Biochem. Sci.* 27:589
52. Medalia O, Weber I, Frangakis AS, Nicastro D, Gerisch G, Baumeister W. 2002. *Science* 298:1209–13
53. Al-Amoudi A, Norlen LP, Dubochet J. 2004. *J. Struct. Biol.* 148:131–35
54. Grünewald K, Desai P, Winkler DC, Heymann JB, Belnap DM, et al. 2003. *Science* 302:1396–98
55. Förster F, Hegerl R. 2007. *Methods Cell Biol.* 79:741–67
56. Roux KH, Taylor KA. 2007. *Curr. Opin. Struct. Biol.* 17:244–52
57. Beck M, Förster F, Ecke M, Plitzko JM, Melchior F, et al. 2004. *Science* 306:1387–90
58. Stoffler D, Feja B, Fahrenkrog B, Walz J, Typke D, Aebi U. 2003. *J. Mol. Biol.* 328:119–30
59. Beck M, Lucic V, Förster F, Baumeister W, Medalia O. 2007. *Nature* 449: 611–15
60. Böhm J, Frangakis AS, Hegerl R, Nickell S, Typke D, Baumeister W. 2000. *Proc. Natl. Acad. Sci. USA* 97:14245–50
61. Frangakis AS, Bohm J, Förster F, Nickell S, Nicastro D, et al. October 2002. *Proc. Natl. Acad. Sci. USA* 99:14153–58
62. Ortiz JO, Förster F, Kürner J, Linaroudis AA, Baumeister W. 2006. *J. Struct. Biol.* 156:334–41
63. Koch MH, Vachette P, Svergun DI. 2003. *Q. Rev. Biophys.* 36:147–227
64. Nagar B, Kuriyan J. 2005. *Structure* 13:169–70
65. Sondermann H, Nagar B, Bar-Sagi D, Kuriyan J. 2005. *Proc. Natl. Acad. Sci. USA* 102:16632–37
66. Yamagata A, Tainer JA. 2007. *EMBO J.* 26:878–90
67. Alber F, Kim MF, Sali A. 2005. *Structure* 13:435–45
68. Parrish JR, Gulyas KD, Finley RL Jr. 2006. *Curr. Opin. Biotechnol.* 17:387–93
69. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, et al. 2005. *Nature* 437:1173–78
70. Michnick SW, Ear PH, Manderson EN, Remy I, Stefan E. 2007. *Nat. Rev. Drug Discov.* 6:569–82
71. Landgraf C, Panni S, Montecchi-Palazzi L, Castagnoli L, Schneider-Mergener J, et al. 2004. *PLoS Biol.* 2:E14
72. MacBeath G, Schreiber SL. 2000. *Science* 289:1760–63

73. Lakey JH, Raggett EM. 1998. *Curr. Opin. Struct. Biol.* 8:119–23
74. Nedelkov D, Nelson RW. 2003. *Trends Biotechnol.* 21:301–5
75. Piehler J. 2005. *Curr. Opin. Struct. Biol.* 15:4–14
76. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, et al. 2007. *Nature* 446:806–10
77. Aloy P, Russell RB. 2002. *Proc. Natl. Acad. Sci. USA* 99:5896–901
78. Fields S. 2005. *FEBS J.* 272:5391–99
79. Bauer A, Kuster B. 2003. *Eur. J. Biochem.* 270:570–78
80. Tackett AJ, Dilworth DJ, Davey MJ, O'Donnell M, Aitchison JD, et al. 2005. *J. Cell Biol.* 169:35–47
81. Krogan NJ, Peng WT, Cagney G, Robinson MD, Haw R, et al. 2004. *Mol. Cell* 13:225–39
82. Cristea IM, Williams R, Chait BT, Rout MP. 2005. *Mol. Cell Proteomics* 4:1933–41
83. Cristea IM, Carroll JW, Rout MP, Rice CM, Chait BT, MacDonald MR. 2006. *J. Biol. Chem.* 281:30269–78
84. Sharon M, Robinson CV. 2007. *Annu. Rev. Biochem.* 76:167–93
85. Sharon M, Taverner T, Ambroggio XI, Deshaies RJ, Robinson CV. 2006. *PLoS Biol.* 4:e267
86. Hainfeld JF, Powell RD. 2000. *J. Histochem. Cytochem.* 48:471–80
87. Rout MP, Aitchison JD, Suprapto A, Hjertaas K, Zhao Y, Chait BT. 2000. *J. Cell Biol.* 148:635–51
88. Pye VE, Beuron F, Keetch CA, McKeown C, Robinson CV, et al. 2007. *Proc. Natl. Acad. Sci. USA* 104:467–72
89. Martin-Benito J, Grantham J, Boskovic J, Brackley KI, Carrascosa JL, et al. 2007. *EMBO Rep.* 8:252–57
90. Golas MM, Sander B, Will CL, Luhrmann R, Stark H. 2005. *Mol. Cell* 17:869–83
91. Sivasubramanian A, Chao G, Pressler HM, Wittrup KD, Gray JJ. 2006. *Structure* 14:401–14
92. Mohd-Sarip A, van der Knaap JA, Wyman C, Kanaar R, Schedl P, Verrijzer CP. 2006. *Mol. Cell* 24:91–100
93. Guan JQ, Almo SC, Reisler E, Chance MR. 2003. *Biochemistry* 42:11992–2000
94. Anand GS, Law D, Mandell JG, Snead AN, Tsigelny I, et al. 2003. *Proc. Natl. Acad. Sci. USA* 100:13264–69
95. Trester-Zedlitz M, Kamada K, Burley SK, Fenyo D, Chait BT, Muir TW. 2003. *J. Am. Chem. Soc.* 125:2416–25
96. Seebacher J, Mallick P, Zhang N, Eddes JS, Aebersold R, Gelb MH. 2006. *J. Proteome Res.* 5:2270–82
97. Sinz A. 2006. *Mass Spectrom. Rev.* 25:663–82
98. Muller EG, Snydsman BE, Novik I, Hailey DW, Gestaut DR, et al. 2005. *Mol. Biol. Cell* 16:3341–52
99. Truong K, Ikura M. 2001. *Curr. Opin. Struct. Biol.* 11:573–78
100. Yan Y, Marriott G. 2003. *Curr. Opin. Chem. Biol.* 7:635–40
101. Bhatnagar J, Freed JH, Crane BR. 2007. *Methods Enzymol.* 423:117–33
102. Park SY, Borbat PP, Gonzalez-Bonet G, Bhatnagar J, Pollard AM, et al. 2006. *Nat. Struct. Mol. Biol.* 13:400–7
103. Aloy P, Bottcher B, Ceulemans H, Leutwein C, Mellwig C, et al. 2004. *Science* 303:2026–29
104. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, et al. 2006. *Nucleic Acids Res.* 34:D291–95
105. Cockell SJ, Oliva B, Jackson RM. 2007. *Bioinformatics* 23:573–81

106. Aloy P, Ceulemans H, Stark A, Russell RB. 2003. *J. Mol. Biol.* 332:989–98
107. Lu L, Lu H, Skolnick J. 2002. *Proteins* 49:350–64
108. Lu L, Arakaki AK, Lu H, Skolnick J. 2003. *Genome Res.* 13:1146–54
109. Henrick K, Thornton JM. 1998. *Trends Biochem. Sci.* 23:358–61
110. Tovchigrechko A, Wells CA, Vakser IA. 2002. *Protein Sci.* 11:1888–96
111. Bordner AJ, Gorin AA. 2007. *Proteins* 68:488–502
112. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. 2005. *Proteins* 60:224–31
113. Fernandez-Recio J, Abagyan R, Totrov M. 2005. *Proteins* 60:308–13
114. Wiehe K, Pierce B, Mintseris J, Tong WW, Anderson R, et al. 2005. *Proteins* 60:207–13
115. Kozakov D, Brenke R, Comeau SR, Vajda S. 2006. *Proteins* 65:392–406
116. Inbar Y, Benyamini H, Nussinov R, Wolfson HJ. 2005. *Phys. Biol.* 2:S156–65
117. Man-Kuang Cheng T, Blundell TL, Fernandez-Recio J. 2007. *Proteins* 68:503–15
118. Mendez R, Leplae R, Lensink MF, Wodak SJ. 2005. *Proteins* 60:150–69
119. Kowalsman N, Eisenstein M. 2007. *Bioinformatics* 23:421–26
120. Tobi D, Bahar I. 2005. *Proc. Natl. Acad. Sci. USA* 102:18908–13
121. McKenna S, Moraes T, Pastushok L, Ptak C, Xiao W, et al. 2003. *J. Biol. Chem.* 278:13151–58
122. Walters KJ, Lech PJ, Goh AM, Wang Q, Howley PM. 2003. *Proc. Natl. Acad. Sci. USA* 100:12694–99
123. Law D, Hotchko M, Ten Eyck L. 2005. *Proteins* 60:302–7
124. Chu F, Shan SO, Moustakas DT, Alber F, Egea PF, et al. 2004. *Proc. Natl. Acad. Sci. USA* 101:16454–59
125. Schulz DM, Ihling C, Clore GM, Sinz A. 2004. *Biochemistry* 43:4703–15
126. Matsuda T, Ikegami T, Nakajima N, Yamazaki T, Nakamura H. 2004. *J. Biomol. NMR* 29:325–38
127. Gaboriaud C, Juanhuix J, Gruez A, Lacroix M, Darnault C, et al. 2003. *J. Biol. Chem.* 278:46974–82
128. Azuma Y, Renault L, Garcia-Ranea JA, Valencia A, Nishimoto T, Wittinghofer A. 1999. *J. Mol. Biol.* 289:1119–30
129. Sachchidanand, Lequin O, Staunton D, Mulloy B, Forster MJ, et al. 2002. *J. Biol. Chem.* 277:50629–35
130. Dobrodumov A, Gronenborn AM. 2003. *Proteins* 53:18–32
131. Dominguez C, Bonvin AM, Winkler GS, van Schaik FM, Timmers HT, Boelens R. 2004. *Structure* 12:633–44
132. Eriksson MA, Roux B. 2002. *Biophys. J.* 83:2595–609
133. Tomaselli S, Ragona L, Zetta L, Assfalg M, Ferranti P, et al. 2007. *Proteins* 69:177–91
134. Ben-Zeev E, Zarivach R, Shoham M, Yonath A, Eisenstein M. 2003. *J. Biomol. Struct. Dyn.* 20:669–76
135. Ben-Zeev E, Eisenstein M. 2003. *Proteins* 52:24–27
136. Fahmy A, Wagner G. 2002. *J. Am. Chem. Soc.* 124:1241–50
137. Korkin D, Davis FP, Sali A. 2005. *Protein Sci.* 14:2350–60
138. Korkin D, Davis FP, Alber F, Luong T, Shen MY, et al. 2006. *PLoS Comput. Biol.* 2:e153
139. Davis FP, Sali A. 2005. *Bioinformatics* 21:1901–7
140. Goddard TD, Huang CC, Ferrin TE. 2007. *J. Struct. Biol.* 157:281–87
141. Frangakis AS, Förster F. 2004. *Curr. Opin. Struct. Biol.* 14:325–31
142. Baker ML, Yu Z, Chiu W, Bajaj C. 2006. *J. Struct. Biol.* 156:432–41
143. Zhou ZH, Baker ML, Jiang W, Dougherty M, Jakana J, et al. 2001. *Nat. Struct. Biol.* 8:868–73

144. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, et al. 2006. In *Current Protocols Bioinformatics*, ed. AD Baxevanis, GA Petsko, LD Stein, GD Stormo, Suppl. 15:5.6.1–30. New York: Wiley
145. Fabiola F, Chapman MS. 2005. *Structure* 13:389–400
146. Topf M, Baker ML, John B, Chiu W, Sali A. 2005. *J. Struct. Biol.* 149:191–203
147. Wriggers W, Milligan RA, McCammon JA. 1999. *J. Struct. Biol.* 125:185–95
148. Volkmann N, Hanein D. 1999. *J. Struct. Biol.* 125:176–84
149. Rossmann MG. 2000. *Acta Crystallogr. D Biol. Crystallogr.* 56(Part 10):1341–49
150. Roseman AM. 2000. *Acta Crystallogr. D Biol. Crystallogr.* 56:1332–40
151. Jiang W, Baker ML, Ludtke SJ, Chiu W. 2001. *J. Mol. Biol.* 308:1033–44
152. Navaza J, Lepault J, Rey FA, Alvarez-Rua C, Borge J. 2002. *Acta Crystallogr. D Biol. Crystallogr.* 58:1820–25
153. Wu X, Milne JL, Borgnia MJ, Rostapshov AV, Subramaniam S, Brooks BR. 2003. *J. Struct. Biol.* 141:63–76
153a. Garzón JI, Kovacs J, Abagyan R, Chacón P. 2007. *Bioinformatics* 23:427–33
154. Halperin I, Ma B, Wolfson H, Nussinov R. 2002. *Proteins* 47:409–43
155. Baker D, Sali A. 2001. *Science* 294:93–96
156. Velazquez-Muriel JA, Sorzano CO, Scheres SH, Carazo JM. 2005. *J. Mol. Biol.* 345:759–71
157. Velazquez-Muriel JA, Valle M, Santamaria-Pang A, Kakadiaris IA, Carazo JM. 2006. *Structure* 14:1115–26
158. Baker ML, Jiang W, Wedemeyer WJ, Rixon FJ, Baker D, Chiu W. 2006. *PLoS Comput. Biol.* 2:e146
159. Chen JZ, Furst J, Chapman MS, Grigorieff N. 2003. *J. Struct. Biol.* 144:144–51
160. Tama F, Brooks CL. 2006. *Annu. Rev. Biophys. Biomol. Struct.* 35:115–33
161. Topf M, Baker ML, Marti-Renom MA, Chiu W, Sali A. 2006. *J. Mol. Biol.* 357:1655–68
162. Frangakis AS, Rath BK. 2006. In *Electron Tomography*, ed. J Frank, pp. 401–16. New York: Springer Verlag
163. Roseman AM. 2003. *Ultramicroscopy* 94:225–36
164. Rath BK, Hegerl R, Leith A, Shaikh TR, Wagenknecht T, Frank J. 2003. *J. Struct. Biol.* 144:95–103
165. Zheng W, Doniach S. 2005. *Protein. Eng. Des. Sel.* 18:209–19
166. Shih AY, Denisov IG, Phillips JC, Sligar SG, Schulten K. 2005. *Biophys. J.* 88:548–56
167. Stuhrmann H. 1970. *Acta Crystallogr. A* 26:297–306
167a. Krukenberg KA, Förster F, Rice LM, Sali A, Agard DA. 2008. *Structure*. In press
168. Chacon P, Moran F, Diaz JF, Pantos E, Andreu JM. 1998. *Biophys. J.* 74:2760–75
169. Walther D, Cohen FE, Doniach S. 2000 . *J. Appl. Crystallogr.* 33:350–63
170. Svergun DI. 1999. *Biophys. J.* 76:2879–86
171. Svergun DI, Petoukhov MV, Koch MH. 2001. *Biophys. J.* 80:2946–53
172. Petoukhov MV, Eady NA, Brown KA, Svergun DI. 2002. *Biophys. J.* 83:3113–25
173. Petoukhov MV, Svergun DI. 2005. *Biophys. J.* 89:1237–50
174. Bernado P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI. 2007. *J. Am. Chem. Soc.* 129:5656–64
175. Tidow H, Melero R, Mylonas E, Freund SM, Grossmann JG, et al. 2007. *Proc. Natl. Acad. Sci. USA* 104:12324–29
176. Wu Y, Tian X, Lu M, Chen M, Wang Q, Ma J. 2005. *Structure* 13:1587–97
177. Grishaev A, Wu J, Trewhella J, Bax A. 2005. *J. Am. Chem. Soc.* 127:16621–28
178. Phizicky E, Bastiaens PI, Zhu H, Snyder M, Fields S. 2003. *Nature* 422:208–15

179. Valencia A, Pazos F. 2002. *Curr. Opin. Struct. Biol.* 12:368–73
180. Brunger AT. 1993. *Acta Crystallogr. D Biol. Crystallogr.* 49:24–36
181. Lim RY, Fahrenkrog B. 2006. *Curr. Opin. Cell Biol.* 18:342–47
182. Yang Q, Rout MP, Akey CW. 1998. *Mol. Cell* 1:223–34
183. Devos D, Dokudovskaya S, Williams R, Alber F, Eswar N, et al. 2006. *Proc. Natl. Acad. Sci. USA* 103:2172–77
184. Devos D, Dokudovskaya S, Alber F, Williams R, Chait BT, et al. 2004. *PLoS Biol.* 2:e380
185. Tama F, Miyashita O, Brooks CL 3rd. 2004. *J. Mol. Biol.* 337:985–99
186. Chandramouli P, Topf M, Ménétret J, Eswar N, Gutell R, et al. 2008. *Structure.* In press
187. Topf M, Lasker K, Webb B, Wolfson H, Chiu W, Sali A. 2008. *Structure* 16:295–307

Annual Review of Biochemistry

Volume 77, 2008

Contents

Prefatory Chapters

Single-Molecule Theme

Recent Advances in Biochemistry

Indexes

Errata

An online log of corrections to *Annual Review of Biochemistry* articles may be found at http://biochem.annualreviews.org/errata.shtml