Chapter 6
Integrative Structure Determination of Protein Assemblies by Satisfaction of Spatial Restraints

Frank Alber, Brian T. Chait, Michael P. Rout, and Andrej Sali

Abstract To understand the cell, we need to determine the structures of macromolecular assemblies, many of which consist of tens to hundreds of components. A great variety of experimental data can be used to characterize the assemblies at several levels of resolution, from atomic structures to component configurations. To maximize completeness, resolution, accuracy, precision and efficiency of the structure determination, a computational approach is needed that can use 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. We illustrate the approach by determining the configuration of the 456 proteins in the nuclear pore complex from Baker's yeast.

6.1 Introduction

Assemblies as functional modules of the cell. Macromolecular assemblies consist of non-covalently interacting macromolecular components, such as proteins and nucleic acids. They vary widely in size and play crucial roles in most cellular processes (Alberts 1998). 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 (Alber et al. 2008; Robinson et al. 2007; Sali 2003; Sali et al. 2003; Sali and Kuriyan 1999). Even a coarse characterization of the configuration of macromolecular components in a complex (Fig. 6.1) helps to elucidate the principles that

F. Alber
Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089-1340, USA
e-mail: alber@usc.edu

A. Panchenko, T. Przytycka (eds.), *Protein-protein Interactions and Networks*,
DOI: 10.1007/978-1-84800-125-1_6, © Springer-Verlag London Limited 2008

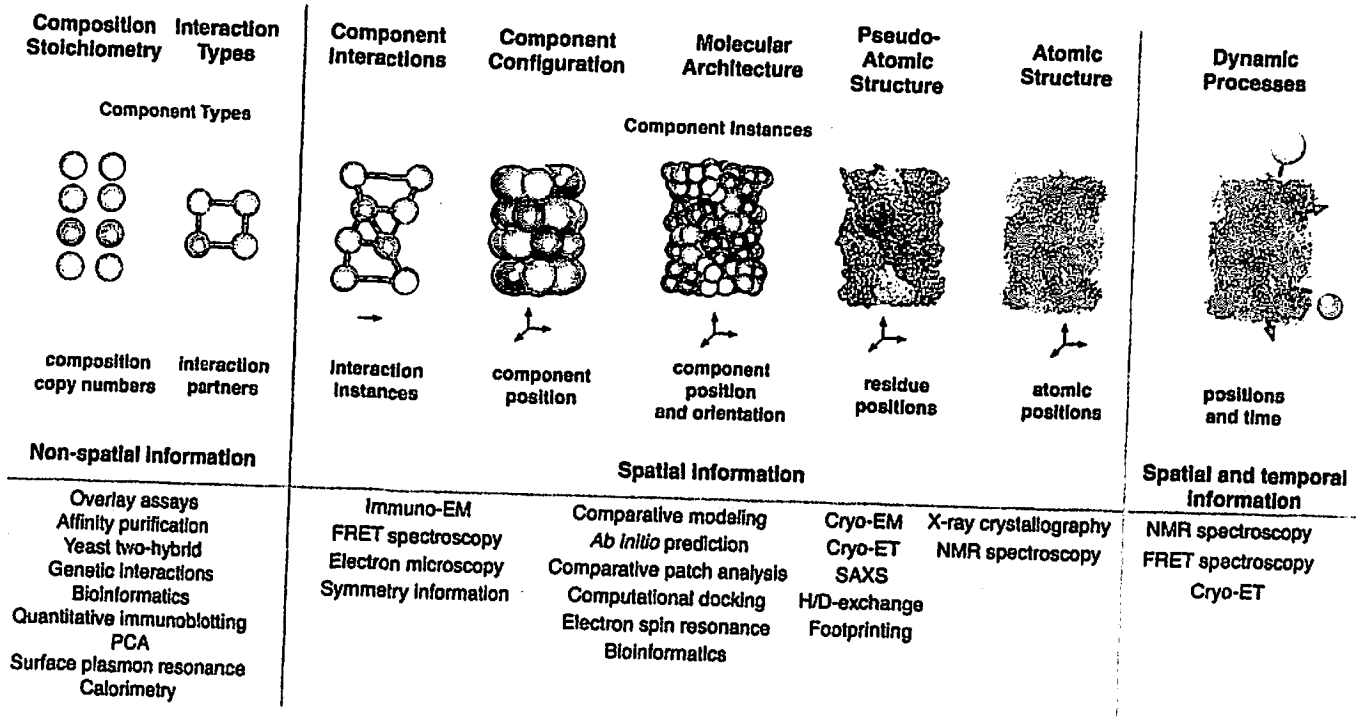

Fig. 6.1 Structural information about an assembly (Alber et al. 2008). Varied experimental methods can determine the copy numbers (stoichiometry) and types (composition) of the components (whether or not components interact with each other), positions of the components, and their relative orientations. 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. PCA, protein-fragment complementation assay; H/D, hydrogen/deuterium; EM, electron microscopy; FRET, fluorescence resonance energy transfer; SAXS, small angle X-ray scattering

underlie cellular processes, in addition to providing a necessary starting point for a higher resolution description.

Scope. Complete lists of macromolecular components of biological systems are becoming available (Aebersold and Mann 2003). However, the identification of complexes between these components is a non-trivial task. This difficulty arises partly from the multitude of component types and the varying lifespan of the complexes (Russell et al. 2004). The most comprehensive information about binary protein interactions is available for the *Saccharomyces cerevisiae* proteome, consisting of ~6,200 proteins. This data has been generated by methods such as the yeast two-hybrid system (Ito et al. 2000; Uetz et al. 2000) and affinity purifications coupled with mass-spectrometry (Collins et al. 2007; Gavin et al. 2006; Krogan et al. 2006). The lower bound on binary protein interactions in yeast has been estimated to be ~30,000 (Russell et al. 2004), corresponding to the average of ~9 protein partners per protein, though not necessarily all at the same time. The number of higher order complexes in yeast is estimated to be ~800, based on affinity purification experiments (Collins et al. 2007; Devos and Russell 2007; Gavin et al. 2006; Krogan et al. 2006). 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 transient interactions are yet to be characterized (Abbott 2002; Alberts 1998).

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 (Alber et al. 2008; Robinson et al. 2007; Sali et al. 2003; Sali and Kuriyan 1999). For example, X-ray crystallography is limited by the ability to grow suitable crystals and to build 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 (Fig. 6.1) (Alber et al. 2007a; Alber et al. 2004; Alber et al. 2008; Harris et al. 1994; Malhotra and Harvey 1994; Robinson et al. 2007; Sali et al. 2003). 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 this data integration.

Review outline. We begin by listing 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. Next, 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 based on any of the individual types of data alone (Alber et al. 2007a; Alber et al. 2008; Robinson et al. 2007; Sali et al. 2003). Finally, we illustrate this approach by its application to the determination of the configuration of 456 proteins in the yeast NPC (Alber et al. 2007a; Alber et al. 2007b).

6.2 Sources of Spatial Information

Different experimental methods produce different types of structural information (Fig. 6.1). This information varies in terms of what 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 yeast two-hybrid system and affinity purification. Relative orientations of components and information about interacting residues can be inferred from cryo-EM, hydrogen/deuterium (H/D) exchange, OH radical footprinting, and chemical-crosslinking. 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 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), pseudo-atomic 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).

6.3 Comprehensive Data Integration by Satisfaction of Spatial Restraints

The experimental data about a structure must be converted to an explicit structural model through computation. Importantly, these computational methods differ in the type of information they use to calculate the assembly structures, rather than how they calculate them once the information is specified.

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 to maximize efficiency, resolution, accuracy, precision, and completeness of the structural coverage of macromolecular assemblies.

In this section, we describe the underlying theory and methods of our hybrid approach to characterizing macromolecular assembly structures. A sample application is provided by the structure determination of the NPC (Alber et al. 2007a; Alber et al. 2007b; Alber et al. 2004; Alber et al. 2005; Sali and Kuriyan 1999) (below).

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, 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 (Fig. 6.2). In this view, models that are consistent with the input

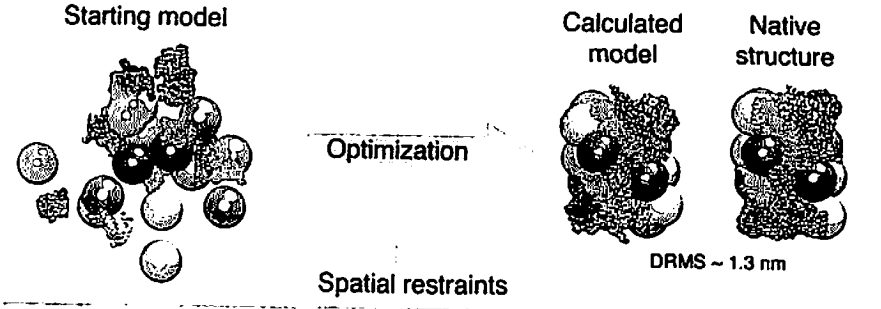

Fig. 6.2 Characterization of an assembly configuration based on data simulated from a known native structure (Alber et al. 2008). In this approach (Alber et al. 2005), the simulated data include protein positions (e.g., from immuno-EM), assembly shape (e.g., from EM), relative proximity of components (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 was 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. DRMS, distance root-mean-square difference between the protein centroids in the determined model and the native structure

information are calculated by optimizing a scoring function. The three components of this approach are (i) a representation of the modeled assembly, (ii) a scoring function consisting of the individual spatial restraints, and (iii) 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 (Fig. 6.2). For a protein assembly, the hierarchy can include atoms, atomic groups, amino acid residues, secondary structure segments, domains, proteins, protein sub-complexes, 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 (Fig. 6.2). 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 our knowledge of any kind as a scoring function whose global optimum corresponds to the native assembly structure (Shen and Sali 2006). 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, $p(C|I)$. $C = (c_1, c_2, \ldots, c_n)$ is the list of the Cartesian coordinates (c_1) 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 information 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, while 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 (Frank 2006); upper distance bounds on residues from different proteins may be obtained from NMR spectroscopy (Fiaux et al. 2002) and chemical cross-linking (Trester-Zedlitz et al. 2003); protein-protein interactions may be discovered by the yeast two-hybrid system (Phizicky et al. 2003) and calorimetry (Lakey and Raggett 1998); two proteins can be assigned to be in proximity if they are part of an isolated sub-complex identified by affinity purification in combination with mass spectrometry (Bauer and Kuster 2003). Increasingly, important restraints will be derived from pairwise molecular docking (Mendez et al. 2005), statistical preferences observed in the structurally defined protein-protein interactions (Davis and Sali 2005), and analysis of multiple sequence alignments (Valencia and Pazos 2002).

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 (Fig. 6.3). The selection of the best alternative interpretation is then achieved as part of the structure optimization process.

Figure 6.3 shows a conditional restraint that encodes protein contacts consistent with an affinity purification experiment (Alber et al. 2007a; Alber et al. 2008; Alber et al. 2005) (Fig. 6.3). In this example, affinity purification identified 4 protein types (yellow, blue, red, green), derived from an assembly containing a single copy of the yellow, blue, and red protein and two copies of the green protein. The sample affinity purification implies that at least 3 of the following 6 possible types of interaction 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 (Fig. 6.3); (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

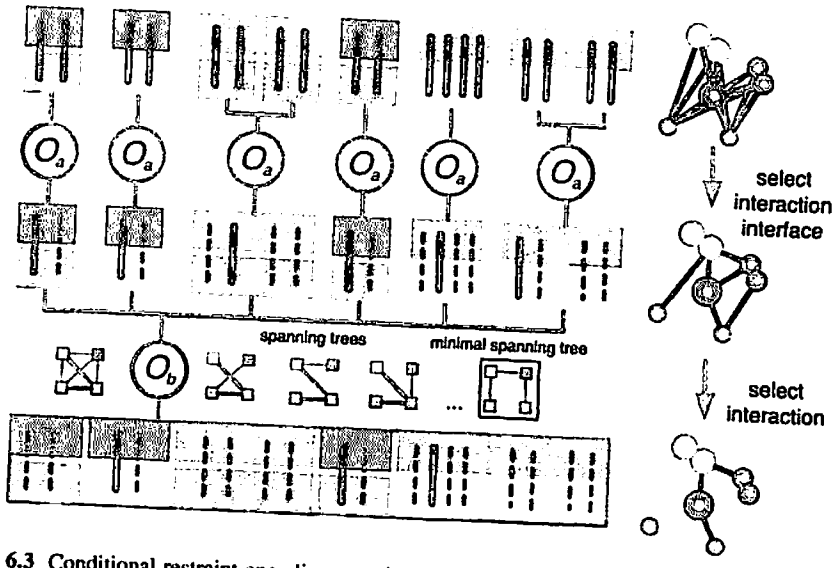

Fig. 6.3 Conditional restraint encoding protein contacts based on an affinity purification experiment that identified 4 protein types (*yellow*, *blue*, *red*, *green*), derived from an assembly containing a single copy of the *yellow*, *blue*, and *red* protein and two copies of the *green* protein (Alber et al. 2007a; Alber et al. 2008; Alber et al. 2005). A single protein is represented by either one bead (*blue* and *green* proteins) or two beads (*yellow* and *red* proteins) (column on the *right*); alternative interactions between proteins are indicated by different edges. Protein contacts are selected in a decision tree-like evaluation process by operator functions O_a and O_b (*left panel*) (see text for a detailed description). *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". 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* to the O_b operator); edge weights correspond to the violations of interaction restraints and quantify how consistent is the corresponding interaction 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)

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 5 restraints (thick red line) at the middle level of the tree (Fig. 6.3). Finally, a minimal spanning tree operator (O_b) finds the minimal spanning tree corresponding to the combination of 3 restraints that are most consistent with the affinity purification (thick red lines in Fig. 6.3). The whole restraint evaluation process is executed at each optimization step based on the current configuration, thus resulting in possibly different subsets of selected optional restraints at each optimization step.

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 our program 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 and 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 real 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: (i) self-consistency of independent experimental data; (ii) structural similarity among all configurations in the ensemble that satisfy the input restraints; (iii) simulations where a native structure is assumed, corresponding restraints simulated from it, and the resulting calculated structure compared with the assumed native structure; (iv) confirmatory spatial data that were not used in the calculation of the structure (e.g., criterion similar to the crystallographic free R-factor (Brunger 1993) can be used to assess both

the model accuracy and the harmony among the input restraints); and (v) patterns emerging from a mapping of independent and unused data on the structure that are unlikely to occur by chance (Alber et al. 2007a; Alber et al. 2007b).

Advantages. The integrative approach to structure determination has several advantages: (i) 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); (ii) it can potentially produce all structures that are consistent with the data, not just one; (iii) the variation among the structures consistent with the data allows us to assess sufficiency of the data and the precision of the representative structure; (iv) it can make the process of structure determination more efficient by indicating what measurements would be most informative.

6.4 Structural Characterization of the Nuclear Pore Complex

Using the approach outlined above, we determined the native configuration of proteins in the yeast nuclear pore complex (NPC) (Alber et al. 2007a; Alber et al. 2007b). NPCs are large (~50 MDa) proteinaceous assemblies spanning the nuclear envelope (NE), where they function as the sole mediators of bidirectional exchange between the nucleoplasmic and cytoplasmic compartments in all eukaryotes (Lim and Fahrenkrog 2006). 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 8 radially arranged "half-spoke" units (Yang et al. 1998). The yeast NPC is build from multiple copies of 30 different proteins, totaling approximately 456 proteins (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 was unknown. A description of the NPC's structure was needed to understand its function and assembly, and to provide clues to its evolutionary origins. Due 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 (Devos et al. 2006).

To determine the protein configuration of the NPC, we collected a large and diverse set of biophysical and biochemical data. The data was derived from six experimental sources (Fig. 6.4): (i) Quantitative immuno-blotting experiments determined the stoichiometry of all 30 nups in the NPC; (ii) hydrodynamics experiments provided information about the approximate excluded volume and the coarse shape of each nup; (iii) immuno-EM provided a coarse localization for each nup along two principal axes of the NPC; (iv) an exhaustive set of affinity purification experiments determined the composition of 77 NPC complexes; (v) overlay experiments determined 5 direct binary nup interactions; and (vi) symmetry considerations and the dimensions of the NE were extracted from cryo-EM. Moreover,

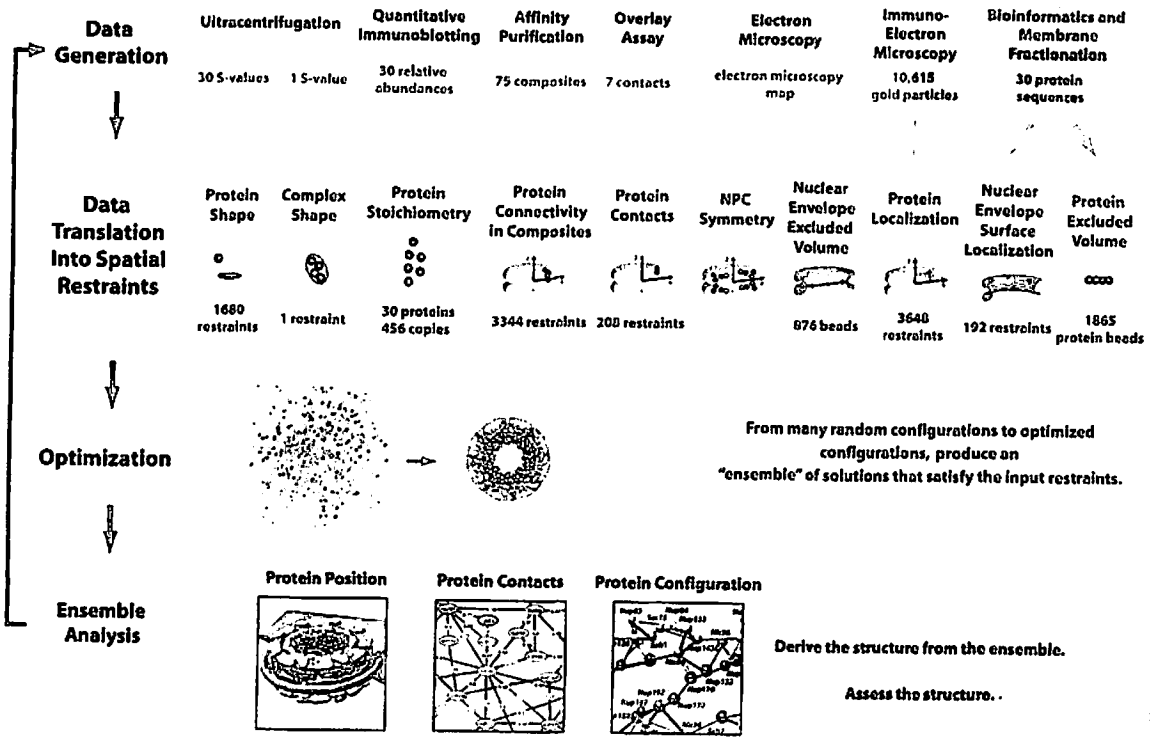

Fig. 6.4 Determining the architecture of the NPC by integrating spatial restraints from proteomic data (Alber et al. 2007a). First, structural data (*red*) are generated by various experiments (*black*). Second, the data and theoretical considerations are expressed as spatial restraints (*blue*). Third, 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, the ensemble is clustered into sets of distinct solutions as well as analyzed in terms of protein positions, contacts, and configuration

bioinformatics analysis provided information about the position of transmembrane helices for the three integral membrane nups. This data was translated into spatial restraints on the NPC (Fig. 6.4).

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 Fig. 6.5. The 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 (Fig. 6.5). To comprehensively sample all possible structural solutions that are consistent with the data, we obtained an "ensemble" of 1,000 independently calculated structures that satisfied the input restraints (Fig. 6.5c). 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 single pronounced maxima for almost all proteins,

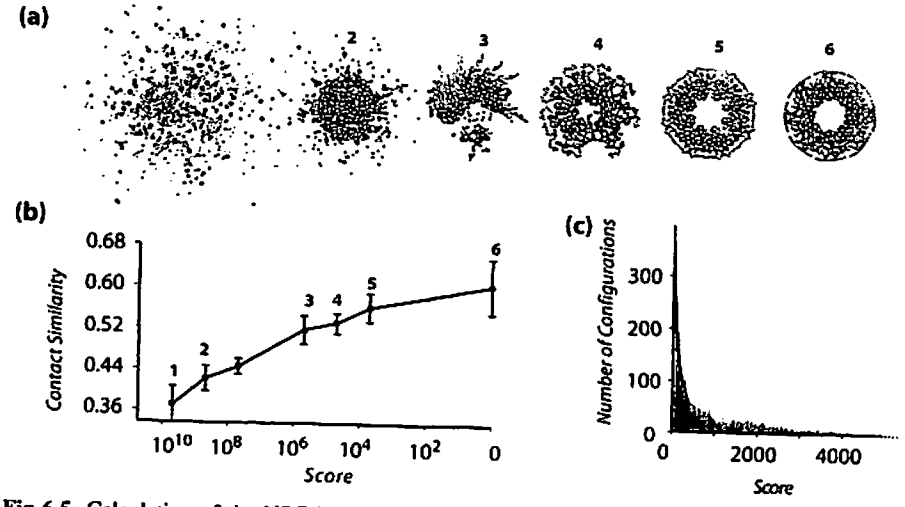

Fig 6.5 Calculation of the NPC bead structure by satisfaction of spatial restraints (Alber et al. 2007a; Alber et al. 2008). (a), Representation of the optimization process as it progresses from an initial random configuration to an optimal solution. (b), The graph shows the relationship between the 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 the 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 satisfy 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

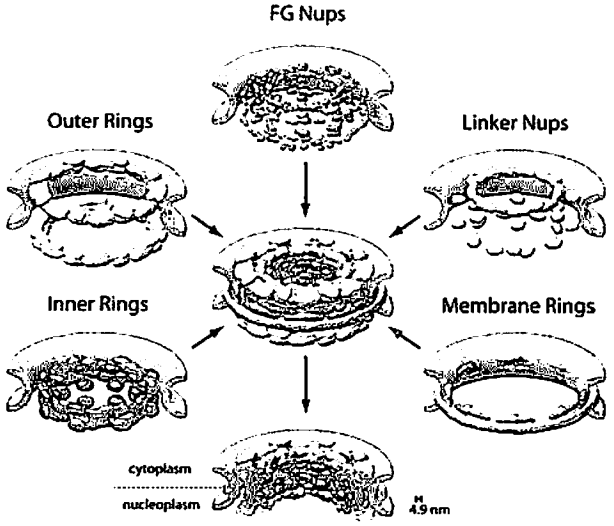

Fig. 6.6 Localization of major substructures and their component proteins in the NPC (Alber et al. 2007b; Alber et al. 2008). The proteins are represented by their localization volumes (Alber et al. 2007a) and have been colored according to their classification into five distinct substructures based on 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*

demonstrating that the input restraints define a single NPC architecture (Fig. 6.6). The average standard deviation for the separation between neighbouring 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 (Fig. 6.7).

Our structure (Fig. 6.5) 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 of its constituent proteins (Devos 2004, 2006) can provide clues to the ancient evolutionary origins of the NPC.

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

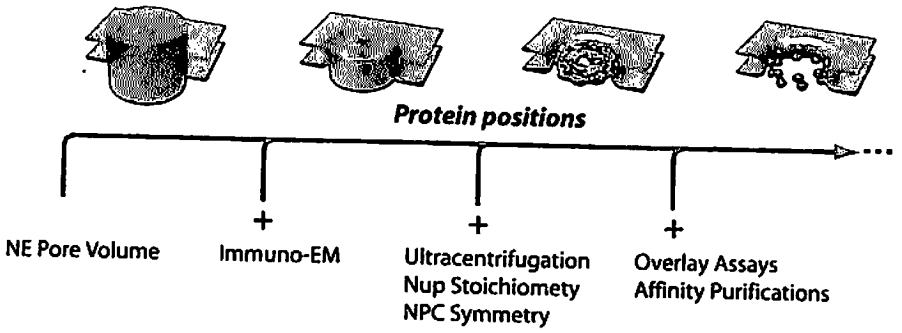

Fig. 6.7 Synergy between varied datasets results into increased precision of structure determination (Alber et al. 2007a; Alber et al. 2008; Robinson et al. 2007). The proteins are increasingly localized by the addition of different types of synergistic experimental information. As an example, each panel illustrates the localization of 16 copies of Nup192 in the ensemble of 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

6.5 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 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.

Acknowledgments We are grateful to Wah Chiu, David Agard, Wolfgang Baumeister, Joachim Frank, Fred Davis, M.S. Madhusudan, Min-yi Shen, Keren Lasker, Daniel Russell, Friedrich Foerster, Dmitry Korkin, Maya Topf and Ben Webb for many discussions about structure characterization by satisfaction of spatial restraints. We are also thankful to Svetlana Dokdovskaya, 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. We 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. This review is based on refs. (Alber et al. 2007a; Alber et al. 2007b; Alber et al. 2008; Robinson et al. 2007).

References

Abbott, A. (2002). The society of proteins. Nature, *417*(6892), 894–896.

Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. Nature, *422*(6928), 198–207.

Alber, F., Dokudovskaya, S., Veenhoff, L., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Rout, M. P. and Sali, A. (2007a). Determining the architectures of macromolecular Assemblies. Nature, *450*(7170), 683–694.

Alber, F., Dokudovskaya, S., Veenhoff, L., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B. T., Sali, A. and Rout, M. P. (2007b). The molecular architecture of the nuclear pore complex. Nature, *450*(7170), 695–701.

Alber, F., Eswar, N. and Sali, A. (2004). Structure determination of macromolecular complexes by experiment and computation. In J. M. Bujnicki (Ed.), *Practical Bioinformatics* (pp. 73–96). Germany: Springer-Verlag.

Alber, F., Foerster, F., Korkin, D., Topf, M. and Sali, A. (2008). Integrating diverse data for structure determination of macromolecular assemblies. Ann Rev Biochem, *77*, in press.

Alber, F., Kim, M. F. and Sali, A. (2005). Structural characterization of assemblies from overall shape and subcomplex compositions. Structure, *13*(3), 435–445.

Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell, *92*(3), 291–294.

Bauer, A. and Kuster, B. (2003). Affinity purification-mass spectrometry. Powerful tools for the characterization of protein complexes. Eur J Biochem, *270*(4), 570–578.

Brunger, A. T. (1993). Assessment of phase accuracy by cross validation: the free R value. Methods and applications. Acta Crystallogr D Biol Crystallogr, *49*(Pt 1), 24–36.

Collins, S. R., Kemmeren, P., Zhao, X. C., Greenblatt, J. F., Spencer, F., Holstege, F. C., Weissman, J. S. and Krogan, N. J. (2007). Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. Mol Cell Proteomics, *6*(3), 439–450.

Davis, F. P. and Sali, A. (2005). PIBASE: a comprehensive database of structurally defined protein interfaces. Bioinformatics, *21*(9), 1901–1907.

Devos, D., Dokudovskaya, S., Williams, R., Alber, F., Eswar, N., Chait, B. T., Rout, M. P. and Sali, A. (2006). Simple fold composition and modular architecture of the nuclear pore complex. Proc Natl Acad Sci U S A, *103*(7), 2172–2177.

Devos, D., Dokudovskaya, S., Williams, S., Alber, F., Williams, R., Chait, B. T., Rout, M. P. Sali, A. (2004) Components of coated vesicles and nuclear pore complexes share a commmon molecular architecture. PLS Biol,2(12), e380.

Devos, D. and Russell, R. B. (2007). A more complete, complexed and structured interactome. Curr Opin Struct Biol, *17*(3), 370–377.

Fiaux, J., Bertelsen, E. B., Horwich, A. L. and Wuthrich, K. (2002). NMR analysis of a 900 K GroEL GroES complex. Nature, *418*(6894), 207–211.

Frank, J. (2006). *Three-Dimensional Electron Microscopy of Macromolecular Assemblies*. Oxford: Oxford University Press.

Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M. A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A. M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B. and Superti-Furga, G. (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature, *440*(7084), 631–636.

Harris, M. E., Nolan, J. M., Malhotra, A., Brown, J. W., Harvey, S. C. and Pace, N. R. (1994). Use of photoaffinity crosslinking and molecular modeling to analyze the global architecture of ribonuclease P RNA. Embo J, *13*(17), 3953–3963.

Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S. and Sakaki, Y. (2000). Toward a protein-protein interaction map of the budding yeast:

A comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins. Proc Natl Acad Sci U S A, *97*(3), 1143–1147.

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrin-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., Beattie, B., Richards, D. P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M. M., Vlasblom, J., Wu, S., Orsi, C., Collins, S. R., Chandran, S., Haw, R., Rilstone, J. J., Gandi, K., Thompson, N. J., Musso, G., St Onge, P., Ghanny, S., Lam, M. H., Butland, G., Altaf-Ul, A. M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J. S., Ingles, C. J., Hughes, T. R., Parkinson, J., Gerstein, M., Wodak, S. J., Emili, A. and Greenblatt, J. F. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature, *440*(7084), 637–643.

Lakey, J. H. and Raggett, E. M. (1998). Measuring protein-protein interactions. Curr Opin Struct Biol, *8*(1), 119–123.

Lim, R. Y. and Fahrenkrog, B. (2006). The nuclear pore complex up close. Curr Opin Cell Biol, *18*(3), 342–347.

Malhotra, A. and Harvey, S. C. (1994). A quantitative model of the Escherichia coli 16 S RNA in the 30 S ribosomal subunit. J Mol Biol, *240*(4), 308–340.

Mendez, R., Leplae, R., Lensink, M. F. and Wodak, S. J. (2005). Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. Proteins, *60*(2), 150–169.

Phizicky, E., Bastiaens, P. I., Zhu, H., Snyder, M. and Fields, S. (2003). Protein analysis on a proteomic scale. Nature, *422*(6928), 208–215.

Robinson, C., Sali, A. and Baumeister, W. (2007). The molecular sociology of the cell. Nature, *450*(7172), 973–982.

Russell, R. B., Alber, F., Aloy, P., Davis, F. P., Korkin, D., Pichaud, M., Topf, M. and Sali, A. (2004). A structural perspective on protein-protein interactions. Curr Opin Struct Biol, *14*(3), 313–324.

Sali, A. (2003). NIH workshop on structural proteomics of biological complexes. Structure, *11*(9), 1043–1047.

Sali, A., Glaeser, R., Earnest, T. and Baumeister, W. (2003). From words to literature in structural proteomics. Nature, *422*(6928), 216–225.

Sali, A. and Kuriyan, J. (1999). Challenges at the frontiers of structural biology. Trends Cell Biol, *9*(12), M20–24.

Shen, M. Y. and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. Protein Sci, *15*(11), 2507–2524.

Trester-Zedlitz, M., Kamada, K., Burley, S. K., Fenyo, D., Chait, B. T. and Muir, T. W. (2003). A modular cross-linking approach for exploring protein interactions. J Am Chem Soc, *125*(9), 2416–2425.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature, *403*(6770), 623–627.

Valencia, A. and Pazos, F. (2002). Computational methods for the prediction of protein interactions. Curr Opin Struct Biol, *12*(3), 368–373.

Yang, Q., Rout, M. P. and Akey, C. W. (1998). Three-dimensional architecture of the isolated yeast nuclear pore complex: functional and evolutionary implications. Mol Cell, *1*(2), 223–234.