Structure, Vol. 13, 435–445, March, 2005, ©2005 Elsevier Ltd All rights reserved. DOI 10.1016/j.str.2005.01.013

Structural Characterization of Assemblies from Overall Shape and Subcomplex Compositions

Frank Alber,[1] Michael F. Kim,[1,2] and Andrej Sali[1,3,*]
[1]Department of Biopharmaceutical Sciences and Department of Pharmaceutical Chemistry and California Institute for Quantitative Biomedical Research
[2]Graduate Program in Biological and Medical Informatics
University of California, San Francisco
QB3
1700 4th Street
San Francisco, California 94143

Summary

We suggest structure characterization of macromolecular assemblies by combining assembly shape determined by electron cyromicroscopy with information about subunit proximity determined by affinity purification. To achieve this aim, structure characterization is expressed as a problem in satisfaction of spatial restraints that (1) represents subunits as spheres, (2) encodes information about the subunit excluded volume, assembly shape, and pulldowns in a scoring function, and (3) finds subunit configurations that satisfy the input restraints by an optimization of the scoring function. Testing of the approach with model systems suggests its feasibility.

Introduction

The structures of a number of large assemblies are being solved at atomic resolution primarily by X-ray crystallography (Ban et al., 2000; Carter et al., 2000; Harms et al., 2001; Zhang et al., 1999) or at lower resolution by electron cryomicroscopy (Frank, 2002; Schmid et al., 2004; Zhang et al., 2003; Yang et al., 1998) and tomography (Baumeister et al., 1999; Beck et al., 2004). Although the atomic structures are more informative, even a low-resolution configuration of subunits in an assembly is useful in biology and provides a starting point for a refinement by higher-resolution methods (Chacon and Wriggers, 2002; Fokine et al., 2004; Gao et al., 2003; Holmes et al., 2003; Topf et al., 2005; Volkmann and Hanein, 1999; Yonekura et al., 2003).

If the resolution of the assembly density map is lower than ~3 nm or the subunit shapes are unknown, the subunit configuration is difficult to determine without additional experiments. In particular, this problem frequently applies to electron tomography, which is especially suitable for studying macromolecular assemblies in their native cellular context (Medalia et al., 2002) but whose resolution is currently limited to less than ~4 nm. To bridge the resolution gap between the assembly shape and the subunit configuration, the assembly density map can be integrated with several additional types of structural information (Alber et al., 2004; Sali et al., 2003). This information includes data from experimental methods, such as chemical crosslinking (Trester-Zedlitz et al., 2003; Malhotra and Harvey, 1994; Young et al., 2000), footprinting (Li et al., 2002), affinity-directed mass spectrometry (Zhao et al., 1996), immunoelectron microscopy (Rout et al., 2000), fluorescence resonance energy transfer (FRET) (Truong and Ikura, 2001), small-angle X-ray and neutron scattering (Koch et al., 2003), site-directed mutagenesis (Wells, 1991), protein arrays (Phizicky et al., 2003), and yeast two-hybrid (Ito et al., 2001; Uetz et al., 2000) as well as theoretical and bioinformatics methods (Aloy et al., 2004; Gray et al., 2003; Russell et al., 2004; Valencia and Pazos, 2002).

In this paper, we focus on characterizing the subunit configuration by combining an assembly density map with one particular source of supplementary information, affinity purification assays. These pull-down experiments depend on a tagged protein subunit (the bait) of a complex. The bait and its noncovalently associated partners (the subcomplex) are first purified by affinity chromatography against the tag and then identified by gel electrophoresis and mass spectroscopy (Aebersold and Mann, 2003; Rout et al., 2000; Cronshaw et al., 2002). Such affinity purification has been used to identify interacting proteins on a large scale in yeast (Gavin et al., 2002; Huh et al., 2003). In contrast to identification of protein interactions, here we exploit the pulldowns for structural characterization. Each affinity purification experiment, in principle, provides some information about spatial relationships among the subunits in the subcomplex. Specifically, all of the proteins identified in a single affinity purification experiment must be located within the expected volume of the subcomplex. Furthermore, each subunit in a subcomplex must interact directly with at least one other subunit in the same subcomplex. For a given assembly, many different subcomplexes can generally be generated by selecting each of the subunits within the assembly as the bait and by varying conditions under which the subcomplexes are purified.

To integrate varied information about the structure of an assembly, we express the structure determination as an optimization approach. In this approach, we need to specify a protein representation, a scoring function, and an optimization method. We use a simplified model with a protein subunit represented by a single sphere. This model can only reveal the configurations of and interactions between subunits, but not their individual conformations nor their relative orientations. Despite these limitations, the proposed representation allows us to encode the affinity purification data and low-resolution assembly density maps as spatial restraints on the subunit configuration, which are then combined into a single scoring function (Figure 1). Next, the scoring function is optimized to find all subunit configurations that satisfy the input restraints. To assess the utility of the combination of the affinity purification data and the assembly density map, the accuracy of the op-

*Correspondence: sali@salilab.org
[3]Lab address: http://salilab.org

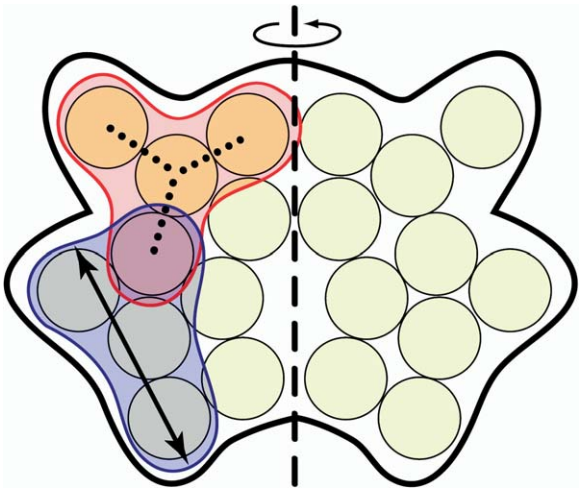

Figure 1. Schematic Representation of the Five Types of Information that Are Assessed with Respect to Their Utility for Assembly Structure Characterization

First, the subunits and their excluded volume are indicated by yellow circles. Second, the assembly shape is indicated by a thick outline. Third, the shapes of two individual subcomplexes, each with four subunits, are shown in red and blue, and the largest diameter of the blue subcomplex is indicated by an arrow (proximity restraint). Fourth, the subunit interactions (connectivity restraint) in the red subcomplex are indicated by dotted lines. Fifth, the symmetry between two parts of the assembly is indicated by a vertical dashed line.

timized configurations was mapped as a function of the variety of simulated restraints for two model assemblies.

Next, we describe in detail an approach to structural characterization by satisfaction of spatial restraints, as well as two model systems and analysis methods used in our calculations (see Approach). We then compare the information content of different combinations of spatial restraints by assessing their predictive power for determining the native assembly structure (see Results). We end by summarizing the main conclusions (see Discussion).

Approach

Structure Characterization by Satisfaction of Spatial Restraints

We express structure characterization as an optimization problem that calculates 3D models consistent with the input information. 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) an optimization of the scoring function to obtain all possible models that satisfy the input restraints. We describe all three components next.

Representation

Each protein subunit is represented as a point. The subunit excluded volume is encoded as a restraint and is described in the next section. The two specific model assemblies used in this paper are described below.

Scoring Function

The most important aspect of structure characterization by satisfaction of spatial restraint is to accurately capture all available input information about the structure of the assembly. We approach this problem by translating all structural information into spatial restraints. We distinguish restraints on five different spatial features (Figure 1): (1) the subunit excluded volume, (2) the assembly shape, (3) the subunit proximity in the subcomplex (the proximity restraint), (4) the subunit connectivity in the subcomplex (the connectivity restraint), and (5) the symmetry. The scoring function is defined as the sum of all individual restraints, described in detail below. In summary, (1) subunit excluded volume restraints are expressed as lower bounds on all pairwise subunit distances; (2) the proximity and (3) connectivity restraints are expressed as pairwise upper distance bounds on the subunits within the subcomplex; (4) the assembly shape restraints are expressed as lower and upper bounds on the absolute subunit coordinates; and (5) the symmetry restraints are expressed as distance restraints on two equivalent parts of the assembly.

In the case of assemblies with multiple copies of the same subunit type (such as the proteasome), there is an ambiguity in the calculation of the proximity and connectivity restraints. For example, there are two copies of each subunit type in the proteasome and four distances between pairs of distinct types. In principle, a restraint on two distinct subunit types could apply to any one of these four pairs. We consider all assignments and only restrain the pair of subunits that leads to the smallest restraint violation.

Subunit Excluded Volume Restraint

The excluded volume restraint imposes a harmonic penalty if the distance between any two subunits is smaller then the sum of their radii (Table 1, row 1).

Assembly Shape Restraint

Subunits can be localized only within a restricted volume in the shape of the target assembly. A harmonic penalty is imposed if the absolute subunit coordinates are below or above the corresponding lower or upper bounds, respectively (Table 1, row 2).

Subcomplex Proximity Restraint

We impose upper distance bounds on all pairs of subunits in a pull-down subcomplex (Table 1, row 3). The upper bound is the largest possible distance between two subunits in a subcomplex and is equal to the maximal diameter of the subcomplex minus the subunit radii. The same subunit pair may appear in multiple subcomplexes and therefore may lead to several upper distance bounds. We keep only the smallest of all pairwise upper bounds.

Subcomplex Connectivity Restraint

Each subunit in a subcomplex must contact at least one other subunit in the subcomplex. For example, in a subcomplex with n components, at least $n - 1$ direct interactions must connect all of its subunits. We refer to this condition as the connectivity restraint of a subcomplex. While the actual subunit contacts are unknown, all valid structural solutions must satisfy this restraint. For a given subcomplex, the restraint is applied with the aid of a minimal spanning tree as follows. We define a fully connected (i.e., complete) graph with

Table 1. Definition of the First Four Restraint Types

Subunit excluded volume restraint	Violated for $f < f_o$, f is the distance between two subunits, f_o is the sum of the subunit radii, and σ is 0.01 nm.
Assembly shape restraint	Lower bound: violated for $f < f_o$, f is the subunit Cartesian coordinate, f_o is the lower bound on this particular subunit coordinate, and σ is 0.1 nm. Upper bound: violated for $f > f_o$, f is the subunit Cartesian coordinate, f_o is the upper bound on this particular subunit coordinate, and σ is 0.1 nm.
Subcomplex proximity restraint	Violated for $f > f_o$, f is the distance between two subunits in a pull-down complex, f_o is the maximal subcomplex dimension, and σ is 0.1 nm.
Subcomplex connectivity restraint	Violated for $f > f_o$, f is the distance between two subunits, f_o is the sum of their radii, and σ is 0.1 nm.

Each restraint term is equal to $(f - f_o)^2/\sigma^2$, where f is the restrained feature, and σ is the parameter that regulates the strength of the term. For upper feature bounds, the score is 0 for $f > f_o$; for lower feature bounds, the score is 0 for $f < f_o$.

the nodes corresponding to the individual subunits and edges with weights equal to the violation of the hypothetical contact (Table 1, row 4). We then find the minimal spanning tree such that the sum of the edge weights is minimal and all subunits are connected to at least one other subunit (Corman et al., 2001). For each edge in the minimal spanning tree, we impose harmonic distance restraints enforcing the direct subunit contacts (Table 1, row 4). At each step of the optimization, we recalculate the fully connected graph and the minimal spanning tree for each subcomplex.

Symmetry Restraints

The similarity between the subunit configurations in each symmetry unit is enforced by imposing a term similar to the distance root mean square (drms), $\Sigma_{ij}\ \omega_{ab}\ (d_{ij}^{a} - d_{ij}^{b})^2$, where d_{ij}^{a} and d_{ij}^{b} are the equivalent distances between two subunits, i and j, in two symmetry-related subunit configurations, a and b, and ω_{ab} is the restraint weight set to 0.2.

Optimization

We generate subunit configurations by simultaneously minimizing violations of all restraints in Cartesian space. The aim is to obtain as many structures as possible that satisfy all input restraints. The generation of these models is stochastic. For each restraint set, we start from at least 10,000 completely randomized subunit configurations. We use an adapted version of the program MODELLER7v0 (Sali and Blundell, 1993).

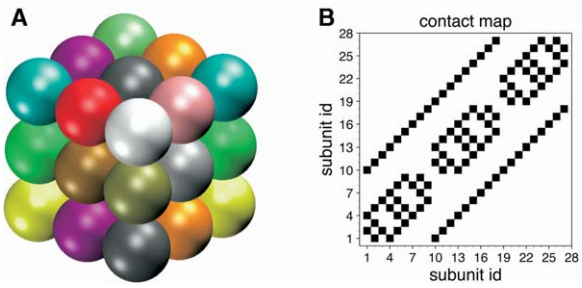

Figure 2. The First Model System

(A) The native cube assembly consisting of 27 different subunits, each of which is represented by a single sphere with a radius of 1 nm. The subunits are located at the grid points of a 3 × 3 × 3 lattice.
(B) The corresponding native contact map with 54 binary subunit contacts. Figure 2A was created with the molecular graphics program VMD (Humphrey et al., 1996).

An optimized structure is obtained from a single optimization run in a series of steps: the initial Cartesian coordinates of all subunits are randomly distributed from –50 to 50 nm, followed by conjugate gradients minimization of up to 500 steps and 50 cycles of simulated annealing molecular dynamics simulation. In each cycle, the temperature of the system is increased from 100 to 1000 K within 50 time steps, kept constant for an additional 100 times steps, and gradually decreased to a temperature of 10 K in 300 time steps. This temperature is kept constant for another 50 time steps, followed by a final optimization by conjugate gradients of up to 1000 steps. After completion of each simulated annealing cycle, the model score is evaluated, and only the structure with the lowest model score is kept.

Model Systems

We use two simple model systems. First, we study a compact assembly consisting of subunits packed in a cube (Figure 2A). Second, we expand our calculations to a more realistic example, a low-resolution model of the proteasome (Figure 3A).

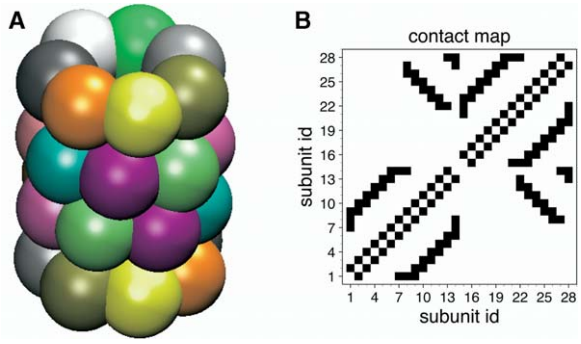

Figure 3. The Second Model System

(A) A low-resolution model of the proteasome with 28 protein subunits. There are 14 different protein types, each occurring twice. Each subunit is represented as a single sphere located at the gravity center of the corresponding protein in the crystal structure of the assembly (Groll et al., 1997). The sphere radii are estimated from the number of residues in each protein (see Approach).
(B) The corresponding native contact map with binary subunit contacts in the low-resolution proteasome structure. Figure 3A was created with the molecular graphics program VMD (Humphrey et al., 1996).

Table 2. Properties of Models Satisfying All Input Restraints that Are Derived from the Six Subcomplex Sets 3–8

Subcomplex Set	3	4	5	6	7	8
Subcomplex Proximity Restraints						
Models satisfying input restraints (%)	100.0	100.0	100.0	100.0	100.0	100.0
Sensitivity[a]	22.2	16.6	18.5	20.3	37.0	37.0
False positive rate[a]	52.2	35.7	41.2	50.0	54.4	71.0
Fraction of correctly predicted models (%)	0.0	0.0	0.0	0.0	0.0	0.0
Drms (nm): smallest, average, largest	1.6, 1.9, 2.2	1.3, 1.6, 1.7	1.7, 1.8, 2.0	1.7, 1.7, 2.2	1.7, 1.9, 2.2	1.6, 1.8, 2.0
Subcomplex Proximity and Assembly Shape Restraints						
Models satisfying input restraints (%)	3.4	1.3	16.0	25.4	36.0	20.7
Sensitivity[a,b] (%)	48.0	61.0	40.7	57.0	62.9	46.3
False positive rate[a,c] (%)	18.8	23.3	46.3	57.5	75.0	77.7
Fraction of correctly predicted models (%)	0.0	1.0	0.0	0.0	0.0	0.0
Drms (nm): smallest, average, largest	0.6, 1.2, 1.5	0.0, 1.1, 1.4	1.1, 1.4, 1.6	1.3, 1.5, 1.7	1.4, 1.6, 1.8	1.4, 1.7, 1.8
Subcomplex Proximity, Subcomplex Connectivity, and Assembly Shape Restraints						
Models satisfying input restraints (%)	0.04	0.1	<0.1	<0.1	<0.1	<0.1
Sensitivity[a,b] (%)	100.0	100.0	100.0	100.0	100.0	100.0
False positive rate[a,c] (%)	0.0	0.0	0.0	0.0	0.0	0.0
Fraction of correctly predicted models (%)	75.0	50.0	100.0	25	33.0	75.0
Drms (nm): smallest, average, largest	0.0, 0.1, 0.5	0.0, 0.2, 0.5	0.0, 0.0, 0.0	0.0, 0.3, 0.4	0.0, 0.2, 0.7	0.0, 0.0, 0.0

See "Cube Model System" in Results. Models are calculated by using subunit excluded volume restraints and subcomplex proximity restraints; subcomplex proximity and the assembly shape restraints; and subcomplex proximity, subcomplex connectivity, and the assembly shape restraints.

[a] Calculated by using the reference frequency cutoff (see Approach).

[b] Sensititvity defined as TP/(TP + FN), where TP is the number of true positive contacts and FN is the number of false negative contacts.

[c] False-positive rate defined as FP/(FP + TP), where FP is the number of false positive contacts and TP is the number of true postive contacts.

Cube Model System

The cube assembly consists of 27 different subunits located at the grid points of a 6 nm × 6 nm × 6 nm lattice (Figure 2A). All subunits are represented as hard spheres with radii of 1 nm. The assembly contains 54 distinct binary contacts shown as a contact map in Figure 2B. For the assembly shape restraint, the shape is a cube with side lengths of 6 nm. For the subcomplex proximity restraint, the maximal distances between subunit centers in subcomplexes with 3–8 subunits are 4, 4.47, 5.66, 6.00, 6.93, and 6.93 nm, respectively.

Proteasome Model System

The proteasome consists of 28 globular proteins of 14 different types that are arranged in two identical pairs of rings (Figure 3A). We approximate each protein by a single sphere with its radius (in nm) estimated from the total protein mass: $r = 0.0726\ M^{1/3}$, where M is the protein mass in Da and the coefficient is determined based on masses and sizes of known protein structures. The sphere center is located at the center of mass of the corresponding protein in the X-ray structure of the proteasome (Groll et al., 1997). For the assembly shape restraint, the shape is a cylinder with a height of 16.2 nm and a radius of 3.3 nm. For the subcomplex proximity restraint, the upper bound is 1.35 times the estimated maximal subcomplex diameter (in nm) from the empirical relationship between the maximal diameter of a subcomplex and its total number of residues: $D = 0.495\ n^{1/3}$, where n is the total number of residues in the subcomplex. The parameter value of 0.495 was derived by fitting the function to the structurally defined protein assemblies in PIBASE (Davis and Sali, 2004), such that 95% of all complexes have predicted maximal diameters that are larger or equal to the actual diameters.

Simulation of Pull-down Subcomplexes

Subcomplexes are generated by an iterative random selection of subunits that are in direct contact with each other in the native structure. A starting point is a subunit that is selected as the bait of the subcomplex. The acceptance of a newly selected subunit is probabilistic; the probability for accepting a subunit is proportional to the inverse cube of the contact shell number, which is the smallest number of subunits that connect the selected subunit with the bait. A uniform selection probability would lead to artificially elongated subcomplexes, as the number of neighbors in higher contact shells grows rapidly.

Generation of Additional Models

For some restraint sets (e.g., derived from subcomplex sets 7 and 8 in Table 2), the optimization protocol was unable to generate a sufficient number of structures that satisfied all the input restraints, even in 500,000 independent runs. In such cases, we increased the

sample size needed for estimating the utility of various restraint sets for structure characterization as follows. We generated 3,000 additional structures from the native structure by swapping subunits between 1 and 10 randomly selected subunit pairs in the assembly. For the proteasome model, each swap involved two pairs of subunits, one in each symmetry unit. If a structure satisfied all input restraints, it was added to the ensemble of good scoring structures generated in the optimization process.

Analysis

Analysis is performed only on models that completely satisfy all input restraints (good scoring models).

Contact Frequencies

A subunit contact is defined if the distance between the two subunits is smaller then the sum of their radii multiplied by a tolerance factor of 1.05. The contact frequency is defined as the ratio of the number of models with the contact and the number of all models.

Receiver Operating Characteristic Analysis

The ability of different restraint sets to predict the native subunit interactions is ranked with the aid of the ROC curves (Theodoridis and Koutroumbas, 1999). For an ensemble of models calculated by a given restraint set, a subunit interaction is predicted if the corresponding contact frequency is sufficiently high (below). The accuracy of the predicted subunit interactions is quantified by calculating the true positive rate (sensitivity) as well as the false positive rate (1-specificity) and plotting them against each other at 16 different cutoff values (the ROC curve). The area under the ROC curve represents the probability of correct classification over the whole range of cutoffs; it can range from 0.5 to 1. An area of 0.5 indicates that the structure calculation could not discriminate between the native and false contacts. If the area under the ROC curve equals 1, the method is able to predict the contact map of the native structure. The closer the ROC curve is to the upper left corner and the closer the integrated area under the curve is to 1, the higher the overall accuracy of the calculations is and the more informative the restraints are about the native contact map of the assembly.

Reference Frequency Cutoff

This cutoff is defined as 56% of the largest contact frequency value present in a contact frequency map. This value was obtained by maximizing the sum of true positives and true negatives for the restraint set derived from subcomplex set 4 (Table 2) and was adopted as a reference value for the analysis of all the restraint sets. Varying the reference cutoff value in a wide range from 30% to 90% does not change the ranking of the restraint sets by their utility in structure characterization. For convenience, the false positive rates and the number of correctly predicted contacts for each restraint set are determined by using the reference frequency cutoff value.

Results

We rely on two simple model systems in which globular protein subunits are represented as single spheres (see Approach). Our aim is to enumerate all subunit interaction networks and configurations that are consistent with subunit excluded volume; protein affinity purification experiments; mass density maps determined by electron cryomicroscopy or tomography, and, when applicable, symmetry (see Approach) (Figure 1). We achieve this aim by using simulated input data sets generated from two simple model systems.

We focus on the utility of affinity chromatography purification for structure characterization. Each experiment reveals the types of proteins present in the pull-down subcomplex and, in principle, contains some information about spatial relationships between subunits in the pull-down subcomplex (see Approach). One such spatial restraint is the upper distance bound on any two subunits in a subcomplex, which we refer to as the "proximity restraint." The dimension of a subcomplex may be derived from hydrodynamic experiments (de la Torre and Bloomfield, 1977), small-angle X-ray scattering (Koch et al., 2003), and negative-stain or electron cryomicroscopy images (Frank, 2002). Another spatial restraint, the "connectivity restraint," specifies that every subunit in a subcomplex must interact with at least one other subunit in the subcomplex. While the actual subunit interaction network is unknown, all valid structural solutions must satisfy this connectivity restraint.

Cube Model System

Our first model system is an assembly of 27 different subunits, represented as single hard spheres of identical radii in a cubic close-packed lattice (Figure 2A). We generated 6 data sets, each composed of 27 simulated pull-down experiments with each of the subunits selected as the bait. Subcomplexes in a data set contain the same number of subunits. We employ data sets with three, four, five, six, seven, and eight subunits per subcomplex (columns 3–8, Table 2). These data sets will allow us to investigate which subcomplex size is most informative about the structure of the assembly. For each data set, we consider three combinations of restraint types: first, we use a combination of the excluded volume restraints for each subunit and the proximity restraints for each of the 27 subcomplexes per data set as the only information for structure characterization (Table 2); second, we add the assembly shape restraint (Table 2); and third, we also add subcomplex connectivity restraints (Table 2). This sequential buildup of the scoring function allows us to isolate the individual contributions to the structural characterization of assemblies.

Subcomplex Proximity Restraints

We begin by considering only subunit excluded volume restraints and subcomplex proximity restraints calculated from the six data sets (columns 3–8, Table 2) containing subcomplexes with 3–8 subunits (see Approach). For each of the six generated restraint sets, at least 10,000 random subunit configurations were optimized in an attempt to find those configurations that satisfy all input restraints (good scoring models). We then predict a subunit interaction if it occurs frequently in the ensemble of good scoring models. Finally, we rely on the Receiver Operating Characteristic (ROC)

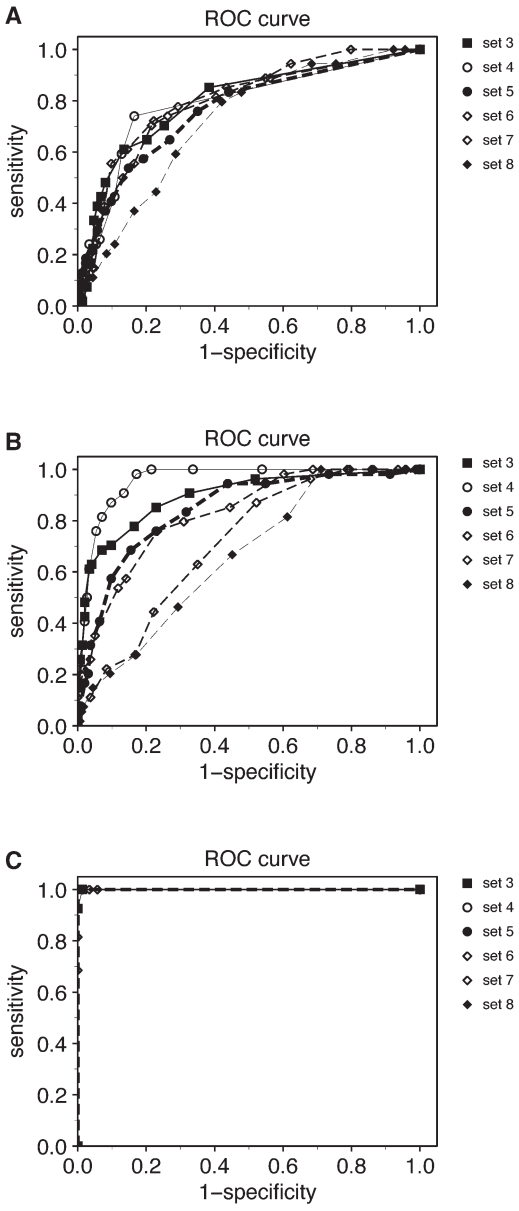

Figure 4. Assessment of Contact Map Prediction

(A–C) ROC curves (see Approach) for six different subcomplex sets (sets 3–8 defined in "Cube Model System" in Results). Models are generated by using subunit excluded volume restraints and (A) subcomplex proximity restraints, (B) subcomplex proximity and assembly shape restraints, and (C) subcomplex proximity, subcomplex connectivity, and assembly shape restraints. The area under the curves is 1 or close to 1.

curves to rank the different restraint sets by their ability to correctly predict the native contacts.

The ROC curves for subcomplex sets 3–7 are similar to each other (Figure 4A). The overall performance is poor, as indicated by the small integrated area under the ROC curves that ranges from 0.7 to 0.8 for all subcomplex sets (Figure 5). Even for the two best-performing subcomplex sets, 3 and 4, only, respectively, 12 and 14 out of the total of 54 native interactions are predicted correctly (the corresponding false positive rates are ~50% and 36%, respectively). This poor performance is also revealed by the 3D structural analysis of the models. The average drms deviation between models and the native structure ranges from 1.6 to 1.9 nm (Table 2). Therefore, it is not possible to correctly determine the assembly structure only by the subunit excluded volume and subcomplex proximity restraints.

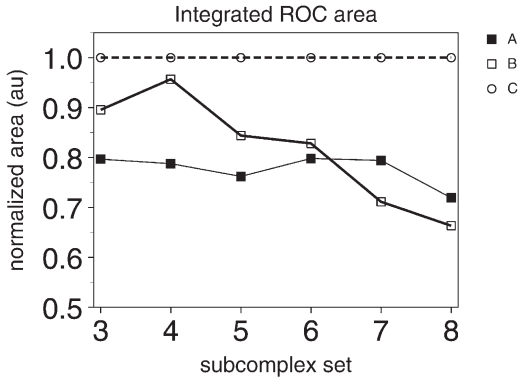

Figure 5. Integrated Area under ROC Curves for Calculations with Restraints Derived from Six Different Subcomplex Sets 3–8

See "Cube Model System" in Results.

(A–C) Models are calculated by using subunit excluded volume restraints and (A) subcomplex proximity restraints, (B) subcomplex proximity and assembly shape restraints, and (C) subcomplex proximity, subcomplex connectivity, and assembly shape restraints.

Subcomplex Proximity and Assembly Shape Restraints

Next, we investigate the effect of adding the assembly shape restraint on the accuracy of our predictions. We use the same subcomplex data sets 3–8, but we now restrict the positions of the subunits to be within the assembly shape (a cube with side length of 6 nm) (see Approach).

With the addition of the assembly shape restraint, the models are generally more compact. For some of the restraint sets, a substantial fraction of the native contacts can now be predicted correctly. For example, 26 of the 54 native contacts occur in 60% of all models calculated from the restraint set 3. The false positive rate is 18.8% (Table 3; Figure 6B). The number of subunits per subcomplex makes a significant difference in the utility of the corresponding restraints, as indicated by the spread of the ROC curves in Figure 4B.

Subcomplex sets with a large number of subunits (e.g., sets 7 and 8) perform worse with the assembly shape restraint than without it (cf. Figures 5A and 5B). For example, for subcomplex set 7, the integrated ROC area for subcomplex sets 7 and 8 decreases from 0.78 to 0.71 and the false positive rate for subunit interaction prediction rises from 57% to 75% (subcomplex set 7 in Table 2). This finding is not surprising, as the estimated diameter of subcomplexes with 7 and 8 subunits is similar to the maximum diameter of the assembly. Therefore, subcomplex sets 7 and 8 do not provide any additional structural information if the assembly shape is

Table 3. Properties of Models Satisfying All Input Restraints that Are Derived from Subcomplex Sets Containing 14 and 28 Subcomplexes

Subcomplex Set	14	28
Sensitivity	96.3	100.0
False positive rate	0.0	0.0
Drms (nm): smallest, average), largest	0.0, 0.8, 1.7	0.0, 0.0, 0.0

The proteasome model system is explained in Results. Subunit-excluded volume, assembly shape, subcomplex proximity, subcomplex connectivity, as well as symmetry restraints are applied (see Approach). See the legend of Table 2 for the definitions of sensitivity and false positive rate.

already specified. However, the increased number of contacts (both native and nonnative) resulting from the reduced accessible volume increases the false positive rate and therefore decreases the prediction accuracy as quantified by a measure that depends on the subunit contacts. While it may be surprising that the accuracy of contact prediction from subcomplex sets 7 and 8 is decreased upon the addition of the assembly shape restraint, other aspects of the predicted structures are improved — for example, the accuracy of the shape prediction (data not shown).

In contrast, for subcomplex sets with a smaller number of subunits (e.g., subcomplex sets 3 and 4), the prediction accuracy is strongly improved upon adding the assembly shape restraint. The highest accuracy is found for subcomplex set 4 (Figure 5), with 33 out of the 54 native contacts correctly determined, in comparison to the prediction of 12 native contacts without the assembly shape restraints. Also, the false positive rate drops from 36% to 23%, and the integrated ROC area increases from 0.8 to 0.96 (subcomplex set 4 in Table 2). Correspondingly, the structural similarity among the models that satisfy the input restraints increases, and their average drms deviation to the native structure is ~1.1 nm (Table 2). Approximately 1% of all models in subcomplex set 4 have all native contacts predicted correctly.

Subcomplex Proximity, Assembly Shape, and Subcomplex Connectivity Restraints

Finally, we investigate the effect of adding the connectivity restraint on the accuracy of our predictions. Using the same subcomplex sets, we now enforce that each subunit in a subcomplex is connected to the rest of the subcomplex subunits via at least one direct contact (subcomplex connectivity restraints in Approach). For the subcomplex sets with a small number of subunits (three and four components), the current optimization scheme provides a sufficient number of models for subsequent analysis. However, for larger subcomplex

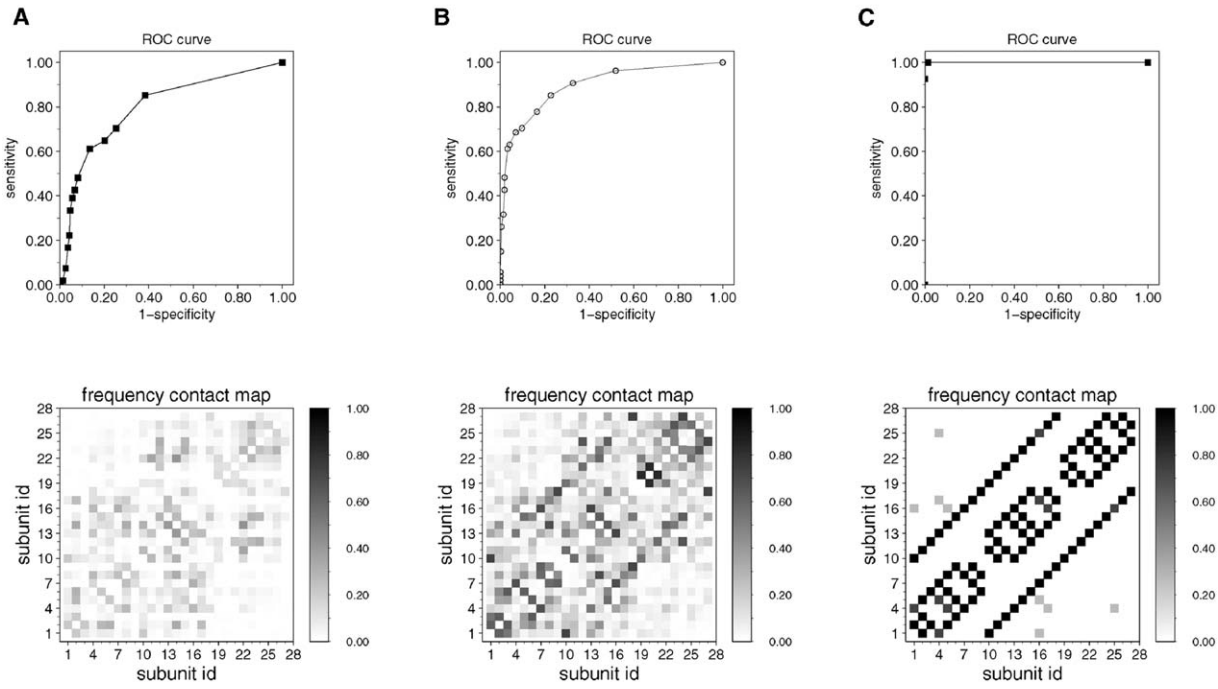

Figure 6. ROC Curves and Contact Frequency Maps for the Cube Model System with Subcomplex Set 3

See "Cube Model System" in Results.

(A–C) Models are calculated by using subunit excluded volume restraints and (A) subcomplex proximity restraints, (B) subcomplex proximity and assembly shape restraints, and (C) subcomplex proximity, subcomplex connectivity, and assembly shape restraints.

sets (between 5 and 8 subunits), we supplement the structures provided by the optimization scheme with additional structures (see Approach) to improve the reliability of the results.

Adding subunit connectivity restraints leads to a dramatic improvement in the accuracy of structure determination. The contact frequency maps for all subcomplex sets are almost identical to the contact map of the native structure (e.g., Figures 2B and 6C). Indeed, for subcomplex set 3, all native contacts are reproduced in the good scoring models with a frequency of at least 75% (50 contacts with a frequency of 100% and 4 contacts with a frequency of 75%). Hence, the integrated ROC area is ~1 for all subcomplex sets (Figure 5). Using the reference frequency cutoff value (see Approach), we are able to determine the complete subunit interaction network of the native structure with a false positive rate of 0 (Figure 5 and Table 2). Structural comparison between the native structure and all models that satisfied the input restraints revealed an average drms deviation ranging from 0.1 to 0.3 nm (Table 2). Indeed, for all of the subcomplex sets, some of the predicted structures differed only by a single interchange of neighboring subunits. Moreover, for the reference frequency cutoff, only models identical to the native structure have the contacts represented in the contact map. Therefore, the native structure can be identified reliably as the most frequently occurring predicted model.

Proteasome Model System

Having demonstrated that it is possible to determine the 3D configuration of a simple model assembly, we turn our attention to the more realistic case of the proteasome.

Given the shape of the proteasome, a soft sphere representation of each of the proteins (one sphere per protein), and a new symmetry restraint (see Approach), we assessed the information content of a relatively modest set of subcomplexes (with 14 and 28 simulated subcomplexes per subcomplex set) (Table 3). Each of these subcomplexes contained between 3 and 5 subunits, with an average of 4 subunits in each subcomplex set. Instead of calculating models by the optimization of the scoring function, we constructed 3000 structures that differed from the native proteasome by a drms of 0.0–4.1 nm (Figure 7) (see Approach). These structures were evaluated by the scoring function. All models with scores less than five times the score of the native structure were included in the analysis (Figure 7).

With 14 subcomplexes, we were able to predict 55 out of the 57 native contacts with an error rate of 0, by using the reference frequency cutoff (Table 3, Figure 8A) (see Approach). As expected, the subcomplex set with 28 subcomplexes performed even better, predicting the complete subunit interaction network (Table 3, Figure 8B). For both cases, the integrated ROC area is ~1, indicating the highly discriminative power of the scoring functions (Figure 8). The scoring function derived from 14 subcomplexes allowed several models that differed only by a single interchange of neighboring spheres. These models differed on average by a drms of 0.8 nm from the native structure. Again, only the native structure contained all predicted direct interactions, which would allow us to determine the native structure without knowing the correct answer in advance.

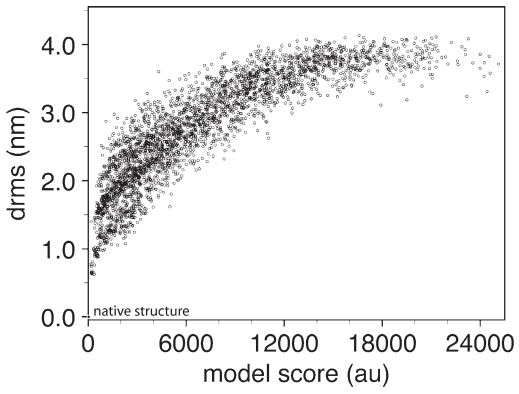

Figure 7. The Structural Similarity between the Proteasome Models and the Native Structure Plotted against the Corresponding Model Score Derived from an Input Data Set Containing 28 Subcomplexes

All models with scores (au, arbitrary units) less than five times the score of the native structure were included in the analysis. The native structure is indicated in the lower left corner.

Discussion

We showed that it is generally possible to determine the subunit packing in assemblies at low resolution by using as sources of spatial information an appropriate representation of the individual subunits, the assembly shape, and only a modest number of subcomplexes (Table 2, Figure 4). This goal is achieved by the satisfaction of spatial restraints that depends on a subunit representation, a scoring function, and an optimization (see Approach).

Information about the coarse shape of the individual subunits can be provided by several methods, including hydrodynamic experiments (de la Torre and Bloomfield, 1977), small-angle X-ray scattering (Koch et al., 2003), negative-stain or electron cryomicroscopy images (Frank, 2002), and bioinformatics. If such analyses are unavailable, the upper bound on the size can be estimated from the mass of a subcomplex. The shape of the assembly can be characterized by a variety of imaging techniques, such as electron cryomicroscopy and tomography. However, these imaging methods sometimes lack the resolution to provide the subunit configuration. We suggest that complementing these imaging techniques with protein affinity purification experiments may provide a way to bridge the resolution gap between assembly shape and subunit configuration.

In our calculations, we used restraints on five different spatial features, including subunit excluded volume, assembly shape, subunit proximity in a subcomplex (proximity restraint), subunit connectivity in a subcomplex (connectivity restraint), and symmetry. None of these restraint types are sufficient on their own for the accurate determination of the native assembly

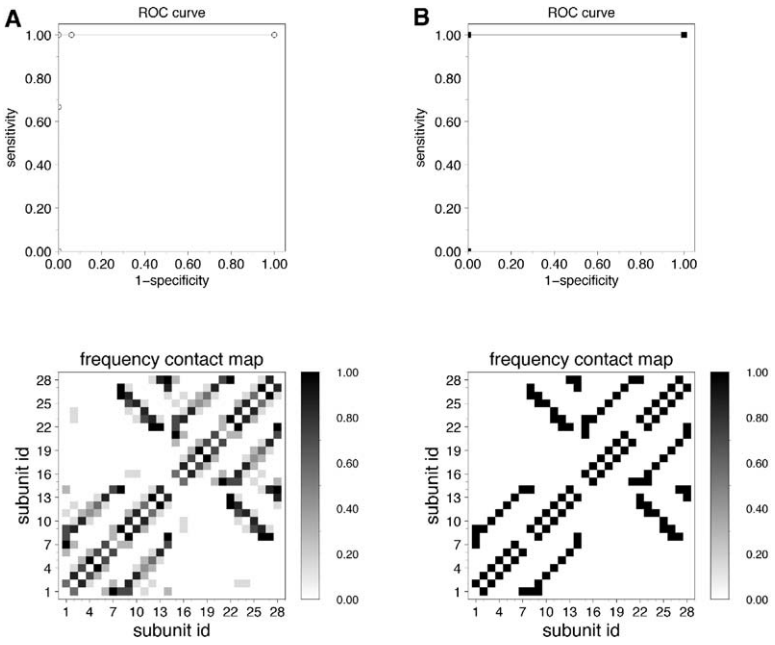

Figure 8. ROC Curves and Contact Frequency Maps for the Proteasome Model

(A and B) Models are calculated by using excluded volume restraints, subcomplex proximity and connectivity restraints, symmetry restraints, and assembly shape restraints (see Approach) with (A) 14 subcomplexes in a subcomplex set and (B) 28 subcomplexes in a subcomplex set (see "Proteasome Model System" in Results).

structure. However, when all of them are integrated into a single scoring function, the correct subunit configuration can be determined. The subcomplex connectivity restraint is particularly useful for accurate structure determination (Table 2 and Figure 2). While the subcomplex proximity restraint is helpful, it is not as informative as the connectivity restraint (Table 2 and Figure 2).

The pull-down restraints generally cannot distinguish between the native structure and its mirror image. Therefore, when the shape of the system is identical to its mirror image, as is the case for the two model systems used here, we cannot distinguish the native structure from its mirror image.

Our analysis depends on sufficiently thorough sampling of the subunit configurations that are consistent with all input restraints. However, once a sufficient sampling is achieved, the analysis is independent of the optimization method. In other words, the assessment of the information content of the input restraints is entirely independent of the sampling procedure used to find good scoring models. The current optimization protocol provides from hundreds to thousands of configurations that satisfy all the restraints derived from most of the subcomplex sets, which we suggest is sufficient for a coarse ranking of the information content of the different restraint sets (Table 2). The exceptions are the restraint sets that include subcomplex connectivity restraints derived only from large subcomplexes (subcomplex sets 5–8 in Table 2), which result in a combinatorial explosion in the number of possible minimal spanning trees per subcomplex. In principle, this expansion of the search space requires more sampling to find good scoring solutions. However, we circumvented this problem by constructing additional good scoring structures based on the native structure (see Approach). It is possible that, in some applications of our approach, a more efficient optimization protocol will be needed to find good scoring structures. For example, if data sets contain subcomplexes of variable sizes, the efficiency of the sampling may be improved by constructing and optimizing the scoring function in several steps, employing the variable target function approach (Braun and Go, 1985) (data not shown). At first, only connectivity restraints derived from the smallest subcomplexes are considered; then, connectivity restraints of larger subcomplexes are added to the scoring function. This procedure leads to a smaller search space for connectivity restraints of large subcomplexes and allows sufficient sampling of good scoring models. This improved optimization strategy should be applicable to most experimental data sets that often contain subcomplexes with a variable number of components.

We assessed the feasibility of determining the configuration of subunits in an assembly by integrating low-resolution spatial information from mass density maps and subunit interactions obtained by pull-down experiments. The key question is whether or not there is sufficient information in such low-resolution data to allow determination of the native structure. The analysis requires the native assembly structure, a low-resolution density map, and many sets of pull-down restraints. As the necessary data are not available for any protein assembly, we performed our analysis on two model systems by using simulated input data sets. This approach allowed us to explore in detail the information content of a large variety of restraint sets, particularly for pull-down experiments.

In the future, testing of our approach could be expanded in a variety of ways. First, we have not exhaustively explored all combinations of different restraint types. For example, we could assess the information content of various combinations of pull-down sizes. Second, we have not yet mapped the accuracy of the structure determination as a function of the error in the simulated restraint sets. This objective can be achieved by using the same approach as described here, except

that some error is introduced in the simulated restraints. Third, we did not study ways to minimize the impact of errors in the input restraints. When the fraction of incorrect restraints is small, we expect that it will be possible to identify incorrect restraints by the inability to find models that are consistent with all of the restraints. We could also employ jack-knifing to identify incorrect restraints. Fourth, we will apply our approach to real assemblies with real data. Large-scale tandem affinity purification experiments may provide a way to do so.

This study is part of our effort to develop and apply a computational system for enumerating structures of protein assemblies that are consistent with all available information from experimental methods, physical theories, and statistical preferences extracted from biological databases (Alber et al., 2004; Sali et al., 2003). We are currently introducing structural representations at multiple levels of resolution. This extension will allow us to use pull-down information together with other sources of spatial information, such as density fitting, computational docking, and crosslinking. The resulting integrated system will maximize efficiency, accuracy, resolution, and completeness of the structural coverage of protein assemblies.

Acknowledgments

We are grateful to Maya Topf, Dmitry Korkin, Fred Davis, Damien Devos, Min-yi Shen, and Ben Webb for many discussions about structure characterization by satisfaction of spatial restraints. We also acknowledge our collaborators, Brian Chait and Mike Rout, for providing biological context that inspired this study. M.F.K. is supported by the Burroughs-Wellcome fund. We acknowledge funding by the National Institutes of Health/National Cancer Institute (R33 CA89810), National Science Foundation EIA-0325004, as well as computer hardware gifts from Sun, Intel, and IBM.

Received: December 5, 2004
Revised: January 12, 2005
Accepted: January 14, 2005
Published: March 8, 2005

References

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

Alber, F., Eswar, N., and Sali, A. (2004). Structure determination of macromolecular complexes by experiment and computation. In Nucleic Acids and Molecular Biology, Volume 15, Practical Bioinformatics, J.M. Bujnicki, ed. (Berlin, Heidelberg: Springer-Verlag), pp. 73–96.

Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A.C., Bork, P., Superti-Furga, G., Serrano, L., and Russell, R.B. (2004). Structure-based assembly of protein complexes in yeast. Science *303*, 2026–2029.

Ban, N., Nissen, P., Hansen, J., Moore, P.B., and Steitz, T.A. (2000). The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. Science *289*, 905–920.

Baumeister, W., Grimm, R., and Walz, J. (1999). Electron tomography of molecules and cells. Trends Cell Biol. *9*, 81–85.

Beck, M., Forster, F., Ecke, M., Plitzko, J.M., Melchior, F., Gerisch, G., Baumeister, W., and Medalia, O. (2004). Nuclear pore complex structure and dynamics revealed by cryoelectron tomography. Science *306*, 1387–1390.

Braun, W., and Go, N. (1985). Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm. J. Mol. Biol. *186*, 611–626.

Carter, A.P., Clemons, W.M., Brodersen, D.E., Morgan-Warren, R.J., Wimberly, B.T., and Ramakrishnan, V. (2000). Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. Nature *407*, 340–348.

Chacon, P., and Wriggers, W. (2002). Multi-resolution contour-based fitting of macromolecular structures. J. Mol. Biol. *317*, 375–384.

Corman, T.H., Leisersoon, C.E., Rivest, R.L., and Stein, C. (2001). Introduction to Algorithms, Second Edition (Cambridge, Massachusettts: MIT Press).

Cronshaw, J.M., Krutchinsky, A.N., Zhang, W., Chait, B.T., and Matunis, M.J. (2002). Proteomic analysis of the mammalian nuclear pore complex. J. Cell Biol. *158*, 915–927.

Davis, F.P. and Sali, A. (2004). PIBASE: a comprehensive database of structurally defined protein domain interfaces. Bioinformatics, in press.

de la Torre, J.G., and Bloomfield, V.A. (1977). Hydrodynamic theory of swimming of flagellated microorganisms. Biophys. J. *20*, 49–67.

Fokine, A., Chipman, P.R., Leiman, P.G., Mesyanzhinov, V.V., Rao, V.B., and Rossmann, M.G. (2004). Molecular architecture of the prolate head of bacteriophage T4. Proc. Natl. Acad. Sci. USA *101*, 6003–6008.

Frank, J. (2002). Single-particle imaging of macromolecules by cryo-electron microscopy. Annu. Rev. Biophys. Biomol. Struct. *31*, 303–319.

Gao, H., Sengupta, J., Valle, M., Korostelev, A., Eswar, N., Stagg, S.M., Roey, P.V., Agrawal, R.K., Harvey, S.C., Sali, A., et al. (2003). Study of the structural dynamics of the *E. coli* 70S ribosome using real-space refinement. Cell *113*, 789–801.

Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141–147.

Gray, J.J., Moughon, S.E., Kortemme, T., Schueler-Furman, O., Misura, K.M., Morozov, A.V., and Baker, D. (2003). Protein-protein docking predictions for the CAPRI experiment. Proteins *52*, 118–122.

Groll, M., Ditzel, L., Lowe, J., Stock, D., Bochtler, M., Bartunik, H.D., and Huber, R. (1997). Structure of 20S proteasome from yeast at 2.4 Å resolution. Nature *386*, 463–471.

Harms, J., Schluenzen, F., Zarivach, R., Bashan, A., Gat, S., Agmon, I., Bartels, H., Franceschi, F., and Yonath, A. (2001). High resolution structure of the large ribosomal subunit from a mesophilic eubacterium. Cell *107*, 679–688.

Holmes, K.C., Angert, I., Kull, F.J., Jahn, W., and Schroder, R.R. (2003). Electron cryo-microscopy shows how strong binding of myosin to actin releases nucleotide. Nature *425*, 423–427.

Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K. (2003). Global analysis of protein localization in budding yeast. Nature *425*, 686–691.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD - Visual Molecular Dynamics. J. Mol. Graph. *14*, 33–38.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA *98*, 4569–4574.

Koch, M.H., Vachette, P., and Svergun, D.I. (2003). Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. Q. Rev. Biophys. *36*, 147–227.

Li, F., Gangal, M., Juliano, C., Gorfain, E., Taylor, S.S., and Johnson, D.A. (2002). Evidence for an internal entropy contribution to phosphoryl transfer: a study of domain closure, backbone flexibility, and the catalytic cycle of cAMP-dependent protein kinase. J. Mol. Biol. *315*, 459–469.

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*, 308–340.

Medalia, O., Weber, I., Frangakis, A.S., Nicastro, D., Gerisch, G., and Baumeister, W. (2002). Macromolecular architecture in eukaryotic cells visualized by cryoelectron tomography. Science *298*, 1209–1213.

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

Rout, M.P., Aitchison, J.D., Suprapto, A., Hjertaas, K., Zhao, Y., and Chait, B.T. (2000). The yeast nuclear pore complex: composition, architecture, and transport mechanism. J. Cell Biol. *148*, 635–651.

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*, 313–324.

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. *234*, 779–815.

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

Schmid, M.F., Sherman, M.B., Matsudaira, P., and Chiu, W. (2004). Structure of the acrosomal bundle. Nature *431*, 104–107.

Theodoridis, S., and Koutroumbas, K. (1999). Pattern Recognition (London: Academic Press).

Topf, M., Baker, M.L., John, B., Chiu, W., and Sali, A. (2005). Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy. J. Struct. Biol. *149*, 191–203.

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*, 2416–2425.

Truong, K., and Ikura, M. (2001). The use of FRET imaging microscopy to detect protein-protein interactions and protein conformational changes in vivo. Curr. Opin. Struct. Biol. *11*, 573–578.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., et al. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. Nature *403*, 623–627.

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

Volkmann, N., and Hanein, D. (1999). Quantitative fitting of atomic models into observed densities derived by electron microscopy. J. Struct. Biol. *125*, 176–184.

Wells, J.A. (1991). Systematic mutational analyses of protein-protein interfaces. Methods Enzymol. *202*, 390–411.

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*, 223–234.

Yonekura, K., Maki-Yonekura, S., and Namba, K. (2003). Complete atomic model of the bacterial flagellar filament by electron cryomicroscopy. Nature *424*, 643–650.

Young, M.M., Tang, N., Hempel, J.C., Oshiro, C.M., Taylor, E.W., Kuntz, I.D., Gibson, B.W., and Dollinger, G. (2000). High throughput protein fold identification by using experimental constraints derived from intramolecular cross-links and mass spectrometry. Proc. Natl. Acad. Sci. USA *97*, 5802–5806.

Zhao, Y., Muir, T.W., Kent, S.B., Tischer, E., Scardina, J.M., and Chait, B.T. (1996). Mapping protein-protein interactions by affinity-directed mass spectrometry. Proc. Natl. Acad. Sci. USA *93*, 4020–4024.

Zhang, G., Campbell, E.A., Minakhin, L., Richter, C., Severinov, K., and Darst, S.A. (1999). Crystal structure of *Thermus aquaticus* core RNA polymerase at 3.3 Å resolution. Cell *98*, 811–824.

Zhang, W., Chipman, P.R., Corver, J., Johnson, P.R., Zhang, Y., Mukhopadhyay, S., Baker, T.S., Strauss, J.H., Rossmann, M.G., and Kuhn, R.J. (2003). Visualization of membrane protein domains by cryo-electron microscopy of dengue virus. Nat. Struct. Biol. *10*, 907–912.