Integrative Structure Modeling of Macromolecular Assemblies from Proteomics Data*

Keren Lasker‡§¶‖, Jeremy L. Phillips‡¶**‡‡, Daniel Russel‡, Javier Velázquez-Muriel‡, Dina Schneidman-Duhovny‡§§, Elina Tjioe‡, Ben Webb‡, Avner Schlessinger‡, and Andrej Sali‡¶‖

Proteomics techniques have been used to generate comprehensive lists of protein interactions in a number of species. However, relatively little is known about how these interactions result in functional multiprotein complexes. This gap can be bridged by combining data from proteomics experiments with data from established structure determination techniques. Correspondingly, integrative computational methods are being developed to provide descriptions of protein complexes at varying levels of accuracy and resolution, ranging from complex compositions to detailed atomic structures. Molecular & Cellular Proteomics 9:1689–1702, 2010.

A 3-D enhanced version of this article is available. The text is identical to this version but includes interactive figures.

Viewing the enhanced version of this article requires the use of a browser plug-in. Please install the plug-in when prompted. http://www.thesgc.org/iSee/MCP/9/8/e2.html

MOTIVATION: STRUCTURES FOR MECHANISTIC UNDERSTANDING OF PROCESSES

The cell contains hundreds of functional macromolecular assemblies responsible for performing critical cellular processes (1, 2). These include, among others, the ribosome (translation) (3, 4), chaperonins (protein folding) (5, 6), RNA polymerase (RNA synthesis) (7), and the proteasome (protein degradation) (8–10). A macromolecular machine is often built around a stable core of proteins that defines the basic function of the complex. This core assembly can be modulated through interactions with peripheral protein components, resulting in a multitude of functionally relevant states (11). A structural description of an assembly in all of its states often facilitates a mechanistic understanding of the corresponding process (3, 12, 13). Thus, a critical challenge in structural biology is to identify biologically relevant states of macromolecular assemblies and to determine the structures of these states at the highest possible resolution.

ASSEMBLY STRUCTURES OFTEN CANNOT BE RESOLVED BY A SINGLE TECHNIQUE

The structures of macromolecular assemblies in their biologically significant states generally cannot be resolved to atomic resolution by a single technique (14). Although x-ray crystallography remains the most powerful approach for visualizing a static snapshot of a complex at atomic resolution, it is limited to samples that can be purified in large quantities and crystallized (15). Similarly, NMR spectroscopy results in an ensemble of structures of a system in solution (16–18), but the technique is limited by the size of the complex and sample availability. Electron microscopy (EM)1 techniques provide an alternative approach for visualizing multiple conformations of complexes in vitro and even within cells (19–22). However, in most cases, the resolution of an electron density map is too low to provide a full mechanistic description of a protein complex. Additional techniques, such as high throughput proteomics methods (23), small angle x-ray scattering (SAXS) (24, 25), and fluorescence resonance energy transfer (FRET) spectroscopy (26), are generally limited by low resolution (14) and at times by low accuracy (27–29) of the corresponding structural information.

INTEGRATIVE STRUCTURE DETERMINATION

The limitations in the resolution, accuracy, and coverage of individual experimental methods can be bridged by simultaneous consideration of multiple types of information. Examples of techniques that specialize in integrating a few types of experimental data include (i) combining electron density maps of complexes with atomic structures of protein components to build high resolution structures of protein complexes (30–34); (ii) using atomic models to estimate the phases required for converting diffraction data into electron density maps (35); (iii) inferring the binary interaction map of a complex from

1 The abbreviations used are: EM, electron microscopy; FRET, fluorescence resonance energy transfer; SAXS, small angle x-ray scattering; NPC, nuclear pore complex; AAA-ATPase, adenosine triphosphatase associated with diverse cellular activities; RNAPII, RNA polymerase II; Rpβ, RNA polymerase II subunit; H-RNAPII, human RNA polymerase II; Y2H, yeast two-hybrid; RMSD, root mean square deviation.
UNIFIED APPROACH FOR INTEGRATIVE MODELING

As outlined above, different studies on different systems will have a variety of different types of available data (Fig. 1 and Table I). Therefore, a unified approach for integrative modeling that can incorporate any type of information about a macromolecular assembly into the determination of its structure is needed. This information may include physical theories, statistical preferences extracted from biological databases, and heterogeneous experimental data at different resolutions, ranging from atomic structures to sets of inter-

affinity purification, mass spectrometry, and comparative modeling data (36); and (iv) incorporating NMR-derived data into protein structure prediction (37, 38).

Recently, a number of macromolecular structures have been resolved by such integrative methods. For instance, the constituent proteins in the nuclear pore complex (NPC) were localized based on the shape and symmetry of the NPC from cryo-EM, positions of the proteins from immuno-EM, relative proximities of proteins from affinity purification, and the shapes of proteins from ultracentrifugation (13, 39). An atomic model of the AAA-ATPase ring of the 26 S proteasome was determined primarily by fitting comparative models of subunits into a single-particle cryo-EM map subject to protein interactions identified by proteomics (40). A structural model for a complete clathrin lattice (41) and a mechanistic model of the clathrin lattice assembly-disassembly cycle driven by chaperone Hsc70 (42) were suggested by combining data obtained by x-ray crystallography and single-particle cryo-EM. The architecture of RNA polymerase II in complex with its initiation factors was determined by combining known crystal structures with data from chemical cross-linking coupled to mass spectrometry (43). An NMR solution structure for the interface between two subunits in the human immunodeficiency virus type 1 capsid was fitted to an electron density map of the whole complex, revealing a relative orientation of subunits different from that in the corresponding crystal structure (44).
acting proteins. We have proposed a single unified approach that can leverage all information to describe a macromolecular structure (14, 39, 45). This approach consists of an iterative series of four steps, including 1) generation of data informative about the structure being determined, 2) design of system representation and translation of the data into spatial restraints, 3) calculation of an ensemble of structures that satisfy the spatial restraints, and 4) an analysis of the ensemble. In this procedure, spatial restraints derived from data about the structure are summed into a scoring function that assesses how well a structural model of an assembly agrees with the data. The scoring function is used to optimize the structural models and to generate a final ensemble of solutions that agrees with the data as much as possible. This four-step approach, by design, benefits from synergy among the input data sets, 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 the final structural models.

**PROTEOMICS AS A KEY DATA SOURCE FOR INTEGRATIVE MODELING**

Proteomics techniques have emerged as a powerful tool for mapping protein interactions in the cell. However, data produced by these techniques are rarely formally incorporated into macromolecular structure determination efforts. Here, we focus on the potential of proteomics techniques to contribute to the integrative modeling of macromolecular assemblies. Specifically, we describe how protein binding and association data can be interpreted as spatial restraints on a protein complex and thus reduce ambiguity in its structural description. These ideas have already been applied to determine the molecular architecture of the NPC (13, 39) and a pseudoatomic model of the 20 S/AAA-ATPase ring of the 26 S proteasome (10, 40, 46). Below, we illustrate our integrative modeling approach by using real experimental data to determine the known architecture of the human RNA polymerase II.

**INTEGRATIVE STRUCTURE CHARACTERIZATION OF HUMAN RNA POLYMERASE II (RNAPII)**

The eukaryotic RNAPII is a central multiprotein machine that synthesizes messenger RNAs and small nuclear RNAs. It is composed of 12 protein subunits with a total molecular mass of 514 kDa (Fig. 2). Ten subunits (Rpb1, Rpb2, Rpb3, Rpb5, Rpb6, Rpb8, Rpb10, Rpb11, and Rpb12) form a structurally conserved core, whereas the Rpb4-Rpb7 heterodimer is located on the periphery (47, 48). Although the atomic structure of the Saccharomyces cerevisiae RNAPII has been solved by x-ray crystallography (49), the human RNAPII (H-RNAPII) has not been determined at atomic resolution mostly because of difficulties in obtaining sufficient quantities of pure sample (50). However, the molecular architecture of the H-RNAPII can be informed by that of its yeast homolog based on the homology between their constituent proteins (50).

Below, we demonstrate that our integrative structure determination procedure can be used to accurately model the known architecture of H-RNAPII using only proteomics-derived protein interactions, an electron density map at 20-Å resolution, comparative models of the protein subunits based on yeast and human crystallographic structures, and geometric complementarity between the interacting subunits. We describe the input data used for the modeling, the translation of these data into spatial restraints, an optimization procedure for determining the models that satisfy the restraints, and an analysis of the resulting set of solutions. We use a previously determined crystallographic structure of the full complex in yeast (51) to evaluate the results.

---

**TABLE I**

<table>
<thead>
<tr>
<th>Restraint</th>
<th>Description</th>
<th>Source of information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excluded volume restraint a</td>
<td>Prevents steric clashes between system particles</td>
<td>Physical first principles</td>
</tr>
<tr>
<td>Geometric complementarity restraint a</td>
<td>Restrains a protein interface to the tightest possible packing</td>
<td>Physical first principles</td>
</tr>
<tr>
<td>Statistical potential restraint</td>
<td>Restrains a structure to have contact frequencies similar to those in structurally defined complexes</td>
<td>Physical first principals, all previously determined protein structures</td>
</tr>
<tr>
<td>Distance restraint a</td>
<td>Restrains the distance between two particles</td>
<td>FRET, BRET, cross-linking, homology to a known structure</td>
</tr>
<tr>
<td>Protein localization restraint</td>
<td>Restrains a protein to a specific position</td>
<td>Immuno-EM, gold labeling, GFP labeling</td>
</tr>
<tr>
<td>Protein connectivity restraint a</td>
<td>Restrains all proteins in a set to interact directly or indirectly</td>
<td>Affinity purification</td>
</tr>
<tr>
<td>Angle restraint</td>
<td>Restrains the angle between three particles</td>
<td>EM, SAXS, homology to a known structure</td>
</tr>
<tr>
<td>Complex diameter restraint</td>
<td>Restrains the distance between the two most distant particles in a protein or complex</td>
<td>EM, SAXS</td>
</tr>
<tr>
<td>Symmetry restraint</td>
<td>Maintains the same configuration of equivalent particles across multiple symmetry units</td>
<td>EM, SAXS, homology to a known structure</td>
</tr>
<tr>
<td>EM quality-of-fit restraint a</td>
<td>Restrains the model to overlap with a density map</td>
<td>EM, SAXS</td>
</tr>
<tr>
<td>Radial distribution function restraint</td>
<td>Restrains the correlation between experimentally measured and computed radial distribution functions</td>
<td></td>
</tr>
</tbody>
</table>

---

a Restraints used to determine the structure of H-RNAPII.
Molecular Architecture from Proteomics

Data Gathering
- MODBASE
- EMDatabank.org
- BioGRID
- Comparative models
- EM density map
- Proteomics data

Data Translation into Spatial Restraints
- Geometric complementarity
- EM quality-of-fit
- Pairwise distance from proteomics
- Connectivity from proteomics

Optimization & Analysis
- Density map segmentation
- MultiFit optimization of EM quality of fit and geometric complementarity
- Filtering by proteomics data
- Best scoring configuration
- Resolved affinity purifications
The percentage of sequence identity between the target subunit and the template as calculated from their alignment used for comparative modeling.

Data Generation by Experiments

Different techniques produce data that differ in types of measured features as well as in the accuracy, resolution, and coverage of the measurements (Fig. 1). An interpretation of the data in terms of a spatial restraint involves identifying the restrained structural components and the allowed values of the restrained feature implied by the data. For example, a result of a cross-linking experiment might be used to restrain the distance between two proteins (40, 52) or within one protein (53); the restraint parameters are a function of the length and flexibility of the cross-linker.

To determine the molecular architecture of the H-RNAPII, we use structural homologs of individual human protein subunits found in the ModBase database (54) (Table II), proteomics data for yeast RNAPII subunits extracted from the BioGRID database (55) (Table III), and an assembly electron density map of H-RNAPII determined at 20-Å resolution by single-particle cryo-EM (50) deposited in the EM data bank (56).

System Representation

The first step in integrative structure determination is deciding on an appropriate representation for the system to be modeled as dictated by the resolution of the available data. At the finest representation granularity, an assembly structure can be represented by particles corresponding to its atoms, each associated with attributes such as position, radius, charge, and mass. Alternatively, a single-particle may be a sphere corresponding to a group of atoms, a whole amino acid residue, a secondary structure segment, a domain, a protein, a “subcomplex” consisting of a subset of proteins in a complete assembly, or even an entire assembly. Given the availability of high accuracy comparative models for the H-RNAPII subunits, we represent the structures of its subunits at atomic resolution. We use atomic models found in the ModBase database of comparative models for domains in known structures (Table II) (57).

Translation of Data into Spatial Restraints

A restraint is a function that reaches its minimum if the restrained feature (e.g., distance) is consistent with the data on which the restraint is based. Beyond that, a restraint can, in principle, have any functional form. For example, a restraint is frequently a harmonic function (of the form $kx^2$ where $x$ is the distance from the mean and $k$ is proportional to the force constant) of the restrained feature. A restrained feature may be any structural attribute of a protein or assembly, including:

Fig. 2. Determining the molecular architecture of human RNAPII. Top, data gathering. Comparative models of the H-RNAPII subunits were obtained from the ModBase database (54). A density map of H-RNAPII at 20-Å resolution (50) was obtained from the EM data bank (56). Proteomics data for S. cerevisiae RNAPII subunits were obtained from BioGRID (Table III) (55). All pairwise direct interactions are visualized in a single graph with solid edges, and each pulldown experiment is presented as a separate graph with dashed edges to indicate the missing underlying binary interaction network. Pulldowns Rpb1-Rpb2-Rpb3-Rpb4-Rpb5-Rpb8 and Rpb1-Rpb2-Rpb3-Rpb8-Rpb10 are missing some edges for clarity. Gray edges indicate interactions present in BioGRID but not in the yeast RNAPII crystallographic structure. Middle, scoring. The scoring function is the sum of the distance (illustrated between Rpb4 and Rpb7), connectivity (illustrated between Rpb1, Rpb2, Rpb3, Rpb8 and Rpb10), EM quality-of-fit (illustrated between the H-RNAPII density map and Rpb1), and geometric complementarity (illustrated between Rpb4 and Rpb7) restraints. Bottom, optimization. The configuration of the subunits in H-RNAPII was optimized using an extension of the divide-and-conquer MultiFit protocol to incorporate proteomics-based restraints. The optimization procedure resulted in a single model that satisfied all of the input restraints.

### Table II

<table>
<thead>
<tr>
<th>Subunit (name, UniProt accession no.)</th>
<th>Sequence identity (%)</th>
<th>Number of residues, residue range</th>
<th>Template (Protein Data Bank code and chain, residue range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rpb1, P24928</td>
<td>55</td>
<td>1970, 11–1475</td>
<td>1/6a H, 7–1445</td>
</tr>
<tr>
<td>Rpb2, P30876</td>
<td>63</td>
<td>1174, 15–1171</td>
<td>2vum B, 20–1216</td>
</tr>
<tr>
<td>Rpb3, P19387</td>
<td>47</td>
<td>275, 1–264</td>
<td>1kw C, 6–263</td>
</tr>
<tr>
<td>Rpb4, O15514</td>
<td>73</td>
<td>2120, 14–1209</td>
<td>1hnn A, 10–73</td>
</tr>
<tr>
<td>Rpb5</td>
<td>52</td>
<td>100, 142, 14–2</td>
<td>2c35 A, 14–142</td>
</tr>
<tr>
<td>Rpb6, P61218</td>
<td>100</td>
<td>100, 142, 14–2</td>
<td>2c35 A, 14–142</td>
</tr>
<tr>
<td>Rpb7, P62487</td>
<td>100</td>
<td>100, 172, 1–171</td>
<td>2c35 B, 1–171</td>
</tr>
<tr>
<td>Rpb8, P52434</td>
<td>100</td>
<td>100, 150, 1–150</td>
<td>2f3i A, 1–150</td>
</tr>
<tr>
<td>Rpb9, P53854</td>
<td>47</td>
<td>100, 125, 1–125</td>
<td>1twf I, 1–150</td>
</tr>
<tr>
<td>Rpb10, P62875</td>
<td>73</td>
<td>73, 67, 1–65</td>
<td>1twf J, 1–65</td>
</tr>
<tr>
<td>Rpb11, P54345</td>
<td>52</td>
<td>1–105, 117</td>
<td>1kw K, 1–105</td>
</tr>
<tr>
<td>Rpb12, P53803</td>
<td>38</td>
<td>1–70, 70</td>
<td>2vum L, 70</td>
</tr>
</tbody>
</table>

* The percentage of sequence identity between the target subunit and the template as calculated from their alignment used for comparative modeling.

### Table III

<table>
<thead>
<tr>
<th>Interacting subunits</th>
<th>Source method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rpb1, Rpb2, Rpb10</td>
<td>Affinity capture-MS (72)</td>
</tr>
<tr>
<td>Rpb1, Rpb2, Rpb3, Rpb4, Rpb5, Rpb8</td>
<td>Affinity capture-MS (72)</td>
</tr>
<tr>
<td>Rpb1, Rpb2, Rpb8</td>
<td>Affinity capture-MS (72)</td>
</tr>
<tr>
<td>Rpb1, Rpb2, Rpb3, Rpb8, Rpb10</td>
<td>Affinity capture-MS (72)</td>
</tr>
<tr>
<td>Rpb1, Rpb2, Rpb6</td>
<td>Affinity capture-MS (72)</td>
</tr>
<tr>
<td>Rpb1, Rpb5</td>
<td>Y2H (58, 59)</td>
</tr>
<tr>
<td>Rpb1, Rpb8</td>
<td>Y2H (60)</td>
</tr>
<tr>
<td>Rpb1, Rpb9</td>
<td>Y2H (58)</td>
</tr>
<tr>
<td>Rpb2, Rpb3</td>
<td>PCA (67)</td>
</tr>
<tr>
<td>Rpb2, Rpb6</td>
<td>PCA (67)</td>
</tr>
<tr>
<td>Rpb2, Rpb10</td>
<td>Y2H (58)</td>
</tr>
<tr>
<td>Rpb3, Rpb11</td>
<td>Y2H (58), reconstituted complex (65), PCA (67)</td>
</tr>
<tr>
<td>Rpb4, Rpb7</td>
<td>Y2H (61–66), reconstituted complex (69)</td>
</tr>
</tbody>
</table>
Contact, proximity, charge, distance, angle, chirality, surface area, volume, excluded volume, shape, symmetry, and localization of particles or sets of particles (Table I). Below, we highlight some restraints in the context of the H-RNAPII structure determination process.

Dealing with Ambiguity

Structural interpretation of data can be ambiguous, especially for proteomics data sets. For instance, if multiple copies of a protein exist in an assembly, a protein-protein interaction derived from a proteomics experiment may not be uniquely assigned to a specific pair of copies. Such ambiguous information must be translated into a restraint that considers all possible structural interpretations of the data; for example, an interaction between two protein types in an assembly with two symmetry units can occur either between the protein copies within each unit or between proteins across the two units (or both). We refer to such restraints as conditional restraints (45).

Distance Restraints from Proteomics

We used direct physical interactions between eight pairs of eukaryotic RNAPII subunits as determined by the yeast two-hybrid (Y2H) system (58–66), protein complementation assays (67), co-localization (47), and complex reconstitution experiments (68, 69) (Table III). These interacting pairs were retrieved from the BioGRID database. Because we aim here to illustrate only what proteomics could do for structure determination, we selected true positive pairwise interactions and ignored the false positives; a discussion of techniques for addressing false positive interactions follows under “Dealing with Incorrect Data, Incomplete Data, and Multiple States”. There are also “indirect” interaction data in BioGRID. However, because BioGRID does not annotate which interactions are physical as opposed to indirect, we encoded as contact distance restraints only those experimentally measured interactions that have been detected by “pairwise” methods listed above.

In general, distance restraints may operate on multiple scales, ranging from the distance between two atoms or residues to the distance between two protein centers in an assembly. For example, if a direct interaction between two proteins has been identified, we may apply a restraint that penalizes deviations from a specified distance between the two protein centers. This distance restraint scores equally all relative orientations between the two proteins with the same intercenter distance. When the shape of the interacting proteins is known, we can achieve a more accurate score at the cost of additional computational time by restraining the distance between the closest pair of particles across the protein-protein interface. Because we do not know a priori which two atoms, residues, or domains are closest to each other, this ambiguity must be handled by a conditional restraint.

Connectivity Restraints from Proteomics

In addition to the pairwise interactions described in the previous section, we also chose to use five sets of physically interacting RNAPII subunits as revealed by affinity purification and mass spectrometry (Table III). We searched three major large scale proteomics data sets (70, 71, 72) for all sets of interacting components that consist of RNAPII subunits only. We then disregarded sets of more than six subunits because such large affinity purification sets are relatively uninformative about the RNAPII structure (their inclusion does not significantly alter the results of our calculations; data not shown). In addition, because the majority of the sets (71 of the 103) were found in the Krogan et al. (72) data set, we used only the Krogan et al. (72) data set for our calculations. For affinity purification data, we know that at least one copy of each protein in a set directly interacts with at least one copy of another protein in the set; however, affinity purification data do not provide information on the stoichiometry of the proteins in the set, the number of complexes with distinct stoichiometry and configuration, or exactly which binary interactions occur, thus resulting in a great deal of ambiguity in the structural interpretation of the results. Because of this ambiguity, each affinity-purified set is encoded as a connectivity restraint that optimizes the assignment of binary interactions to proteins in the set along with the configuration of proteins (39). A putative binary interaction network for the proteins that best satisfies all available data for the system is assigned during each evaluation of the connectivity restraint during the optimization procedure.

Quality-of-fit Restraint from an Electron Density Map

The fit of a model into an assembly density map is usually assessed by a cross-correlation measure between the assembly density and the model smoothed to the resolution of the map (22). Here, the configurations of the H-RNAPII subunits were restrained to fit an electron density map of the H-RNAPII complex (50).

Excluded Volume Restraint

Molecules take up space that cannot be occupied by other molecules. This space filling property provides a key restraint on the conformations of the assembly. If the atomic structure is known, as is the case for H-RNAPII, the van der Waals radius for each atom is typically used to define the excluded volume (73). When the structure of a molecule is not known, it can be represented by a sphere; the volume of the sphere can be estimated from its composition (e.g. the number of residues in a protein (74)).

Geometric Complementarity Restraint from First Principles

Protein-protein interfaces are typically geometrically complementary, characterized by tight packing with little space
between them. This geometric complementarity is commonly used as a restraint in protein-protein docking (75, 76). Because atomic models are used for H-RNAPII subunit structures, this consideration was enforced with an explicit restraint. The geometric complementarity restraint may be less informative if used on coarsely represented subunits.

**Additional Restraints**

Although not applied in our integrative structure determination of H-RNAPII, many additional restraint types can also be used.

**Radial Distribution Restraint**

An approximate radial distribution function of an assembly can be measured by an SAXS experiment (24, 25). Correspondingly, the SAXS restraint on a model can penalize the difference between the experimental and computed radial distribution functions (77). This restraint was used, for example, to select among several putative configurations of domains for the chaperone Hsp90 (78).

**Symmetry Restraint**

Symmetry is a recurrent theme in macromolecular assembly structures (79–81). For example, cyclic, helical, dihedral, and icosahedral symmetries are found in many important molecular machines such as viruses, the NPC, and chaperonins. The similarity between corresponding particle configurations in each symmetry unit can be enforced by imposing a restraint that maintains the same particle-particle distances within each unit (39, 82).

**Physical Energy and Statistical Potential Restraints**

Positions and orientations of interacting proteins can also be restrained by potentials based on the laws of physics (83–86) as well as statistical potentials extracted from databases of known protein structures (87–92). For example, a statistical potential can be derived from the observed distance distributions or contact frequencies of different atom type pairs in structurally defined proteins or complexes (93–96).

**Combining Restraints into a Scoring Function**

Once the data sets are encoded as restraints, they are combined into a scoring function, usually the sum of all the restraints. In this sum, the degree of uncertainty encoded by each restraint is effectively its weight. Ideally, the restraint on a spatial feature should be a probability density function on the feature given the corresponding measurement (39); for example, the lower and upper bounds on a distance should reflect the uncertainty of the corresponding distance measurement and its interpretation.

**Calculation of an Ensemble of Structures by Satisfaction of Spatial Restraints**

Next, all structural models that minimize the scoring function and therefore fit the original data must be found. An optimization procedure performs a search through the space of all possible macromolecular complex configurations by minimizing the violations of all restraints simultaneously. It is helpful to have many optimization methods available and to choose one that works best with a given representation and set of restraints. We have implemented several different optimizers as part of the Integrative Modeling Platform package. These optimizers can be classified as whole-system and divide-and-conquer optimizers.

**Whole-system Optimizers**

In this class of optimizers, an algorithm usually starts with a random initial configuration. The space of conformations is then explored iteratively by computing the next assembly configuration based on the values of all restraints for the configuration in the current optimization step with the intent of moving closer to the minimum value of the scoring function. Optimizers in this class include traditional conjugate gradients (97), quasi-Newton (98) and molecular dynamics schemes (99), Monte Carlo procedures as well as more sophisticated methods such as self-guided Langevin dynamics (100), and the replica exchange protocol (101). Because of the stochastic nature of these optimizations and the need to find all good scoring solutions, many independent runs are generally performed, each starting with a different random initial configuration.

**Divide-and-conquer Optimizers**

Divide-and-conquer optimizers can separate the particles and restraints in a system into smaller “suboptimizations,” ultimately resulting in more rapid sampling of structures. We have recently suggested a general divide-and-conquer approach to more efficiently sample protein assembly configurations (32). In this approach, the set of variables is decomposed into relatively uncoupled but potentially overlapping subsets that can be sampled independently of each other (i.e. are not required to be sampled together in a single calculation and can be sampled in parallel) and then efficiently gathered to compute the global minimum. The strength of this approach is derived from the decomposition procedure, which helps to reduce the size of the search space from exponential in the number of components in the whole system to exponential in the number of components in the largest subset. Similar ideas have been used for various modeling tasks such as side chain packing (102–104), sequence-structure threading (103), ab initio RNA folding (105), and prediction of quaternary structures of multiprotein complexes (106).
Use of Restraints to Restrain the Search Space for Optimization

Efficiency can be increased by designing an optimization scheme to avoid considering configurations that clearly violate a subset of the data. Examples include segmenting an electron density map for the entire assembly into components that likely correspond to individual proteins prior to fitting the assembly proteins into the map (32), eliminating geometrically unlikely protein-protein docking solutions (75, 107), and restricting the search space to symmetric configurations (108, 109).

Human RNAPII Optimization

For our H-RNAPII example, we used the sum of the distance, connectivity, EM quality-of-fit, and geometric complementarity restraints described above as a scoring function. The configuration of the subunits in H-RNAPII was optimized using an extension of the divide-and-conquer MultiFit protocol (Fig. 2) (32, 33). We began by segmenting the electron density map into 12 regions, each one of which served to localize one of the 12 constituent H-RNAPII proteins. This procedure resulted in 479,001,600 (12!) possible H-RNAPII subunit configurations. Next, we eliminated all H-RNAPII subunit configurations that did not satisfy a majority of the proteomics restraints (Table III), keeping only 2,576 configurations for further refinement. We then refined each of these 2,576 configurations to optimize the EM quality-of-fit and geometric complementarity restraints using the standard MultiFit protocol (32); 63 of the 2,576 configurations resulted in refined models with “good” scores. These models had equivalent positions for Rpb1, Rpb2 and Rpb3; however, the models varied in the positions of the remaining subunits. Finally, we filtered the 63 models by all proteomics restraints, resulting in a single model that satisfied all proteomics restraints as well as the EM quality-of-fit and geometric complementarity restraints (Fig. 3).

Analysis of the Ensemble

Precision

There are three possible outcomes of an optimization procedure. First, if only a single structural model satisfies all restraints and thus all input information, there is probably sufficient data for prediction of the unique native state. Second, if two or more different models are consistent with the restraints, the data are insufficient to define the single native state, or there are multiple significantly populated states. 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. For example, it might be that a complex exists in several functional states and that the available data cover more than one of them.

In the case of the H-RNAPII model, optimization resulted in a single model that satisfied all the data. Thus, sufficient information was available to predict the positions and orientations of the H-RNAPII subunits. The ensemble of possible models in the absence of proteomics data was much larger (2,576 coarse configurations) and defined the structure far less precisely. Therefore, proteomics data were crucial for providing an unambiguous determination of a precise molecular architecture of H-RNAPII.

Accuracy

Assessing the accuracy of a structure, defined as the difference between the model and the native structure, is difficult but important (45). It is impossible to know with certainty the accuracy of the proposed structure without knowing the real native structure. Nevertheless, our confidence can be modulated by five considerations: (a) self-consistency of independent experimental data; (b) structural similarity among all configurations in the ensemble that satisfy the input restraints; (c) simulations where a native structure is assumed, corresponding restraints are simulated from it, and the resulting calculated structure is compared with the assumed native structure; (d) confirmatory spatial data that were not used in the calculation of the structure (e.g. a criterion similar to the crystallographic free R-factor (110) can be used to assess both the model accuracy and the harmony among the input restraints); and (e) patterns emerging from a mapping of independent and unused data on the structure that are unlikely to occur by chance (13, 39).

In the case of H-RNAPII, we can estimate the accuracy directly because we know the crystallographic structure of the yeast RNAPII, which is likely to be highly similar to that of H-RNAPII (50) (c.f. the high degree of sequence similarity between yeast and human subunit orthologs (Table II) and the high correlation coefficient of 0.65 between the crystallographic yeast RNAPII structure and the electron density map of H-RNAPII). The H-RNAPII model clearly recapitulates the molecular architecture of yeast RNAPII (Fig. 3), preserving all of its protein interactions. More quantitatively, the subunits in the H-RNAPII model share a Cα root mean square deviation (RMSD) of only 11.4 Å with the human subunits individually superposed on their orthologs in the yeast RNAPII structure.

Dealing with Incorrect Data, Incomplete Data, and Multiple States

Proteome-wide protein-protein interaction maps have been produced by high throughput assays, such as affinity purification (11, 71) and yeast two-hybrid system (111–116). However, these data sets can be limited in three respects (117–119). First, the data can be incomplete in the sense that a
number of interactions insufficient to describe the studied system were detected. Second, the data can be inaccurate in the sense that some detected interactions do not apply to the studied system. Third, the data can be “frustrated” in the sense that different subsets of the data apply to compositionally and/or conformationally different states of the studied system. For example, prior to filtering, a significant fraction of the affinity purification data for RNAPII subunits corresponds to false positive interactions (defined as a set of interacting subunits that do not have a connecting interaction path in the crystallographic structure of the complex (51)). In particular, 31, 35, and 0% of the 71, 26, and six affinity purification sets with two or more RNAPII subunits as reported by Krogan et al. (72), Gavin et al. (71), and Ho et al. (70) were false positives, respectively. In addition, 33% of the 12 reported binary interactions extracted from the BioGRID database were false positives.

A reasonable goal of structural modeling is to find the minimum number of system states that account for the observed data. If the data sets are correct and complete and describe a single state of the system, the optimization procedure should, in principle, result in a single solution that satisfies the data. If the data sets are inaccurate or incomplete, irrespective of the number of system states, the sampling should result in different states, some of which may or may not satisfy all the data. Next, we describe these possible outcomes in more detail.

Correct, Complete Data, Single State—The optimization procedure should result in a single solution that satisfies all restraints. If the data set is redundant, it is possible to cross-validate the solution by rerunning the modeling procedure using only random subsets of the data (120).

Correct, Incomplete Data, Single State—The optimization procedure should produce multiple solutions, all of which should satisfy all restraints. For example, this situation may occur when the proteomics data do not apply to all subunits of a system or only cover a small subset of interactions. It is possible to identify the least precisely localized components of the system within the set of solutions, directing future experiments for the largest possible gain in the next iteration of integrative modeling.

Incorrect, Complete Data, Single State—The optimization procedure should produce multiple solutions, each satisfying a fraction of the restraints. If there are redundant correct data, it may be possible to identify the conflicting incorrect data by cross-validation.

same two views as in a–d further illustrate the high similarity between the model and the reference. In the coarse representation, sets of 30 contiguous residues are shown as a single bead. III, i and j, protein contact maps for the H-RNAPII model and the reference structure (white, no contact; gray, weak contact; black, contact). The maps are essentially identical, differing only in the interactions of Rpb6 with Rpb2 and Rpb3, and Rpb1 with Rpb12.
Incorrect, Incomplete Data, Single State—The optimization procedure should produce multiple solutions, each satisfying a fraction of the restraints. It is difficult to identify the incorrect data as well as to detect a solution corresponding to the correct state. This situation arose in a preliminary attempt to model the molecular architecture of the 19 S regulatory particle of the 26 S proteasome (46). In that case, we have concluded that additional data are required.

Multiple States—Even when all data are correct and complete, the optimization procedure may be inadequate and produce multiple solutions, each satisfying only a fraction of the restraints. The same outcome is obtained when using incorrect data. Thus, multiple states are difficult to deconvolve from incorrect data (such as false positive interactions from proteomics).

In conclusion, when no solution is found that satisfies all data, it is difficult to identify the correct state(s). Formally, a similar problem exists in protein structure determination based on NMR spectroscopy. There, structural features, such as interatomic distances and dihedral angles, are obtained experimentally and used in the form of spatial restraints for finding the set of structural models that satisfies these restraints. One approach to dealing with incorrect data for one or more states looks at the frequency with which each restraint is violated in an ensemble of calculated structures (121, 122); if a given restraint is violated often, the bounds on the restraint can be loosened. Other approaches use cross-validation to assess the completeness of the experimental restraints (123). Another development, the inferential structure determination method, formulates structure determination as an inference problem, handling incorrect and incomplete data as well as multiple states in a Bayesian framework (43). Adaptations of these methods and development of new methods should improve future handling of incorrect and incomplete data in integrative structure determination of conformationally and compositionally heterogeneous assemblies.

DISCUSSION

As illustrated above, proteomics techniques can now facilitate the characterization of the structure of macromolecular assemblies via integrative modeling. We have demonstrated that by using atomic subunit structures, an electron density map of their assembly, and proteomics data restraining relative subunit proximities we can extend the scope of macromolecular structure determination beyond what is possible with single methods. Specifically, using the RNAP II structure as an example, we have shown that proteomics data, although traditionally not considered a source of formal structural information, can play a key role in assembly structure determination.

One key challenge for integrating proteomics data into structure determination remains the treatment of assemblies that exist in multiple functional states, corresponding to different configurations and compositions of the assembly. Although integrative methods can already restrain the structure of the modeled assembly based on all available information, some of the proteomics data may in fact apply to only a subset of all functional states of the assembly. For example, proteomics techniques often detect peripheral interactions that are not part of the core assembly but could be vital for one of the biologically relevant states. Thus, future protocols need to be able to simultaneously determine structures for all biologically relevant states. These methods will need to associate specific interactions with specific functionally relevant states of an assembly as well as remove false positive interactions that are not relevant to a given state.

As the quantity and variety of experimental data about macromolecular assemblies grows, integrative structure determination will be vital for characterization of these machines and the corresponding cellular processes. Methods are needed that are more accurate in translation of heterogenous data into spatial restraints as well as combination of these restraints into a scoring function. New sampling and optimization schemes should improve the accuracy and level of detail with which we can describe assemblies. In addition, as a generalization of treating systems with multiple configurations and compositions, we should address the challenge of characterizing the dynamics of macromolecular assemblies by satisfying both spatial and temporal restraints for a system of multiple components. As integrative structure determination techniques advance, we will be able to describe an increasing number of key cellular structures, progressing toward a comprehensive structural, temporal, and logical model of the cell.

Acknowledgments—We thank Frank Alber, Michael P. Rout, Brian Chait, Wolfgang Baumeister, and Friedrich Förster for discussions about integrative structure determination based on proteomics data; Haim Wolfson for collaborating on optimization methods; and Hannes Braberg and Javier Fernandez-Martinez for discussing interpretation of proteomics data.

* This work was supported, in whole or in part, by National Institutes of Health Grants R01 GM54762, U54 RR022220, PN2 EY016525, and R01 GM083960 (to A. Sali).
¶¶ Both authors contributed equally to this work.
|| Supported by the Clore Foundation Ph.D. Scholars program and carried out research in partial fulfillment of the requirements for the Ph.D. degree at Tel Aviv University. To whom correspondence may be addressed. E-mail: kerenl@salilab.org.
‡‡ To whom correspondence may be addressed. E-mail: jeremy@salilab.org.
 §§ Supported by a Weizmann Institute Advancing Women in Science postdoctoral fellowship.
¶¶¶ Supported by the Sandler Family Supporting Foundation, National Science Foundation Grant IIS-0705196, Ron Conway, Mike Homer, Hewlett-Packard, NetApp, IBM, and Intel. To whom correspondence may be addressed: University of California, MC 2552, Byers Hall at Mission Bay, Suite 503B, 1700 4th St., San Francisco, CA 94158. Tel.: 415-514-4227; Fax: 415-514-4231; E-mail: salli@salilab.org.
Molecular Architecture from Proteomics


rithms in Bioinformatics (Guido, R., and Gusfield, D., eds) pp. 185–200, Springer-Verlag, London


