Annual Review of Biochemistry

Integrative Structure Modeling: Overview and Assessment

Merav Braitbard,[1] Dina Schneidman-Duhovny,[1,2] and Nir Kalisman[1]

[1]Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel; email: nirka@mail.huji.ac.il

[2]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel; email: dina.schneidman@mail.huji.ac.il

Annu. Rev. Biochem. 2019. 88:113–35

First published as a Review in Advance on March 4, 2019

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

https://doi.org/10.1146/annurev-biochem-013118-111429

Keywords

integrative modeling, macromolecular assemblies, protein structure, cross-linking, mass spectrometry, cryo–electron microscopy, cryo-EM

Abstract

Integrative structure modeling computationally combines data from multiple sources of information with the aim of obtaining structural insights that are not revealed by any single approach alone. In the first part of this review, we survey the commonly used sources of structural information and the computational aspects of model building. Throughout the past decade, integrative modeling was applied to various biological systems, with a focus on large protein complexes. Recent progress in the field of cryo–electron microscopy (cryo-EM) has resolved many of these complexes to near-atomic resolution. In the second part of this review, we compare a range of published integrative models with their higher-resolution counterparts with the aim of critically assessing their accuracy. This comparison gives a favorable view of integrative modeling and demonstrates its ability to yield accurate and informative results. We discuss possible roles of integrative modeling in the new era of cryo-EM and highlight future challenges and directions.

Contents

1. INTRODUCTION

Structural biology was dominated for many years by X-ray crystallography and NMR spectroscopy, which contributed thousands of atomic-resolution structures to the Protein Data Bank (PDB) (1). The difficulty in obtaining crystallographic structures, especially for large protein complexes, prompted the more recent development of alternative structural characterization techniques. These include cryo–electron microscopy (cryo-EM), which has recently undergone a resolution revolution (2); mass spectrometry–based techniques, such as cross-linking coupled to mass spectrometry (XL-MS) (3) and hydrogen–deuterium exchange (HDX) (4); small-angle X-ray scattering (SAXS) (5); Förster resonance energy transfer (FRET) (6); and sequence analysis (7). Integrative structure modeling (ISM) has the purpose of combining information from multiple techniques through computation to yield structural models that have better accuracy, resolution, and precision than those possible with any one technique alone (8–10). The rationale for integrative modeling is that the strengths of different techniques complement one another. For example, some methods provide the global shape of a complex, whereas others report on the local proximity of subunits within it.

Cryo-EM: cryo–electron microscopy

XL-MS: cross-linking coupled to mass spectrometry; also known as CX-MS

SAXS: small-angle X-ray scattering

FRET: Förster resonance energy transfer

ISM: integrative structure modeling

Input information: experimental data and any additional information

During the past 15 years, ISM has been applied to a wide variety of structural systems, ranging in size from complexes of two subunits to giant assemblies comprising hundreds of subunits. The structural information also ranges in level of detail from all-atom models to topological maps. All ISM strategies share a common four-stage process: (*a*) gathering input information, (*b*) designing model representation and converting information into a scoring function, (*c*) sampling good-scoring models, and (*d*) analyzing models and information. We begin this review by describing these steps in more detail (Section 2).

At their time of publication, the accuracy of many integrative models was assessed on the basis of how well they converged to a single solution and how consistent this solution was with the available data. In the second part of this review (Section 3), we examine the accuracy of ISM more critically, by comparing the published models with more recent structural information.

2. EXPERIMENTAL AND COMPUTATIONAL ASPECTS OF INTEGRATIVE MODELING

ISM iterates through the process of gathering data, proposing structural models, and then gathering more data to validate and refine those models. All ISM strategies share a common four-stage process:

1. Gathering input information: This information consists of data from wet laboratory experiments such as those listed above, statistical tendencies such as atomic statistical potentials, molecular mechanics force fields, and anything else that can be converted into a score on features of a structural model.
2. Designing the model representation and scoring function: The resolution of the representation depends on the quantity and resolution of the available information and should be commensurate with the resolution of the final models; different parts of a model may be represented at different resolutions, and a part may be represented at several different resolutions simultaneously. The scoring function evaluates whether or not a given model is consistent with the input information, taking into account the uncertainty in the information.
3. Sampling good-scoring models: Multiple models are generated, and their scores are improved by subjecting them to a variety of optimization algorithms (e.g., Monte Carlo, gradient decent, etc.). These algorithms change the coordinates of different components of the models in directions that overall improve their scores. The goal is for the models to thoroughly sample the possible conformations of the system. Consequently, the better-scoring models represent the more likely conformations.
4. Analyzing models and information: The ensemble of good-scoring models need to be clustered and analyzed to ascertain ensemble precision and accuracy and to check for inconsistent information. Analysis can also suggest which are likely to be the most informative experiments to perform in the next iteration.

Ensemble: a set of structural models, each one of which is consistent with the data

Ensemble precision: structural variability among models in the ensemble

Accuracy: the difference between the structural model and the true structure(s)

The actual implementations of the ISM approach vary greatly across users and systems. Users may choose one of several comprehensive software packages that cover all of these stages such as the Integrative Modeling Package (IMP) (10). Alternatively, users may implement each stage with a different program or employ their own computer code. A compilation of available software packages is listed in Section 2.5. Additionally, the second part of this review (Section 3) presents a range of published ISM results from which readers may learn of implementations that suit their modeling needs.

2.1. Information Sources in Integrative Modeling

Almost any type of information can be encoded into a scoring function and used in ISM. We describe here the most commonly used types of information (**Table 1**).

2.1.1. Subunit structures. Ideally, ISM would assemble a model for a complex from all-atom structures of its subunit components. Accordingly, one should aim to obtain the highest possible structural coverage for the modeled sequences. Experimental structures of individual subunits determined by X-ray crystallography or NMR spectroscopy are most suitable. If experimental structures are not available, comparative modeling is routinely used (11, 12). Comparative modeling can substantially improve the structural coverage, because reliable templates (sequence identity >30%) are available for ~50% and ~80% of all human and bacterial proteins, respectively (13, 14). To increase structural coverage even further, one can apply ab initio modeling, preferably

Table 1 Types of structural information

Modality	Locality	Number of distance restraints	Accuracy	Conformational heterogeneity?	In vivo possibility?
Cryo-EM single particle analysis	Global	N/A	Atomic resolution to 30 Å	Yes	No
SAXS	Global	N/A	~30 Å (shape dependent)	Yes	No
X-ray crystallography	Local	N/A	Atomic resolution	No	No
NMR spectroscopy	Local	N/A	Atomic resolution	Yes	No
Comparative modeling	Local	N/A	Template dependent	N/A	No
Ab initio modeling	Local	N/A	Variable	N/A	No
XL-MS	Local	10s–1,000s	15–20 Å	Yes	Yes
Correlated mutations	Local	10s–1,000s	6–10 Å	Yes	Yes
Pairwise interactions	Local	1–10	10–20 Å	No	Yes
FRET	Local	1–10	~60 Å	Yes	Yes

Abbreviations: cryo-EM, cryo–electron microscopy; FRET, Förster resonance energy transfer; N/A, not applicable; SAXS, small-angle X-ray scattering; XL-MS, cross-linking coupled to mass spectrometry.

guided by contact information from sequence coevolution analysis (7, 15). It is often the case that some sequence regions are left uncovered by these approaches. Such regions are then handled in the ISM process by coarse representations (Section 2.2).

2.1.2. Cryo–electron microscopy. Cryo-EM is becoming the most common method for structural characterization of large macromolecular complexes (16). It is also the source of global structural information in many ISM studies. The main work mode of cryo-EM in structural biology is single-particle analysis [>75% of the maps in the electron microscopy (EM) data bank (17)], which gives the three-dimensional (3D) electron density map of a complex. In this mode, the particles are confined within a thin layer of solution and frozen so rapidly that they become embedded in amorphous ice in random orientations. Imaging of the particles with a transmission electron microscope yields micrographs of their two-dimensional (2D) projections. As the particles are oriented randomly relative to the electron beam, each projection provides an image of the particle from a different direction. Dedicated software packages then reconstruct a 3D volume from thousands of individual projections (18, 19). The procedure is similar to computerized tomography in medicine, which renders the 3D volume of a skull from multiple X-ray exposures around the head.

Cryo-EM is advantageous for the study of large complexes because it does not require crystallization. Conformational flexibility is a major obstacle, which can sometimes be addressed by reconstruction of multiple conformations within the framework of cryo-EM. In most cases, the resolutions of structures obtained by cryo-EM are inferior to those obtained by crystallography. Until approximately five years ago, the resolution of electron density maps from single-particle analysis was typically 20–30 Å. At such resolution, the maps described the global shape of a macromolecular complex but revealed little information about the arrangement of subunits and domains within the map. This situation prompted the development of ISM to incorporate sources of local information and reveal the detailed architectures of the complexes. A dramatic improvement, a so-called resolution revolution (2), resulted from the introduction of direct electron detectors (20). New 3D reconstruction software packages such as RELION (18) and cryoSPARC (19) were

developed to analyze the resulting information. Consequently, more than half of the electron density maps reported last year had resolutions better than 6 Å, with many claiming less than 4 Å (21).

Besides single particle analysis, other working modes of cryo-EM are also relevant to ISM. Electron cryo-tomography is imaging 3D volumes of whole cells or cell slices and can be used to obtain the electron density maps of macromolecular complexes in situ. Therefore, it is well suited for the study of large assemblies that are hard to purify intact such as the nuclear pore complex (22) or adhesion focal points (23). EM imaging of complexes labeled with antibodies enables localization of specific subunits in the density maps (24). Finally, the use of 2D projections per se is also applicable to ISM (25–27) and is beneficial for cases in which the particles are embedded in the ice in a limited number of orientations that are insufficient for accurate 3D reconstruction.

2.1.3. Cross-linking coupled to mass spectrometry. XL-MS (also known as CX-MS) is a rapid and efficient experimental approach to obtain distance restraints between residue pairs within a complex (3, 28, 29). It is based on the incubation of the complex of interest under physiological conditions with a bifunctional cross-linking reagent. Following proteolytic cleavage, further analysis by mass spectrometry identifies pairs of protein residues that underwent cross-linking. Every identified cross-link indicates that the two cross-linked residues were in proximity within the context of the complex. XL-MS therefore creates a distance restraint between the two residues for use in structural modeling. Typically, tens to hundreds of cross-links can be identified from a complex of even moderate size, leading to a similar number of distance restraints. For modeling purposes, the possible range of distances between the Cα atoms of cross-linked residues was determined by analysis of XL-MS on known structures (3, 30). For the widely used cross-linking reagent disuccinimidyl suberate (DSS), this distance was established to be between 10 and 25 Å (3). Interestingly, shorter cross-linking reagents only lead to slightly shorter bridging distances, as these distances are mainly determined by protein flexibility and side-chain lengths (30).

XL-MS has several important advantages that are complementary to other structural approaches. Foremost, it is applicable to nearly any protein complex, regardless of size or flexibility (31, 32), including membrane assemblies (33–35). Because flexibility and solubility are major obstacles to crystallography and cryo-EM, XL-MS is the method of choice for some systems. Another advantage of XL-MS is its potential to probe conformational changes (36–39). At the same time, XL-MS also has three inherent disadvantages. First, the assignment of cross-links can be ambiguous. The data set may contain a mixture of cross-links from multiple conformational states or multiple copies of the same subunit. Computational modeling must take these types of ambiguity into account. Second, cross-link locations are often distributed unevenly across the complex, with some interacting subunits having few or no cross-links. This is often caused by the lack of reactive amino acid residues at their interfaces and can be sometimes addressed by alternative cross-linking chemistries. Finally, cross-linking may lead to the formation of aggregates and identification of non-native contacts. This problem can be avoided by the use of low concentrations of cross-linking reagents, but care must be taken during the experiment and modeling to ensure that the data are derived from a non aggregated state.

Over the past decade, improvements in mass spectrometry instrumentation and better availability of analysis software greatly facilitated the use of XL-MS. In fact, any laboratory with access to a proteomics facility can now employ this rapid and inexpensive approach for structural studies (40). Together with its other advantages, we foresee a growing use of XL-MS in ISM.

2.1.4. Small-angle X-ray scattering. SAXS is becoming a widely used technique for low-resolution structural characterization of molecules in solution (5, 41–47). A key strength of SAXS

is that it is informative about the shapes of macromolecules as large as 1,000 Å in the 10–50-Å resolution range at near physiological conditions. The experiment is performed with an ∼1.0 mg/mL sample in an ∼15 μL volume and usually takes only a few minutes on a well-equipped synchrotron beamline (42). Moreover, SAXS profiles can be rapidly collected for a variety of experimental conditions, such as ligand-bound and unbound protein samples, ligand titration series, and different temperatures or pH values. The SAXS profile of a macromolecule, I(q), is computed by subtracting the SAXS profile of the buffer from the SAXS profile of the macromolecule in the buffer. The profile can be converted into an approximate distribution of pairwise atomic distances of the macromolecule (i.e., the pair-distribution function) via a Fourier transform. In contrast to static structures solved by X-ray crystallography, SAXS profiles, although limited in resolution, contain information about conformational and compositional states of the system in solution (48–50). The major challenge of computational modeling is to determine the set of states that is consistent with the data (48, 51, 52).

2.1.5. Sequence information. The growth in the number of sequences has greatly outpaced the growth in the number of PDB structures due to high-throughput sequencing technologies (53). Sequence analysis has been traditionally used to define evolutionary conserved positions that indicate importance for folding, function, interactions, and dynamics (54). Growth in the number of sequences enabled analysis of coevolving residue positions. Because specific residue interactions are required for stable and functional proteins, coevolution is observed for contacting amino acids. Methods based on sequence coevolution for prediction of structural contacts have been very successful in ab initio structure prediction (7, 15, 55–57). Moreover, analysis of coevolving positions can be used to predict protein–protein interactions, structures of transmembrane proteins, conformations of disordered regions, and mutation effects (56, 58, 59).

2.1.6. Förster resonance energy transfer. FRET is a powerful technique for studying protein structure, dynamics, and interactions both in vitro and in living cells (6, 60). FRET occurs when two spectrally matched fluorescent molecules are in proximity and excitation energy is transferred from the donor to the acceptor fluorophore through nonradiative dipole–dipole coupling (61). The efficiency of this process is inversely proportional to the sixth power of the distance between donor and acceptor, making FRET extremely sensitive to small changes in distance. It has been used to probe distances over the range of 1 to 10 nm, resulting in spatial restraints for modeling of both structure and dynamics (62, 63). Similarly to XL-MS and SAXS, single-molecule FRET provides rich information about the conformational states of a system (64).

2.2. Designing the Model Representation and Scoring Function

Once structural information has been obtained from the methods described above, ISM may be performed. One begins ISM by deciding on model representation and converting information into scoring functions, followed by sampling good-scoring models and analyzing models and information. Experimental information can be used in all of these steps. For example, all-atom models are usually used in model representation, cross-links can be used in scoring functions, EM maps can be used to restrict sampling, and SAXS profiles can be used for model validation.

2.2.1. Model representation. System representation is a central part of computational modeling (**Figure 1**). It is influenced by several factors, particularly the availability of structural coverage for the subunits. At the most detailed level, subunit structures can be represented by atomic

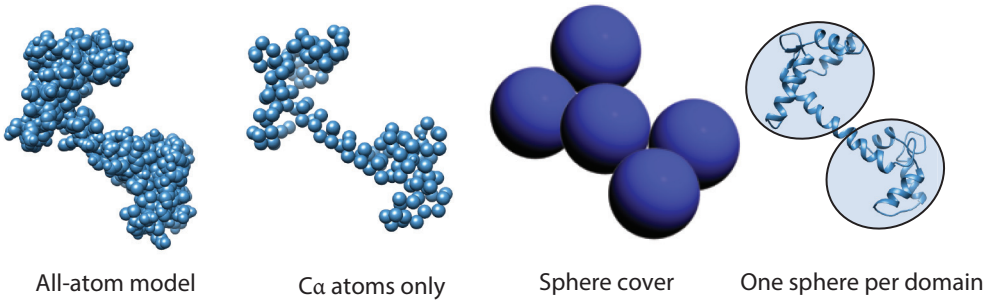

Figure 1

Selected representation options for one subunit.

positions. However, even when this level of detail is possible, one can decide on a coarser representation for ISM depending on the available information and sampling requirements. For example, if cross-linking information is available, the natural representation is one particle per residue (centered at the Cα atom), because the scoring function only measures the distances between Cα atoms. Reduced representation enables more sampling iterations at the same computational time frame. When high-resolution structures are not available, coarse-grained spheres (one sphere per several residues, per secondary structure, or even per domain) or volumes are required. In the case of partial structural coverage, a mixed representation is often used: Subunits with high-resolution structural models are represented in all-atom detail, whereas subunits without structures are represented by coarse-grained spheres.

2.2.2. Scoring functions. In ISM, the scoring function is a combination of several terms depending on the available information and system representation. In this section, we first describe more general scoring functions that are applicable to almost any system, followed by information-specific scores.

2.2.2.1. Excluded volume. Molecules cannot overlap with each other in space, as steric clashes between the atoms are impossible. The goal of the excluded volume score is to enforce this property on the generated models. In many cases, the representation and sampling are not sufficiently accurate, so the excluded volume score has to allow for a small amount of steric overlap. This is often achieved by a smoothed, repulsive van der Waals potential (65).

2.2.2.2. Shape complementarity. Geometric shape complementarity is a necessary condition for formation of protein–protein interfaces, because interfaces in solved structures are characterized by tight packing. A shape complementarity score is commonly used in protein–protein docking by employing algorithms that are based on fast Fourier transforms, which express shape complementarity as a correlation (66, 67). The geometric complementarity restraint is less discriminative when used at coarser representations that do not approximate shape accurately.

2.2.2.3. Physico-chemical complementarity. In addition to shape complementarity, the physico-chemical complementarity between the proteins in the assembly can be assessed (68). Most of the scoring functions in this category require atomic- or residue-level representation. They can be classified into two major types: energy functions based primarily on a molecular mechanics force field [such as the scoring functions of RosettaDock (65)] and knowledge-based statistical potentials based on distributions of intermolecular features in large databases (69).

2.2.2.4. Distance-based scoring functions. Multiple types of information can be converted into distance-based scores. These scores can be divided into constraints and restraints. Constraints require that the specified distance is within the predefined threshold and discard all the models that violate it. Restraints are a softer version of constraints. They penalize violated distances but do not discard the models. The most prevalent use of constraints is to enforce connectivity between successive residues along a protein chain or, in case of coarser representations, between successive spheres or volumes. Protein–protein interaction data are usually also converted into distance-based scoring terms as restraints (70).

Data points from XL-MS, coevolving sequence positions, and FRET define pairs of residues in a specific distance range. Each of these pairs contributes one distance restraint that is measured between the Cα or Cβ atoms of the corresponding residues. The upper limit on this distance is method dependent, with typical values listed in **Table 1**. The simplest scoring scheme is to count the number of distance violations in a model (71). More sophisticated scores use harmonic or Gaussian functions (26) around the optimal distance of the restraint as well as sigmoidal and Lorentz functions (72). The most objective scoring of models is in principle achieved by a Bayesian-based function. It was successfully adopted for NMR spectroscopy data (73), cysteine cross-linking (36), chemical cross-linking (37), and FRET spectroscopy (63). The Bayesian scoring function can account for most sources of uncertainty in data without overfitting but is more difficult to implement.

2.2.2.5. EM-based scoring functions. The fit between the model and the 3D cryo-EM density map is usually assessed by computing the cross-correlation coefficient (74, 75). When 2D class averages are used, the scoring process is more involved (25, 26). First, hundreds of 2D projections are calculated from the model. Next, each measured class average is compared with all the calculated projections, and the best matching projection is selected. The score is then the sum of cross-correlation coefficients between each class average and its corresponding projection.

2.2.2.6. SAXS-based scoring functions. Modeling approaches that include SAXS information usually compare the theoretical SAXS profile calculated from a model with the experimental SAXS profile. Theoretical SAXS profile calculations from atomic models require spherical averaging to account for the random orientation of the macromolecule in solution. The observed scattering profile is the difference between the scattering of the target macromolecule with its ordered hydration layer and the excluded volume that takes into account the missing scattering of bulk solvent. Therefore, methods for calculating SAXS profiles have to account for the excluded volume of bulk solvent and the hydration layer. As a result, the approaches for profile computation generally differ between methods by how they implement spherical averaging, treatment of the excluded volume, and treatment of the hydration layer (76–79). The χ^2 score is usually used to assess the fit between experimental and computational SAXS profiles.

2.3. Sampling Good-Scoring Models

Effective sampling of large systems is a challenging task due to the many degrees of freedom. Each subunit contributes six degrees of freedom (three rotational and three translational variables) to the system. The total number of degrees of freedom is therefore $6(N - 1)$, where N is the number of subunits. Parts represented by spheres only have translational degrees of freedom. Sampling methods can be classified into discrete and continuous search. The discrete search limits the subunits to a finite set of positions in space (e.g., a grid set within a density map). In the continuous search, there are no restrictions on subunit positions. Sampling methods for continuous

search space can be further divided into two major categories: (*a*) randomized optimization methods, such as conjugate gradients or Monte Carlo, simulated annealing, or genetic algorithms; and (*b*) deterministic sampling strategies, such as divide-and-conquer or docking-based methods.

Multistate models: models that specify two or more coexisting structural states or values for any other parameter

Randomized optimization usually starts with a random initial configuration. The space of conformations is then explored iteratively, by computing the next assembly configuration on the basis of the scores of the current configuration. The major advantage of this sampling strategy is that it can be applied with any scoring function, and as a result, any information type can be used. Owing to the stochastic nature of randomized optimizations, many independent runs are required, each starting from a different random initial configuration. The thoroughness of sampling can be indicated by showing that new independent runs do not result in significantly different good-scoring solutions (convergence test) (80). However, finding all good-scoring solutions is impossible to guarantee, because passing the convergence test is a necessary but not sufficient condition for thorough sampling. It is possible that the pathway to the minimum score corresponding to the native state is narrow and thus difficult to find. Moreover, randomized sampling is not efficient because it initially iterates through many poor-scoring configurations.

In contrast, deterministic sampling methods focus on sampling relevant configurations and avoiding configurations that clearly violate a subset of the data. For example, given the EM density map, it is possible in some cases to divide it into segmented regions, find good-scoring assignments between subunits and the segmented regions, and refine the assemblies locally (81). In the absence of the density map, it is possible to define pairwise configurations between the subunits using docking approaches. The assemblies of the subunits are than enumerated from the docking results using combinatorial enumeration strategies (82) or genetic algorithms. This approach works well for distance-based information (cross-links, covariation, and FRET), as distance restraints can be easily considered by docking algorithms, reducing the number of possible pairwise and, in turn, multi-subunit configurations. In the case of symmetric assemblies, it is possible to restrict the transformation space to symmetric configurations only (83, 84).

2.4. Analysis of Models and Information

Input information and an ensemble of output models are analyzed together to estimate ensemble precision. Based on the number of good-scoring models and the consistency between the models and information, three outcomes of modeling are possible. In the first outcome, only a single model (or a cluster of similar models) satisfies all input information. Therefore, sufficient information probably exists for determining the correct structure (with the precision corresponding to the variability within the cluster). In the second outcome, two or more different models are consistent with the information. In this case, the input information is insufficient for structure determination or there are multiple significantly populated states. Additional experiments can provide the information to narrow down the possible solutions. In the third outcome, no model satisfies all the input information. There can be one or more reasons for this result: Some input information or its implementation in the scoring function is incorrect, the representation needs to include additional degrees of freedom, and/or sampling needs to be improved (regardless of the outcome of the convergence test above).

The input information often describes a heterogeneous sample (**Table 1**), indicating multiple structural states. In this case, the scoring function and the sampling should take the heterogeneity into account and generate multistate models. In addition, the number of states needs to be determined. Frequently, the principle of Occam's razor suggests using the smallest number of states sufficient to explain the input information without data overfitting. An example of this approach is the minimal ensemble method in molecular modeling based on SAXS or NMR data (40).

However, sometimes Occam's razor is not applicable. For example, even though a SAXS profile of an intrinsically disordered protein may be matched by a sum of profiles for the minimal ensemble structures, the system is likely to exist in a large ensemble of widely different states (85). Such cases are indicated by similarity between distributions of structural properties, such as the radius of gyration, of the best-scoring multistate models.

2.5. Available Software

Supplemental Material >

Software packages are available for various steps of integrative modeling (**Supplemental Tables 1–3**). It is common in an ISM project to employ several of these programs in succession. Alternatively, some of the packages cover the entire ISM process. Most notably, the Integrative Modeling Package (IMP) (10) was designed specifically for ISM and supports many types of information, various representations, and scoring functions. HADDOCK (86) and PatchDock (26, 87) are also frequently used in cases for which significant coverage of all-atom structures are available for the protein subunits. We further refer interested readers to the next section, in which we review a range of published ISM results. These specific examples may instruct researchers on the implementation options for various modeling needs.

3. REVIEW OF PUBLISHED INTEGRATIVE MODELS

We now describe in more detail a selection of published integrative models. It is not an exhaustive list but rather a limited compilation aiming to demonstrate the range of structural questions to which ISM is applicable. In all the cases that we describe, a higher-resolution structure (from crystallography, cryo-EM, or NMR) of the same system was published after the release of the integrative model. Comparison of the model and structure allows assessment of the overall strengths and weaknesses of ISM in a critical manner. Many of the higher-resolution structures were obtained by cryo-EM, reflecting the great technical advance that has occurred in that field over the past decade. We divide the cases reviewed here according to the level of detail in their representation, starting from all-atom models and ending with models that describe only the architecture of the system. A tabulated summary of these comparisons is also available in **Supplemental Table 4.**

3.1. All-Atom Models

All-atom representation is possible when reliable structures are available for most of the system components. These subunit models are obtained by X-ray crystallography, NMR spectroscopy, cryo-EM for larger individual components, comparative models (11, 88), or ab initio structure predictions (89, 90). Assembly of the all-atom components into the full complex is then guided by integration of information from other structural sources. Technically, this assembly is done by established protein–protein docking applications that are augmented by the external data (26, 91–93). In most cases, the information is converted into restraints that are incorporated into the sampling algorithms and scoring functions of these applications. The advantage of atomic representation is that shape and physico-chemical complementarity between the subunits can be considered, resulting in a more accurate scoring function.

3.1.1. Phosphoinositide 3-kinase heterodimer. Phosphoinositide 3-kinases (PI3Ks) play a key role in a variety of cellular processes and are mutated in many cancer types. The class IA PI3Kα consists of the p110α catalytic subunit and the p85α regulatory subunit with oncogenic mutations identified in both subunits. The PI3Kα model was constructed using the X-ray structures of the

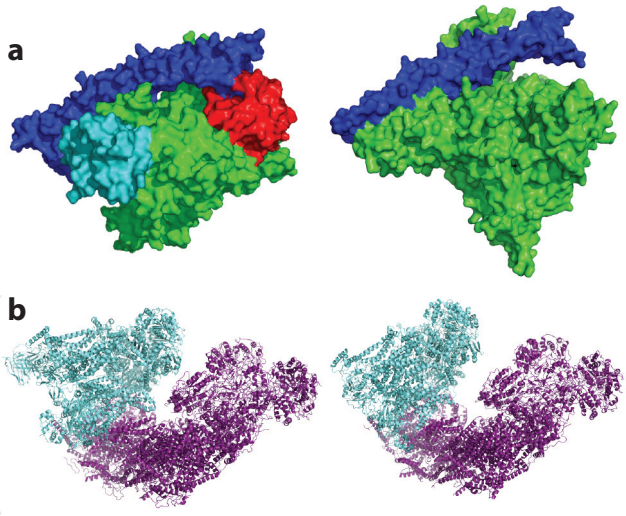

Figure 2

Examples of all-atom integrative models. (*a*) Integrative model (*left*) (94) and crystal structure at 3.0-Å resolution (*right*) (95) of the complex between the p110α catalytic subunit (*green*) and the complex of the p110α ABD domain with the p85α iSH2 domain (*blue*). The p85α nSH2 domain (*red*) and the p85α cSH2 domain (*cyan*) are only present in the integrative model. (*b*) Integrative model (*left*) (34) and the cryo–electron microscopy structure at 7.8-Å resolution (*right*) (96) of the mitochondrial respiratory supercomplexes I (*purple*) and III (*light blue*).

four subunits: p110α (without the ABD domain), the complex of the p110α ABD domain and the p85α iSH2 domain, and the p85α nSH2 and cSH2 domains. PatchDock and CombDock with sequence connectivity constraints between the domains were used to generate a set of models consistent with the constraints (94). These docking calculations pointed to just a single cluster of best-scoring models comprising all of the subunits, except for the p85α cSH2 domain, which could not be placed reliably (**Figure 2*a***). The resulting model shed light on the mechanism of two classes of cancer mutations in the interface between the catalytic and regulatory subunits. The structure of the complex was later solved by X-ray crystallography without the nSH2 and cSH2 domains (95), confirming the integrative model.

3.1.2. Mitochondrial respiratory complexes I and III. The respirasome is a mitochondrial supercomplex consisting of oxidative phosphorylation complexes I, III, and IV. The assembly of complexes I and III was modeled by a protein–protein docking method (PatchDock) using shape complementarity, distance restraints from three cross-links, and the confinement of the complexes to the plane of the membrane (34). The resulting supercomplex is in excellent agreement with a cryo-EM structure of the same system (96), with less than 2.5 Å deviation between them (**Figure 2*b***). Of note, the distance restraints between the two subunits originated from cross-linking of intact mitochondria. The integrative model of the supercomplex thus strengthened the relevance of the cryo-EM structure to the in vivo state.

3.2. Combinatorial Modeling: A Special Case of All-Atom Modeling

For certain biological systems, the native structure is assumed to be one of a finite set of configurations that can be enumerated exhaustively. A fitness score is then assigned to each of these models

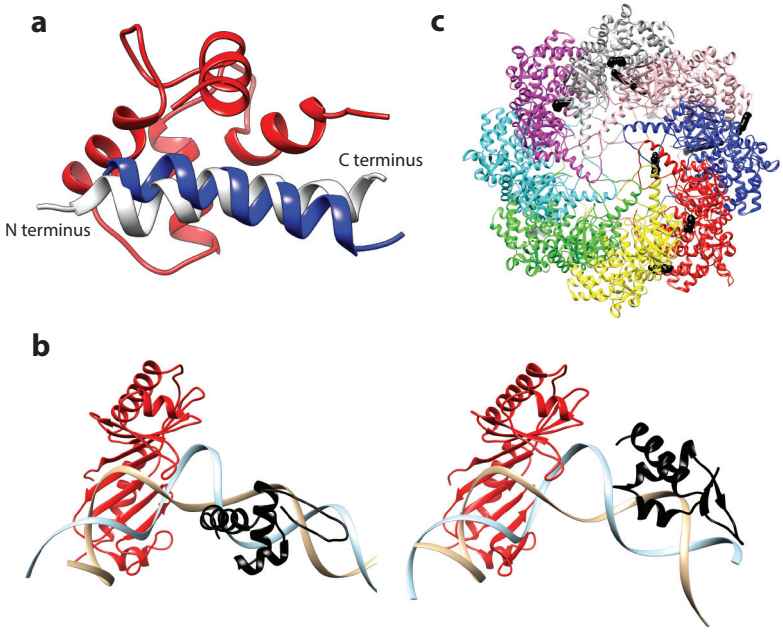

Figure 3

Examples of combinatorial modeling cases. (*a*) The model of a helix from Munc13 (*blue*) bound to calmodulin (*red*) (97) is superimposed on the corresponding NMR solution structure (*light gray*) (98). This orientation of the helix fits the data better than the opposite orientation. (*b*) Integrative model (*left*) (99) and cryo–electron microscopy structure at 4.7-Å resolution (*right*) (100) of the winged-helix domain of Tfg2 (*black*) on the DNA around the TATA-binding protein (*red*) in the transcription pre-initiation complex. Other subunits are not shown for clarity. The model had the best fit to the data out of 35 positions along the DNA path. (*c*) Top view of a model of the CCT chaperonin with each subunit colored by type. The back side of the complex is not shown for clarity. There are 40,320 possible arrangements of the subunits within the complex. The native arrangement was singled out by combinatorial homology modeling and cross-linking coupled to mass spectrometry (71, 101). Some of the inter-subunit cross-links are modeled to scale (*black van der Waals spheres*). Note the small size of the cross-links compared with the overall architecture of the particle.

on the basis of its compatibility with the available structural data. The best-fitting model is thus clearly and objectively singled out. We describe three examples of such systems (**Figure 3**).

3.2.1. Munc13 peptide–calmodulin complex. The correct orientation of the Munc13 helical peptide in the calmodulin cavity was selected from two options (97). At the time of modeling, the structures of all the components were available, but their relative orientations were not known. Dimova et al. (97) modeled both the parallel and antiparallel orientations and compared them with distance restraints from cross-links identified between calmodulin and the helix (**Figure 3*a***). They found that only the antiparallel model satisfied the restraints, and indeed, their model was validated a year later by an independent NMR study (98).

3.2.2. Tfg2 winged-helix domain and DNA–transcription pre-initiation complex. A more elaborate case involved ISM of the winged-helix domain of Tfg2 on DNA in the context of the large transcription pre-initiation complex (99) (**Figure 3*b***). The domain was modeled in 35 different positions along the DNA path. Only one model was compatible with the cross-linking data to neighboring subunits and the steric hindrance by the nearby TATA binding protein. A later

cryo-EM study at much higher resolution (100) confirmed that the location of this domain was accurate.

3.2.3. CCT chaperonin. Finally, an outstanding example of combinatorial modeling is the eukaryotic CCT chaperonin. This conserved complex comprises eight different subunits that are highly similar to one another in structure and sequence. Consequently, the subunit order within the complex could not be resolved, even though the overall shape of the particle was solved at low resolution. To determine the subunit order, all of the 40,320 (= 8 factorial) possible arrangements of the complex were modeled and then compared with distance restraints from XL-MS experiments (71, 101). This approach clearly demonstrated that only one model satisfied these restraints (**Figure 3*c***). The model was later verified conclusively by an independent crystallographic study (102). It should be noted that, at the time of the modeling, several other subunit arrangements were hypothesized to be correct, and all turned out to be significantly different from the one found by ISM. This result emphasizes the potential of ISM to yield nontrivial structural insights that are unaffected by prior bias.

Aside from the assurance that the search space is completely sampled, combinatorial modeling, if applicable, is favorable for two other reasons. First, it quantifies how much better the best-fitting model is compared with the structural alternatives, thus giving a confidence estimate to the selection. Second, in most cases, the models are built independently of the data used for selection. This approach greatly reduces the risk of over-interpreting the data. Clearly, combinatorial modeling is possible only when abundant structural information is available to restrict the system. In the cases reviewed here, all-atom structures were available for all the subunits, and additional information was available for their global architectures. This is usually not the case in ISM, but one should be on the lookout for specific subsystems or structural questions for which combinatorial modeling may apply.

3.3. Mixed-Representation Models

As their name implies, these models combine all-atom structures for specific parts of the assembly and a coarser representation for the rest of it. Very often, such cases involve a low-resolution cryo-EM density map and structural coverage for several of the larger subunits. ISM then proceeds in two steps. In the first step, the all-atom components of the system are docked into the density map by standard correlation (74, 75, 103, 104). These then serve as anchors for the second step, in which the remaining subunits are modeled around them by incorporating local structural information such as XL-MS. Methods for this type of modeling in published studies range from manual modeling (105, 106), through dedicated applications (99), to general platforms such as IMP (107).

3.3.1. Polycomb repressive complex 2. **Figure 4*a*** shows a mixed-representation integrative model of the human polycomb repressive complex 2 (PRC2) that was obtained with a cryo-EM map at 21-Å resolution (105). Despite the low resolution of the map, the shape was sufficiently defined to dock two of the subunits (RPB4 and EED). Other subunits were then placed around them to satisfy 39 distance restraints from XL-MS. At the time, all-atom models were not available for 55% of the protein mass. Therefore, several subunits were not modeled explicitly. Instead, their general locations in the map, as suggested by the cross-links to neighboring components, were marked by coloring of the electron density envelope. We commend this careful approach, which seems highly appropriate given the amount of available structural data. The final integrative model agrees well with the recently published cryo-EM structure of the same system at 3.9-Å

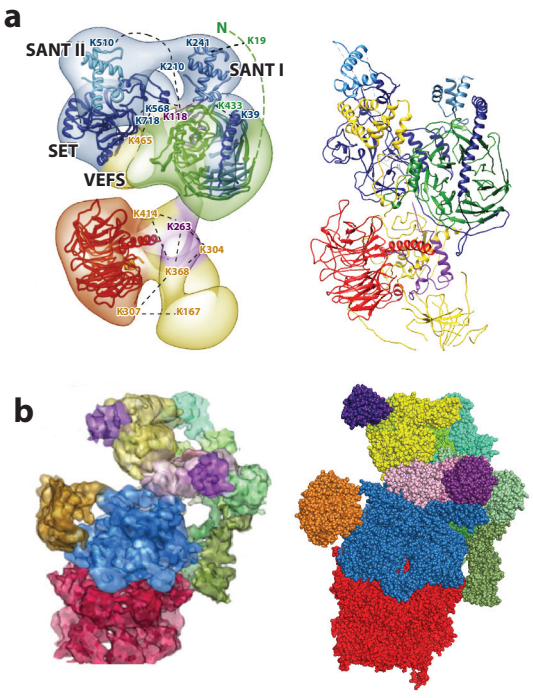

Figure 4

Examples of mixed-representation models, which combine all-atom and coarse-grained components. The integrative models are shown on the left, and higher-resolution cryo–electron microscopy (cryo-EM) structures are shown on the right. The subunit colors match between model and structure. (*a*) The model (105) and cryo-EM structure at 3.9-Å resolution (108) of the human polycomb repressive complex 2 comprising five subunits. Some of the cross-links identified between the subunits are marked on the model (EZH2 and its domains SANT I, SANT II, and SET, *shades of blue*; EED, *green*; RPB4, *red*; SUZ12 and its VEFS domain, *yellow*; AEBP2, *purple*). (*b*) The model (107) and cryo-EM structure at 4.8-Å resolution (110) of the yeast proteasome 19S regulatory particle comprising 19 subunits (RPN1, *orange*; RPN2, *yellow*; RPN8/RPN11 heterodimer, *pink*; RPN10, *purple*; RPN13, *dark purple*, *upper left*; PCI subunits, *shades of green*; AAA-ATPase heterohexamer, *blue*; core particle, *red*).

resolution (108). The locations of all the subunits in the integrative model are accurate, including the locations of subunits that were not modeled explicitly. However, the integrative model deviates from the cryo-EM structure in the rotations of several subunits, in some cases by more than 90°.

3.3.2. The 19S proteasome regulatory particle. **Figure 4*b*** shows an integrative model of the 19S regulatory particle of the yeast proteasome that was obtained using a cryo-EM map of ∼12-Å resolution (107). In the first step, half of the subunits with known atomic structures were docked into the cryo-EM map. The docking mainly resolved the region at the lower half of the map and was assisted by additional information from deletion studies. All-atom models were not available for 12 of the subunits that made up 40% of the protein mass. These subunits were modeled as spheres, each representing 50 residues. In the second step, the IMP package was used to model the remaining subunits, guided by local information from 57 cross-links and 47 pairwise interactions (107). The resulting integrative model was in excellent agreement with a concurrent study (109) and a later higher-resolution structure (110), both by cryo-EM. Of note are the exceptionally

accurate orientations of the subunits in the integrative model, due in part to the relatively high resolution of the map.

3.4. Architectural Models

The available structural data are often limited compared with the many degrees of freedom within the system. These cases involve a combination of a large-assembly, low-resolution electron density map, low structural coverage of the sequences with all-atom models, and a low ratio of cross-links per residue. Under such circumstances, integrative modeling is possible only with coarse representations outlining the probable location of each subunit. Accordingly, ISM can only reveal the architecture rather than the structure of the system. Currently, only the IMP package (10) is capable of handling this level of abstraction, as demonstrated in the following examples.

3.4.1. Transcription factor II H complex. An architectural integrative model of the transcription factor II H (TFIIH) complex was determined using all-atom model coverage for half of the protein mass, a cryo-EM density map at 30-Å resolution, and approximately one cross-link per 15 residues (111). A simplified representation with spheres representing 30 residues each was used in the modeling. In accord with the sparse data, the best model eventually reported only the relative subunit localizations as they extend along an arc-like architecture (**Figure 5*a***). Comparison of the integrative model to the cryo-EM structure at 4.4-Å resolution (112) shows good agreement with the subunit order along the arc. The domain positions and orientations of the two largest subunits (Rad3 and Ssl2) are also modeled well. However, the orientations of the Tfb4 subunit and the Ssl1\Tfb1\Tfb2 cluster differ by nearly 120° between the model and the cryo-EM structure. Yet, one should note that the cryo-EM structure was part of a much larger complex (the pre-initiation complex), whereas the data used for modeling came from TFIIH alone. Therefore, these discrepancies may arise partly from conformational changes between these two states rather than modeling inaccuracies.

3.4.2. Nup84 complex. The seven-subunit coat nucleoporin Nup84 complex from yeast was modeled with cross-link data (one cross-link per 19 residues) and a low-resolution EM 2D class average (27). Structures and comparative models were available for 80% of the protein mass, which were represented by one sphere per residue. The remaining mass was represented by one sphere per 20 residues (**Figure 5*b***). Comparison of the integrative model with a low-resolution crystal structure (113) shows good agreement of the subunit order along the arms of the Y-like shape. In fact, it is plausible that any disagreement between the model and the structure is the result of multiple conformations of this flexible complex. The integrative model is also more complete than the crystal structure, for which one subunit (NUP133) was not resolved.

3.4.3. Nuclear pore complex. One of the first and largest integrative models was reported in 2007 by Alber et al. (22) of the nuclear pore complex (**Figure 5*c***). The model comprises 57 domains that recur in an eight-fold symmetry for a total of 456 components per complex. This model was globally confined by a toroidal density of low resolution (120 Å). Additional inputs to the model included localization in the density for some of the subunits based on immuno-labeling EM and approximately 200 pairwise interactions between subunits (1 interaction per 300 residues). At the time, the percentage of protein mass for which all-atom structures were available was low, and XL-MS was only emerging. Accordingly, modeling relied entirely on representation of domains by spherical beads. Over the past decade, the structural information available for the nuclear pore complex has drastically expanded, and ISM was used to determine the architecture of the nuclear

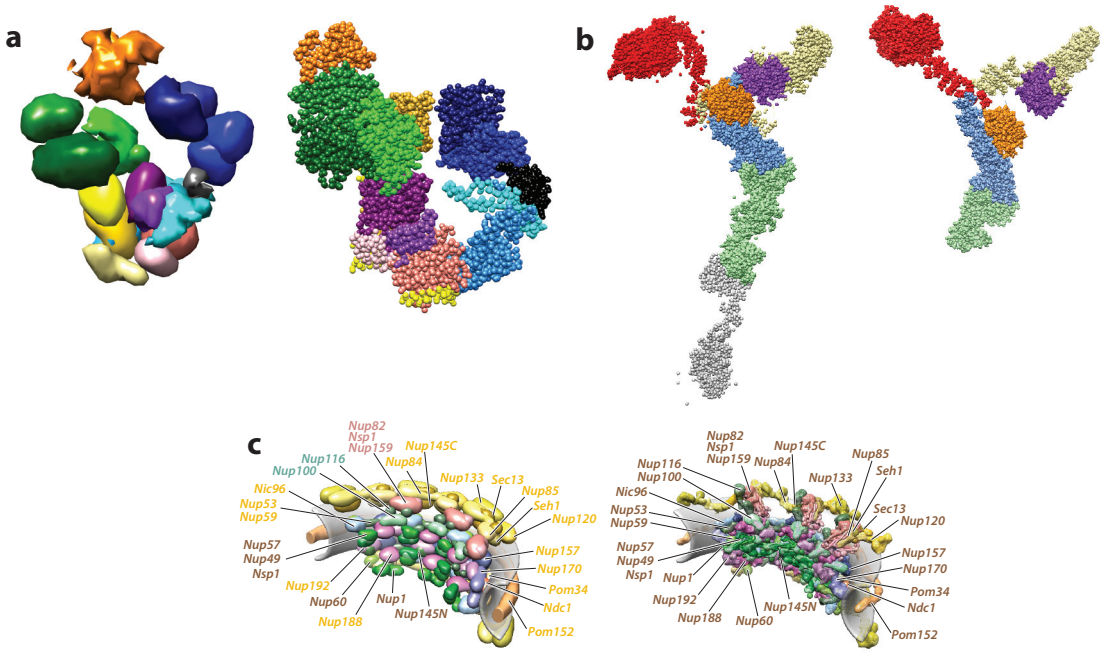

Figure 5

Examples of architectural integrative models. The integrative models are shown on the left, and higher-resolution structures are shown on the right. The subunit colors match between the models and structures. (*a*) Integrative model (111) and cryo–electron microscopy (cryo-EM) structure at 4.7-Å resolution (100) of the yeast transcription factor II H complex comprising eight subunits (Ssl2 domains, *shades of dark blue*; Rad3 domains, *shades of green*; Tfb1 domains, *shades of yellow*; Tfb2 domains, *cyan*, *light blue*; Ssl1 domains, *shades of purple*; Tfb3, *orange*; Tfb4 domains, *pink*, *salmon*; Tfb5, *black*). (*b*) Integrative model (27) and crystal structure at 7.4-Å resolution (113) of the yeast coat nucleoporin complex comprising seven subunits (EC13, *orange*; NUP145, *blue*; SEH1, *purple*; NUP85, *light yellow*; NUP120, *red*; NUP84, *green*; NUP133, *gray*). The crystallographic work did not include the flexible NUP133. (*c*) Sections of the topological map (22) and the recent integrative model with higher resolution (116) of the nuclear pore complex, comprising 456 subunits. The two models agree well with each other, though the more recent one is defined at an order-of-magnitude higher precision.

pore in increasing detail (114, 115). Although an all-atom structure of the nuclear pore is still a goal for the future, an integrative model of significantly higher resolution was recently determined (116) for the same system that was modeled in 2007. The new model relied on an electron density map at 28-Å resolution, 50% coverage of the protein mass by all-atom structures, and nearly 3,100 cross-links (1 cross-link per 18 residues). Comparison of the 2007 and 2018 models reveals how strikingly accurate the former is. Only three subunits in the 2007 model are in a substantially different location compared with the newer model. Even for these subunits, the localization to the nuclear side of the pore is still correct. The main differences between the models are again in the rotations of several subunits, which in some cases are substantial (>90°).

3.5. Summary of Integrative Modeling Accuracy

The cases reviewed here give a general estimate for the accuracy of integrative modeling. The *positions* of subunit centers within the complex are determined to better than 5 Å in all-atom models, 5–10 Å in mixed-representation models, and 10–20 Å in architectural models. These are surprisingly good figures, which mark ISM as an excellent tool for the study of large protein architectures. At the same time, the determination of relative *rotations* for certain subunits within the complex is

less accurate and can deviate by more than 45°. Although part of these deviations can be explained by the flexibility of the complexes, this is nonetheless a weaker point of ISM. As relative rotations are mainly determined by local structural modalities (such as cross-links), devising better ways to use them in ISM should be emphasized.

An almost trivial observation in ISM is that the more data are available, the higher the accuracy of the model. This is especially true for subunit rotations, which were substantially less accurate in regions where the input information was sparse. It is hard to say exactly how much data are required for successful integrative modeling. Yet, on the basis of the cases reviewed here and on our own experience, we estimate that accurate models are likely if the following inputs are available: (*a*) density map with resolution better than 20 Å, (*b*) all-atom models covering more than 70% of the protein mass, and (*c*) one distance constraint per 20 residues, or better. In any case, users must be aware of the main pitfall of ISM—over-interpreting the data. We note that, in this regard, the authors of the ISM studies reviewed here were conservative in their interpretations.

3.6. Functional Insights from Integrative Models

What can architecture tell us about the function of a complex? Globally, the integrative model is informative about the subunit localization within the complex, even at coarse representation. This localization immediately sorts the subunits into core and periphery, thus identifying the central structural subunits. The measurement of distance between any two subunits is also immediately apparent from such models. Finally, if the functions of individual subunits are known, the subunit localization may reveal the partitioning of a complex into defined submodules with specific functions. Locally, the integrative model can identify new interactions between subunits. Such interactions can then be the focus of follow-up studies using other techniques.

Examples of functional questions that were answered by ISM abound. A few are listed here:

1. A mechanism for DNA opening in the transcription pre-initiation complex was suggested by the location of the Ssl2 helicase in an integrative model (99).
2. ISM revealed the partitioning of the CCT chaperonin into ATP-hydrolyzing and ATP-inert sectors (102).
3. The protein SUZ12 was identified at the core of the PRC2 complex (105).
4. The internal symmetries of the nuclear pore complex were calculated (22).
5. The distance between ubiquitin receptors in the integrative model of the proteasome suggested an explanation for polyubiquitination requirement in degradation (107).
6. Disease-causing mutations were mapped to new subunit interfaces identified in the TFIIH complex (111).

As in ISM itself, the risk of overinterpretation is also present in regard to functional insights from integrative models. ISM users must therefore be careful to limit their inferences to the accuracy and precision of their models.

3.7. Future Directions for Integrative Modeling

Many biological systems for which ISM was believed to be the best methodological option for structural characterization were solved by cryo-EM in the last five years. We therefore ask: What is the role of ISM in the new era of high-resolution cryo-EM? To answer this, we note that only 50% of the recently published EM maps had resolutions lower than 6 Å, and the majority of the maps were published without an accompanying PDB model (21). Moreover, it is common for maps of a certain resolution to have large parts that are much lower in resolution or that are even

unresolved. Clearly, these are cases for which ISM is the approach of choice. In addition, ISM can be applied to systems that, at least in the near future, are still not well resolved by cryo-EM, such as large membrane complexes, filament assemblies, very small complexes, and highly flexible complexes.

ISM currently faces several challenges. In our opinion, the most pressing is to improve the performance and ease-of-use of available software, especially for cases of architectural modeling. Often they require running many scripts in succession, with occasional obscure error messages that are hard to interpret even for expert users. A second challenge is to formalize a standard of reporting for ISM results. Efforts in this direction are ongoing (117, 118). Another challenge is modeling the dynamics of complexes, which requires ensembles of conformations. These are all doable tasks, the completion of which will greatly expand the use of ISM in structural biology.

Finally, we highlight two prospective directions in ISM that we believe will be influential. The first is the use of in situ and in vivo data as inputs to the ISM process. Such inputs include cryo-tomograms of cellular sections, in vivo XL-MS (33, 34), and super-resolution light microscopy. Although these data are currently inferior in the resolution and number of restraints compared with what is achieved with purified complexes, they nonetheless hold the potential to yield models that describe the native state better. A second direction is the use of high-yield XL-MS methodology for the structure determination of small globular proteins (72, 119). Although the computational aspects of this approach still need substantial work, it holds the potential to become an NMR-like structural method.

In this review, we have undertaken a critical assessment of the ISM field by comparison of published models with more recent structures at higher resolution. We believe this comparison gives a favorable view of ISM and demonstrates its ability to yield nontrivial structural insights. We hope that these successes will motivate an even wider use of ISM and development of the software infrastructure to support it.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors acknowledge support from the Israel Science Foundation (grants 1466/18 to D.S.-D. and 1768/15 to N.K.).

LITERATURE CITED

1. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–42
2. Kühlbrandt W. 2014. The resolution revolution. *Science* 343:1443–44
3. Rappsilber J. 2011. The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *J. Struct. Biol.* 173:530–40
4. Kochert BA, Iacob RE, Wales TE, Makriyannis A, Engen JR. 2018. Hydrogen-deuterium exchange mass spectrometry to study protein complexes. *Methods Mol. Biol.* 1764:153–71
5. Petoukhov MV, Svergun DI. 2013. Applications of small-angle X-ray scattering to biomacromolecular solutions. *Int. J. Biochem. Cell Biol.* 45:429–37

6. Jares-Erijman EA, Jovin TM. 2003. FRET imaging. *Nat. Biotechnol.* 21:1387–95
7. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, et al. 2011. Protein 3D structure computed from evolutionary sequence variation. *PLOS ONE* 6:e28766
8. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, et al. 2007. Determining the architectures of macromolecular assemblies. *Nature* 450:683–94
9. Robinson CV, Sali A, Baumeister W. 2007. The molecular sociology of the cell. *Nature* 450:973–82
10. Russel D, Lasker K, Webb B, Velazquez-Muriel J, Tjioe E, et al. 2012. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLOS Biol.* 10:e1001244
11. Sali A, Blundell TL. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234:779–815
12. Dunbrack RL Jr. 2006. Sequence comparison and protein structure prediction. *Curr. Opin. Struct. Biol.* 16:374–84
13. Bienert S, Waterhouse A, de Beer TA, Tauriello G, Studer G, et al. 2017. The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Res.* 45:D313–19
14. Khafizov K, Madrid-Aliste C, Almo SC, Fiser A. 2014. Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. *PNAS* 111:3733–38
15. Ovchinnikov S, Park H, Varghese N, Huang P-S, Pavlopoulos GA, et al. 2017. Protein structure determination using metagenome sequence data. *Science* 355:294–98
16. Fernandez-Leiro R, Scheres SH. 2016. Unravelling biological macromolecules with cryo-electron microscopy. *Nature* 537:339–46
17. Lawson CL, Patwardhan A, Baker ML, Hryc C, Garcia ES, et al. 2016. EMDataBank unified data resource for 3DEM. *Nucleic Acids Res.* 44:D396–403
18. Scheres SH. 2012. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* 180:519–30
19. Punjani A, Rubinstein JL, Fleet DJ, Brubaker MA. 2017. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* 14:290–96
20. Li X, Mooney P, Zheng S, Booth CR, Braunfeld MB, et al. 2013. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods* 10:584–90
21. Patwardhan A. 2017. Trends in the Electron Microscopy Data Bank (EMDB). *Acta Crystallogr. Sect. D* 73:503–8
22. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, et al. 2007. The molecular architecture of the nuclear pore complex. *Nature* 450:695–701
23. Patla I, Volberg T, Elad N, Hirschfeld-Warneken V, Grashoff C, et al. 2010. Dissecting the molecular architecture of integrin adhesion sites by cryo-electron tomography. *Nat. Cell Biol.* 12:909–15
24. Wu S, Avila-Sakar A, Kim J, Booth DS, Greenberg CH, et al. 2012. Fabs enable single particle cryoEM studies of small proteins. *Structure* 20:582–92
25. Velázquez-Muriel J, Lasker K, Russel D, Phillips J, Webb BM, et al. 2012. Assembly of macromolecular complexes by satisfaction of spatial restraints from electron microscopy images. *PNAS* 109:18821–26
26. Schneidman-Duhovny D, Rossi A, Avila-Sakar A, Kim SJ, Velazquez-Muriel J, et al. 2012. A method for integrative structure determination of protein-protein complexes. *Bioinformatics* 28:3282–89
27. Shi Y, Fernandez-Martinez J, Tjioe E, Pellarin R, Kim SJ, et al. 2014. Structural characterization by cross-linking reveals the detailed architecture of a coatomer-related heptameric module from the nuclear pore complex. *Mol. Cell. Proteom.* 13:2927–43
28. Sinz A. 2014. The advancement of chemical cross-linking and mass spectrometry for structural proteomics: from single proteins to protein interaction networks. *Expert Rev. Proteom.* 11:733–43
29. Leitner A, Faini M, Stengel F, Aebersold R. 2016. Crosslinking and mass spectrometry: an integrated technology to understand the structure and function of molecular machines. *Trends Biochem. Sci.* 41:20–32
30. Leitner A, Joachimiak LA, Unverdorben P, Walzthoeni T, Frydman J, et al. 2014. Chemical cross-linking/mass spectrometry targeting acidic residues in proteins and protein complexes. *PNAS* 111:9455–60

31. Greber BJ, Bieri P, Leibundgut M, Leitner A, Aebersold R, et al. 2015. The complete structure of the 55*S* mammalian mitochondrial ribosome. *Science* 348:303–8
32. Robinson PJ, Trnka MJ, Bushnell DA, Davis RE, Mattei PJ, et al. 2016. Structure of a complete Mediator-RNA polymerase II pre-initiation complex. *Cell* 166:1411–22.e16
33. Wu X, Chavez JD, Schweppe DK, Zheng C, Weisbrod CR, et al. 2016. In vivo protein interaction network analysis reveals porin-localized antibiotic inactivation in *Acinetobacter baumannii* strain AB5075. *Nat. Commun.* 7:13414
34. Schweppe DK, Chavez JD, Lee CF, Caudal A, Kruse SE, et al. 2017. Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry. *PNAS* 114:1732–37
35. Debelyy MO, Waridel P, Quadroni M, Schneiter R, Conzelmann A. 2017. Chemical crosslinking and mass spectrometry to elucidate the topology of integral membrane proteins. *PLOS ONE* 12:e0186840
36. Molnar KS, Bonomi M, Pellarin R, Clinthorne GD, Gonzalez G, et al. 2014. Cys-scanning disulfide crosslinking and Bayesian modeling probe the transmembrane signaling mechanism of the histidine kinase, PhoQ. *Structure* 22:1239–51
37. Street TO, Zeng X, Pellarin R, Bonomi M, Sali A, et al. 2014. Elucidating the mechanism of substrate recognition by the bacterial Hsp90 molecular chaperone. *J. Mol. Biol.* 426:2393–404
38. Walzthoeni T, Joachimiak LA, Rosenberger G, Rost HL, Malmstrom L, et al. 2015. xTract: software for characterizing conformational changes of protein complexes by quantitative cross-linking mass spectrometry. *Nat. Methods* 12:1185–90
39. Yu C, Mao H, Novitsky EJ, Tang X, Rychnovsky SD, et al. 2015. Gln40 deamidation blocks structural reconfiguration and activation of SCF ubiquitin ligase complex by Nedd8. *Nat. Commun.* 6:10053
40. Slavin M, Kalisman N. 2018. Structural analysis of protein complexes by cross-linking and mass spectrometry. *Methods Mol. Biol.* 1764:173–83
41. Putnam CD, Hammel M, Hura GL, Tainer JA. 2007. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q. Rev. Biophys.* 40:191–285
42. Hura GL, Menon AL, Hammel M, Rambo RP, Poole FL 2nd, et al. 2009. Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nat. Methods* 6:606–12
43. Tiede DM, Mardis KL, Zuo X. 2009. X-ray scattering combined with coordinate-based analyses for applications in natural and artificial photosynthesis. *Photosynth. Res.* 102:267–79
44. Whitten AE, Trewhella J. 2009. Small-angle scattering and neutron contrast variation for studying biomolecular complexes. *Methods Mol. Biol.* 544:307–23
45. Rambo RP, Tainer JA. 2013. Accurate assessment of mass, models and resolution by small-angle scattering. *Nature* 496:477–81
46. Rambo RP, Tainer JA. 2013. Super-resolution in solution X-ray scattering and its applications to structural systems biology. *Annu. Rev. Biophys.* 42:415–41
47. Trewhella J. 2016. Small-angle scattering and 3D structure interpretation. *Curr. Opin. Struct. Biol.* 40:1–7
48. Bernado P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI. 2007. Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.* 129:5656–64
49. Rambo RP, Tainer JA. 2011. Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. *Biopolymers* 95:559–71
50. Hura GL, Budworth H, Dyer KN, Rambo RP, Hammel M, et al. 2013. Comprehensive macromolecular conformations mapped by quantitative SAXS analyses. *Nat. Methods* 10:453–54
51. Pelikan M, Hura GL, Hammel M. 2009. Structure and flexibility within proteins as identified through small angle X-ray scattering. *Gen. Physiol. Biophys.* 28:174–89
52. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A. 2016. FoXS, FoXSDock and MultiFoXS: single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res.* 44:W424–29
53. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. 2013. The next-generation sequencing revolution and its impact on genomics. *Cell* 155:27–38
54. Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, et al. 2003. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics* 19:163–64

55. Marks DS, Hopf TA, Sander C. 2012. Protein structure prediction from sequence variation. *Nat. Biotechnol.* 30:1072–80
56. Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. 2012. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 149:1607–21
57. Wang S, Sun S, Li Z, Zhang R, Xu J. 2017. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLOS Comput. Biol.* 13:e1005324
58. Hopf TA, Scharfe CP, Rodrigues JP, Green AG, Kohlbacher O, et al. 2014. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife* 3:e03430
59. Hopf TA, Marks DS. 2017. Protein structures, interactions and function from evolutionary couplings. In *From Protein Structure to Function with Bioinformatics*, ed. DJ Rigden, pp. 37–58. Dordrecht, Neth.: Springer. 2nd ed.
60. Selvin PR. 2000. The renaissance of fluorescence resonance energy transfer. *Nat. Struct. Biol.* 7:730–34
61. Forster T. 1948. Intermolecular energy transfer and fluorescence. *Ann. Phys.* 2:55–75
62. Brunger AT, Strop P, Vrljic M, Chu S, Weninger KR. 2011. Three-dimensional molecular modeling with single molecule FRET. *J. Struct. Biol.* 173:497–505
63. Bonomi M, Pellarin R, Kim SJ, Russel D, Sundin BA, et al. 2014. Determining protein complex structures based on a Bayesian model of in vivo Förster resonance energy transfer (FRET) data. *Mol. Cell. Proteom.* 13:2812–23
64. Lerner E, Cordes T, Ingargiola A, Alhadid Y, Chung S, et al. 2018. Toward dynamic structural biology: Two decades of single-molecule Förster resonance energy transfer. *Science* 359:eaan1133
65. Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, et al. 2003. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* 331:281–99
66. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. 1992. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *PNAS* 89:2195–99
67. Ritchie DW, Kemp GJ. 2000. Protein docking using spherical polar Fourier correlations. *Proteins* 39:178–94
68. Moal IH, Moretti R, Baker D, Fernandez-Recio J. 2013. Scoring functions for protein–protein interactions. *Curr. Opin. Struct. Biol.* 23:862–67
69. Dong GQ, Fan H, Schneidman-Duhovny D, Webb B, Sali A. 2013. Optimized atomic statistical potentials: assessment of protein interfaces and loops. *Bioinformatics* 29:3158–66
70. Lasker K, Phillips JL, Russel D, Velazquez-Muriel J, Schneidman-Duhovny D, et al. 2010. Integrative structure modeling of macromolecular assemblies from proteomics data. *Mol. Cell. Proteom.* 9:1689–702
71. Kalisman N, Adams CM, Levitt M. 2012. Subunit order of eukaryotic TRiC/CCT chaperonin by cross-linking, mass spectrometry, and combinatorial homology modeling. *PNAS* 109:2884–89
72. Belsom A, Schneider M, Fischer L, Brock O, Rappsilber J. 2016. Serum albumin domain structures in human blood serum by mass spectrometry and computational biology. *Mol. Cell. Proteom.* 15:1105–16
73. Rieping W, Habeck M, Nilges M. 2005. Inferential structure determination. *Science* 309:303–6
74. Chacon P, Wriggers W. 2002. Multi-resolution contour-based fitting of macromolecular structures. *J. Mol. Biol.* 317:375–84
75. Garzon JI, Kovacs J, Abagyan R, Chacon P. 2007. ADP_EM: fast exhaustive multi-resolution docking for high-throughput coverage. *Bioinformatics* 23:427–33
76. Svergun D, Barberato C, Koch MHJ. 1995. CRYSOL—a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystallogr.* 28:768–73
77. Schneidman-Duhovny D, Hammel M, Sali A. 2010. FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res.* 38(Suppl):W540–44
78. Schneidman-Duhovny D, Kim SJ, Sali A. 2012. Integrative structural modeling with small angle X-ray scattering profiles. *BMC Struct. Biol.* 12:17
79. Schneidman-Duhovny D, Hammel M, Tainer JA, Sali A. 2013. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* 105:962–74

80. Viswanath S, Chemmama IE, Cimermancic P, Sali A. 2017. Assessing exhaustiveness of stochastic sampling for integrative modeling of macromolecular structures. *Biophys. J.* 113:2344–53
81. Lasker K, Topf M, Sali A, Wolfson HJ. 2009. Inferential optimization for simultaneous fitting of multiple components into a cryoEM map of their assembly. *J. Mol. Biol.* 388:180–94
82. Inbar Y, Benyamini H, Nussinov R, Wolfson HJ. 2005. Combinatorial docking approach for structure prediction of large proteins and multi-molecular assemblies. *Phys. Biol.* 2:S156–65
83. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. 2005. Geometry-based flexible and symmetric protein docking. *Proteins* 60:224–31
84. Andre I, Bradley P, Wang C, Baker D. 2007. Prediction of the structure of symmetrical protein assemblies. *PNAS* 104:17656–61
85. Carter L, Kim SJ, Schneidman-Duhovny D, Stohr J, Poncet-Montange G, et al. 2015. Prion protein—antibody complexes characterized by chromatography-coupled small-angle X-ray scattering. *Biophys. J.* 109:793–805
86. Dominguez C, Boelens R, Bonvin AM. 2003. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* 125:1731–37
87. Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ. 2005. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res.* 33:W363–67
88. Soding J. 2005. Protein homology detection by HMM–HMM comparison. *Bioinformatics* 21:951–60
89. Simons KT, Kooperberg C, Huang E, Baker D. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* 268:209–25
90. Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. *BMC Bioinform.* 9:40
91. Orban-Nemeth Z, Beveridge R, Hollenstein DM, Rampler E, Stranzl T, et al. 2018. Structural prediction of protein models using distance restraints derived from cross-linking mass spectrometry data. *Nat. Protoc.* 13:478–94
92. van Dijk AD, Boelens R, Bonvin AM. 2005. Data-driven docking for the study of biomolecular complexes. *FEBS J.* 272:293–312
93. Lyskov S, Gray JJ. 2008. The RosettaDock server for local protein-protein docking. *Nucleic Acids Res.* 36:W233–38
94. Miled N, Yan Y, Hon WC, Perisic O, Zvelebil M, et al. 2007. Mechanism of two classes of cancer mutations in the phosphoinositide 3-kinase catalytic subunit. *Science* 317:239–42
95. Huang CH, Mandelker D, Schmidt-Kittler O, Samuels Y, Velculescu VE, et al. 2007. The structure of a human p110α/p85α complex elucidates the effects of oncogenic PI3Kα mutations. *Science* 318:1744–48
96. Letts JA, Fiedorczuk K, Sazanov LA. 2016. The architecture of respiratory supercomplexes. *Nature* 537:644–48
97. Dimova K, Kalkhof S, Pottratz I, Ihling C, Rodriguez-Castaneda F, et al. 2009. Structural insights into the calmodulin–Munc13 interaction obtained by cross-linking and mass spectrometry. *Biochemistry* 48:5908–21
98. Rodriguez-Castaneda F, Maestre-Martinez M, Coudevylle N, Dimova K, Junge H, et al. 2010. Modular architecture of Munc13/calmodulin complexes: dual regulation by Ca^{2+} and possible function in short-term synaptic plasticity. *EMBO J.* 29:680–91
99. Murakami K, Elmlund H, Kalisman N, Bushnell DA, Adams CM, et al. 2013. Architecture of an RNA polymerase II transcription pre-initiation complex. *Science* 342:1238724
100. Schilbach S, Hantsche M, Tegunov D, Dienemann C, Wigge C, et al. 2017. Structures of transcription pre-initiation complex with TFIIH and Mediator. *Nature* 551:204–9
101. Leitner A, Joachimiak LA, Bracher A, Monkemeyer L, Walzthoeni T, et al. 2012. The molecular architecture of the eukaryotic chaperonin TRiC/CCT. *Structure* 20:814–25
102. Kalisman N, Schroder GF, Levitt M. 2013. The crystal structures of the eukaryotic chaperonin CCT reveal its functional partitioning. *Structure* 21:540–49
103. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. 2004. UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25:1605–12

104. van Zundert GCP, Melquiond ASJ, Bonvin A. 2015. Integrative modeling of biomolecular complexes: HADDOCKing with cryo-electron microscopy data. *Structure* 23:949–60
105. Ciferri C, Lander GC, Maiolica A, Herzog F, Aebersold R, Nogales E. 2012. Molecular architecture of human polycomb repressive complex 2. *eLife* 1:e00005
106. Murakami K, Tsai KL, Kalisman N, Bushnell DA, Asturias FJ, Kornberg RD. 2015. Structure of an RNA polymerase II preinitiation complex. *PNAS* 112:13543–48
107. Lasker K, Forster F, Bohn S, Walzthoeni T, Villa E, et al. 2012. Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. *PNAS* 109:1380–87
108. Kasinath V, Faini M, Poepsel S, Reif D, Feng XA, et al. 2018. Structures of human PRC2 with its cofactors AEBP2 and JARID2. *Science* 359:940–44
109. Lander GC, Estrin E, Matyskiela ME, Bashore C, Nogales E, Martin A. 2012. Complete subunit architecture of the proteasome regulatory particle. *Nature* 482:186–91
110. Luan B, Huang X, Wu J, Mei Z, Wang Y, et al. 2016. Structure of an endogenous yeast 26S proteasome reveals two major conformational states. *PNAS* 113:2642–47
111. Luo J, Cimermancic P, Viswanath S, Ebmeier CC, Kim B, et al. 2015. Architecture of the human and yeast general transcription and DNA repair factor TFIIH. *Mol. Cell* 59:794–806
112. Greber BJ, Nguyen THD, Fang J, Afonine PV, Adams PD, Nogales E. 2017. The cryo-electron microscopy structure of human transcription factor IIH. *Nature* 549:414–17
113. Stuwe T, Correia AR, Lin DH, Paduch M, Lu VT, et al. 2015. Architecture of the nuclear pore complex coat. *Science* 347:1148–52
114. von Appen A, Kosinski J, Sparks L, Ori A, DiGuilio AL, et al. 2015. In situ structural analysis of the human nuclear pore complex. *Nature* 526:140–43
115. Kosinski J, Mosalaganti S, von Appen A, Teimer R, DiGuilio AL, et al. 2016. Molecular architecture of the inner ring scaffold of the human nuclear pore complex. *Science* 352:363–65
116. Kim SJ, Fernandez-Martinez J, Nudelman I, Shi Y, Zhang W, et al. 2018. Integrative structure and functional anatomy of a nuclear pore complex. *Nature* 555:475–82
117. Sali A, Berman HM, Schwede T, Trewhella J, Kleywegt G, et al. 2015. Outcome of the first wwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure* 23:1156–67
118. Vallat B, Webb B, Westbrook JD, Sali A, Berman HM. 2018. Development of a prototype system for archiving integrative/hybrid structure models of biological macromolecules. *Structure* 26:894–904.e2
119. Schneider M, Belsom A, Rappsilber J. 2018. Protein tertiary structure by crosslinking/mass spectrometry. *Trends Biochem. Sci.* 43:157–69

Annual Review of Biochemistry

Volume 88, 2019

Contents

Errata

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