Published in final edited form as:

Science. 2013 February 22; 339(6122): 913–915. doi:10.1126/science.1228565.

Integrative Structural Biology

Andrew B. Ward, Andrej Sali, and Ian A. Wilson

Biological assemblies and machines often elude structural characterization, hampering our understanding of how they function, how they evolved, and how they can be modulated. A number of macromolecular assemblies have been reconstructed over the years by piecemeal efforts, such as fitting high-resolution crystal structures of individual components into lower resolution electron microscopy (EM) reconstructions of the entire complex (1). Although notable successes have been achieved in this way, ambiguous or conflicting models can still arise (2–4). Thus, structural and computational biologists have been looking for new ways to put all of the pieces back together. Sophisticated integrative approaches are being developed (5, 6) that combine information from different types of experiments, physical theories, and statistical analyses to compute structural models of multicomponent assemblies and complex biological systems.

In addition to the conventional biophysical techniques of x-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, EM, and small-angle x-ray scattering (SAXS), a growing number of experimental methods can also provide valuable information about the structures and dynamics of proteins and their assemblies. These methods include sequence comparisons of related proteins, copurification, hydrogen-deuterium exchange mass spectrometry (HDXMS), single-molecule fluorescence, atomic force microscopy, analytical spectroscopy (both electron paramagnetic resonance and double electron-electron resonance), light scattering, chemical cross-linking, and mutagenesis (see the figure).

The individual pieces of data gathered using different techniques provide restraints on the conformation, position, and orientation of the components in an assembly or biological system (5). Relative to the use of any single set or type of data, simultaneous use of all such restraints can markedly improve the accuracy, precision, and completeness of a model, especially when high-resolution structural data of the entire complex are not available.

Because of the many degrees of freedom in macromolecular structures and the difficulty of combining disparate data, models must be computed with algorithms that sample as many solutions as possible given the computing power available. These algorithms are driven by a scoring function consisting of the individual spatial restraints, and are analogous to methods in use in x-ray crystallography and NMR spectroscopy, which also generate models by minimizing differences between experimentally measured data and data computed from a model. Assessing how to best combine and weigh different types of data from multiple sources is a prerequisite for constructing structural models of increasingly larger and more dynamic macromolecular complexes.

A useful test of a model is whether it explains all data points within their own error bars and whether the entire data set is redundant, meaning that a subset of the data can be omitted without any impact on the model. In such a case, the confidence in the model, the data, and the parameters used for modeling can be high. When a subset of the data points cannot be satisfied by a single model because the data were collected from a heterogeneous sample and/or the data are noisy, more sophisticated methods for combining individual restraints are needed. In such cases, emphasis is placed on evaluating models in an objective manner, using Bayesian (7) and other statistical methods that explicitly take into account the noise in the data and/or multiple structures in the sample.

Integrative, restraint-based approaches can be used whenever a challenging structural biology problem is encountered, from an individual protein to a small macromolecular machine to a large multicomponent cellular assembly. Thus, integrative approaches span wide resolution ranges and bridge observations made from the atomic to the cellular level. The following three examples illustrate the power of these new methods in generating models at different levels of resolution.

Some of the most successful applications of integrative approaches have resulted from combining sparse experimental observations with computation to generate atomic-level models of macromolecules. Rosetta (8), a platform for modeling protein structures, works by exhaustive calculations under a set of assumptions about the underlying geometry and chemistry of peptides. These assumptions reduce the nearly infinite sampling necessary to fold a one-dimensional sequence of amino acids into a defined shape. Experimental restraints from NMR (9) or EM (10) can further narrow the search and help to converge on more accurate models. For example, Loquet *et al.* used solid-state NMR, EM, and Rosetta to build an atomic-level model of the bacterial type III secretion needle used to inject its proteins into host cells (11). The model revealed details of the supramolecular interfaces of the component protomers, providing a structural understanding of this machine that had eluded characterization by single techniques.

Two recent independent studies of the molecular architecture of the 26S proteasome exemplify the value of integrative approaches for medium-resolution structures. Lander *et al.* combined EM reconstructions and x-ray crystal structures (12), whereas Lasker *et al.* used restraints from a variety of data sets (EM, x-ray crystallography, chemical crosslinking, and proteomics) and employed the Integrative Modeling Platform package (6, 13) to build an almost identical model of the 26S proteasome (14). The Lasker, *et al.* model was tested by systematically removing some input data, recalculating a model, and assessing it against the omitted data. Although neither model resolved all interactions at an atomic level, they provided a detailed understanding of the arrangement of the component subunits and were therefore informative about the evolution and function of the 26S proteasome.

At low resolution, chromatin can be modeled through integrative approaches. In this way, Duan *et al.* constructed a three dimensional model of the yeast genome (15), uncovering the topology and spatial relationships of different chromosomal elements. For this study, the restraints were garnered from cross-linking, restriction enzyme digestion, ligation, and deep sequencing, thereby revealing the three-dimensional structure of the genome at a level of detail not accessible to any conventional imaging method typically used to study assemblies of this size. These inferential, cellular-scale approaches enable comparison of normal and aberrant cells and may eventually serve an important diagnostic role in medicine.

As integrative methods evolve, structural models can be revisited and new data incorporated, so that the model is continuously improved and revised using the latest information (13). Integrative software tools should therefore be flexible enough to incorporate new data and/or restraints. The Protein Data Bank (<http://www.pdb.org>) is facilitating this process by acting as a curator for a variety of structural data from different methods as well as models based on these data.

Any experimental observation can in principle be converted into a restraint for building ever more complex models. The reach and impact of structural biology can thus be extended to a wider and more diverse audience. Using these new computational and bioinformatics approaches to collect and integrate diverse pieces of structural and experimental data, Humpty Dumpty can be put back together again.

Acknowledgments

Supported by the Protein Structure Initiative of the National Institute of General Medical Sciences, the CHAVI-ID and HIVRAD (HIV Vaccine Research and Design) programs of the National Institute of Allergy and Infectious Diseases, and IAVI. 10.1126/science.1228565

References and Notes

1. Lander GC, Saibil HR, Nogales E. *Curr Opin Struct Biol.* 2012; 22:627. [PubMed: 22835744]
2. Wang HW, et al. *Nat Struct Mol Biol.* 2009; 16:1148. [PubMed: 19820710]
3. Lau PW, et al. *Nat Struct Mol Biol.* 2012; 19:436. [PubMed: 22426548]
4. Henderson R, et al. *Structure.* 2012; 20:205. [PubMed: 22325770]
5. Alber F, et al. *Annu Rev Biochem.* 2008; 77:443. [PubMed: 18318657]
6. Russel D, et al. *PLoS Biol.* 2012; 10:e1001244. [PubMed: 22272186]
7. Rieping W, et al. *Science.* 2005; 309:303. [PubMed: 16002620]
8. www.rosettacommons.org.
9. Sgourakis NG, et al. *J Am Chem Soc.* 2011; 133:6288. [PubMed: 21466200]
10. DiMaio F, et al. *J Mol Biol.* 2009; 392:181. [PubMed: 19596339]
11. Loquet A, et al. *Nature.* 2012; 486:276. [PubMed: 22699623]
12. Lander GC, et al. *Nature.* 2012; 482:186. [PubMed: 22237024]
13. <http://integrativemodeling.org>.
14. Lasker K, et al. *Proc Natl Acad Sci U S A.* 2012; 109:1380. [PubMed: 22307589]
15. Duan Z, et al. *Nature.* 2010; 465:363. [PubMed: 20436457]

Figure 1.