Published in final edited form as:
Curr Opin Struct Biol. 2014 October ; 0: 96–104. doi:10.1016/j.sbi.2014.08.001.

Uncertainty in Integrative Structural Modeling

Dina Schneidman-Duhovny[a,*], **Riccardo Pellarin**[a], and **Andrej Sali**[a,b,*]

[a]Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA 94158, USA

[b]Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences (QB3), University of California, San Francisco, San Francisco, CA 94158, USA

Abstract

Integrative structural modelling uses multiple types of input information and proceeds in four stages: (i) gathering information, (ii) designing model representation and converting information into a scoring function, (iii) sampling good-scoring models, and (iv) analyzing models and information. In the first stage, uncertainty originates from data that are sparse, noisy, ambiguous, or derived from heterogeneous samples. In the second stage, uncertainty can originate from a representation that is too coarse for the available information or a scoring function that does not accurately capture the information. In the third stage, the major source of uncertainty is insufficient sampling. In the fourth stage, clustering, cross-validation, and other methods are used to estimate the precision and accuracy of the models and information.

Keywords

macromolecular assemblies; integrative modeling; protein structure; accuracy; precision; uncertainty

Introduction

To understand and modulate biological processes, we need their spatiotemporal models. These models can be computed based on input information about the structure and dynamics of the system of interest, including physical theories, statistical inference from databases of known sequences and structures, as well as a large variety of experimental methods. A structural model of a molecule is defined by the relative positions and orientations of its components (*eg*, atoms, pseudo-atoms, residues, secondary structure elements, domains, and subunits). All structural characterization approaches correspond to finding models that best fit input information, as can be judged by a scoring function; when the scoring function

*Corresponding authors: 1700 4th Street, Byers Hall 503B, University of California, San Francisco, San Francisco, CA 94158; tel 415-514-4227; web http://salilab.org; dina@salilab.org and sali@salilab.org.

includes experimental data, it quantifies the difference between the observed data and the data computed from the model. Therefore, structural characterization can be described as a four-stage process: (i) gathering input information, (ii) designing model representation and converting information into a scoring function, (iii) sampling good-scoring models, and (iv) analyzing models and information. For example, in X-ray crystallography a model consists of atomic positions, and the scoring function assesses the agreements (i) between the computed and observed structure factors *via* the R_{free} parameter [1] as well as (ii) between the model geometry and the ideal geometry implied by a molecular mechanics force field *via* the potential energy of the model.

To use a model well, we need to assess its accuracy (stage iv above). Assessment standards and corresponding tools have already been developed for X-ray crystallography [2] and Nuclear Magnetic Resonance (NMR) spectroscopy [3], while they are still evolving for electron microscopy (EM) [4], Small Angle X-ray Scattering [5,6], and comparative modeling [7]. Standard validation of the crystallographic and NMR entries in the Protein Data Bank (PDB) [8] includes assessing geometrical features such as stereochemistry and packing, fit of the model to the experimental data, and the quality of the data itself. In the EM field, Fourier Shell Correlation (FSC) is commonly used to estimate map resolution [4,9,10]. Recently, new validation methods for EM maps were suggested, including tilt pair analysis [11], gold-standard FSC curves [4], high-resolution noise substitution [12,13], and ResLog plots [14•]. In SAXS data validation, the χ-free criterion was recently proposed [15••], inspired by R_{free} in crystallography. Protein aggregation can be revealed in the Guinier plot, inter-particle interference can be detected by measuring SAXS profiles at multiple concentrations, and conformational heterogeneity is to some degree reflected in the Kratky or Porod-Debye plots [16]. Estimating the accuracy of comparative models is still challenging, but methods based on a variety of criteria do exist [7,17,18].

No single experimental method is guaranteed to produce a satisfactory structure for a given system. Nevertheless, structure determination can often benefit from an integrative (hybrid) approach, where information from multiple experimental datasets is used to compute all structural models that are consistent with the available data [19–22]. Data from X-ray crystallography, EM, NMR spectroscopy, SAXS, cross-linking combined with mass spectrometry (MS), Förster resonance energy transfer (FRET) spectroscopy, double electron-electron resonance (DEER), and hydrogen–deuterium exchange (HDX) is frequently used in integrative structure determination (Table 1). Sometimes integrative models are assessed based on clustering of models, modeling with simulated data, observation of non-random patterns in the models, and modeling with subsets of data [20]. However, a set of standards for validating integrative models has not yet been developed [22].

It is essential for appropriate use of a structural model to estimate errors in the model as well as the data used to compute it. Model error is defined as the difference between the model and true structure. It originates from several different sources. First, input data can be sparse, noisy, ambiguous, or incoherent (Glossary). Second, the system representation can be too coarse, resulting in some input information being ignored. Third, the scoring function may not accurately capture the input information or the input information is insufficient to

identify the true structure. Fourth, sampling may not find the true structure due to many degrees of freedom used to represent the system. Because the true structure is unknown in real applications, model error is also unknown. However, the lower bound on the model error can often be estimated as the precision of the set of models consistent with the input information. Here, we describe the origins of uncertainty in each stage of integrative modeling, and suggest how to quantify and minimize it.

Stage 1: Gathering information

Spatial information about a given system can include data from experiments such as those listed above, statistical propensities such as atomic statistical potentials extracted from known protein structures, and physical laws, such as interatomic interactions approximated by a molecular mechanics force field. This information is used to represent the system as well as to sample and rank its possible configurations. There are four sources of uncertainty in the information, as follows.

Data sparseness

The data sparseness measures the amount of information in the data relative to the number of degrees of freedom in the model; the amount of information in the data depends on the number of data points and their precision as well as their interdependence. Data sparseness affects the precision of the model [23]. For example, for a protein-protein complex mapped by a single cross-link, if each protein is represented by a single sphere, the data sparseness is 1 data point per 1 degree of freedom; if the proteins are represented by their rigid atomic structures, the data sparseness is 1 data point per 6 degrees of freedom (3 rotations and 3 translations). In X-ray crystallography, the data sparseness can be quantified by the number of reflections divided by the number of atoms in the unit cell. In NMR spectroscopy, the data sparseness is usually quantified by the number of NOE restraints per residue. In SAXS, the data sparseness of a SAXS profile is defined using the Nyquist-Shannon sampling theorem: given the maximum dimension (d_{max}), the sampling theorem determines that the number of unique, evenly distributed observations for a maximum scattering vector (q_{max}) is given by $(d_{\max} q_{\max})/\pi$. The problem with sparse data is that there are more free parameters than observations, which may lead to an over-interpretation of the data (over-fitting).

Data error

Error of the data is the sum of random and systematic measurement errors. The magnitude of random error can best be assessed by multiple repeated measurements; systematic errors for a given type of data can be estimated by benchmarks relying on known structures. For example, in X-ray crystallography the random error is caused by random variations among the crystals as well as noise in the X-ray flux, detector and electronics, while the systematic error can result from radiation damage, conformational heterogeneity of the sample, and crystal packing defects [24]. In SAXS, the random error sources are similar to those in crystallography, while the systematic error can result from sample aggregation and radiation damage [5].

Data ambiguity

Data ambiguity is the uncertainty in assigning data points to specific components of the system. For example, it is generally not possible to assign which of the three methyl protons gave rise to an observed NOE signal [25]. Another example is the ambiguity of assigning a cross-link to a specific instance of a protein when the complex contains multiple instances of it. In contrast, diffraction and scattering data are a function of all components of a system and thus not ambiguous.

Data incoherence

Data incoherence is a result of compositional or conformational heterogeneity of one or more samples used to generate one or more datasets for modeling; for example, a system may exist as a mixture of two states in an NMR solution experiment or it may exist in different states in X-ray (crystal) and SAXS (solution) experiments. As a result, the measured data will be a mixture of contributions from each state. The ability to disentangle different states depends on the precision and accuracy of the data; for example, conformational differences smaller than the precision of the data may be difficult to detect.

Stages 2 and 3: Converting input information into system representation, scoring function, and sampling

Input information about the structure of the system can be used (i) to select the set of variables that represent the system (system representation), (ii) to rank the different configurations (scoring function), and (iii) to search for good scoring solutions (sampling). It is often most computationally efficient, although not always possible, to encode information into the representation; in contrast it is generally most straightforward, but least efficient, to encode information into the scoring function. For example, in protein-protein docking, maximization of shape complementarity can be encoded into a scoring function that is then optimized by a generic optimization method. Alternatively, maximization of shape complementarity can also be encoded more efficiently through a representation consisting of shape descriptors, such as surface curvature, resulting in faster sampling by generating only configurations of subunits with complementary shape descriptors [26].

Representation

The representation of a system is defined by all the variables that need to be determined based on input information, including the assignment of the system components to geometric objects, such as points and spheres. A simple example is Cartesian coordinates for points corresponding to the individual atoms. More complex representations can assign a component to other geometric primitives (*eg*, spheres, ellipsoids, and 3D Gaussian density functions) and include additional degrees of freedom, such as the number of states in the system and their weights. For instance, in a high-resolution representation, a sphere can represent a single atom, while in a coarse-grained representation it may correspond to a residue. Coarse-graining can be used to encode the uncertainty arising from both static and dynamic variability. Moreover, in a “rigid body”, the relative positions of the primitives (*eg*, atoms in a domain) can be constrained, for example based on a crystallographic structure. In

most applications, the representation is determined before any other computations and is not changed. The resolution of the representation should be commensurate with the input information. In some cases, it is beneficial to represent different parts of a structure with different representations or a part may be described with several different inter-linked representations simultaneously (*ie*, multi-scale representation); in such a case, information can be applied to restrain the model by using the most convenient representation [27].

When defining the representation, we usually have to balance between the requirements of scoring and sampling. We need a representation that is sufficiently detailed for accurately assessing a match between a model and the input information. For example, when using chemical cross-linking information, we need to choose between representing the cross-linker explicitly with all of its atoms [28] or implicitly as a function of the distance between the cross-linked residues. To minimize data sparseness, we also need a representation that is sufficiently coarse, given the invariably limited information content of the data. Finally, the representation should also be sufficiently coarse to allow for exhaustive sampling of good scoring models in a feasible timeframe. While it would be best to be able to compute an optimal representation based on the input information, this is not yet possible.

Scoring

Most generally, the scoring function ranks alternative models based on the evidence provided by the input information. The scoring function should take into account the uncertainty in the input information, including sparseness, error, ambiguity, and incoherence. For example, a scoring function could evaluate whether or not a given model fits the data within its error bars. Ambiguity in the data assignment should also be accounted for by the scoring function. For example, to address the ambiguity in methyl proton assignment for an observed NOE signal, the signal is often assigned to the center of mass of the three methyl protons [25]. For a complex with multiple copies of the same protein, a cross-link can be assigned to the copy of the protein that satisfies it best [20]. When a sample is heterogeneous (*ie*, data are incoherent), a scoring function should rank instances of a model, each one of which consists of multiple structures (multi-state model). For example, protein heterogeneity in a crystal can be modeled using snapshots of molecular dynamics simulations [29••]. Protein dynamics in solution, as measured by SAXS and NMR spectroscopy, can be modeled by fitting multiple weighted conformations to the data [30–34]. In EM single particle reconstruction, heterogeneity can be addressed by multi-model reconstruction using multi-stage clustering [35].

The most objective ranking of models is in principle achieved by a Bayesian scoring function [36]. The Bayesian approach estimates the probability of a model, given information available about the system, including both prior knowledge and newly acquired experimental data. When modeling heterogeneous systems, model M includes a set of N modeled structures $X = \{X_i\}$, their population fractions in the sample $\{w_i\}$, and potentially additional parameters (*eg*, the unknown data errors). The posterior probability $p(M|D, I)$ of model M given data D and prior knowledge I is

$$p(M|D,I) \propto p(D|M,I) \cdot p(M|I)$$

where the *likelihood function* $p(D|M, I)$ is the probability of observing data D given M and I; and the *prior* $p(M|I)$ is the probability of model M given I. The likelihood function is based on the *forward model* $f(X)$ that predicts the data point that would have been observed for structure(s) X in the absence of experimental error, and a *noise model* that specifies the distribution of the deviation between the experimentally observed and predicted data points. The *Bayesian scoring function* is defined as $S(M) = -\log[p(D|M, I)\cdot p(M|I)]$ which ranks the models the same as the posterior probability. The most probable models are found by selecting the best scoring models sampled from the posterior distribution.

The Bayesian scoring function can account for most sources of uncertainty in data without over-fitting. It was successfully adopted for NMR spectroscopy data [36,37], and recently cryo-EM density maps [38,39••]. Bayesian structure determination based on sparse NOE measurements produces more accurate structures and better estimates of precision than standard NMR structure determination methods [36,37,40]. In single particle EM reconstruction, the Bayesian approach results in density maps with higher resolution than those from standard reconstruction methods using the same input datasets [38,39]; moreover, high-resolution maps can be obtained from only a few thousand of particles [41–43]. Recently, the BioEM method for Bayesian analysis of individual EM images that can deal with conformational heterogeneity was developed [44]. Bayesian scoring functions have also been developed for cysteine cross-linking [45], chemical cross-linking [46], FRET spectroscopy [47], and atomic statistical potentials [48].

The Bayesian approach is more objective than traditional scoring functions in a number of respects: (i) inference of unknown quantities, such as data error and state weights, (ii) combination of different types of information, (iii) inference of multiple structures, (iv) estimate of model precision, and (v) “marginalization” of parameters that are difficult to determine. The main disadvantage is that the model is more elaborate (*cf*, noise model and priors) and a more exhaustive sampling of structural and parameter space is required.

Sampling

A variety of optimization methods (*eg*, conjugate gradients), sampling algorithms (*eg*, Monte Carlo), and even exhaustive enumeration (*eg*, Fast Fourier Transform) can be used to find models consistent with input information. The major source of uncertainty in this stage is insufficient sampling due to the ruggedness and high dimensionality of the scoring function landscape that needs to be sampled. As a result, it is almost never certain that the best scoring models were sampled. For stochastic sampling, such as the Monte Carlo algorithm, the thoroughness of sampling can be indicated by showing that new independent runs (*eg*, using random starting configurations and different random number generator seeds) do not result in significantly different good-scoring solutions (“convergence test”) [49]. Passing such a test is a necessary but not sufficient condition for thorough sampling; a positive outcome of the test may be misleading if, for example, the landscape contains only

a narrow, and thus difficult to find, pathway to the pronounced minimum corresponding to the native state.

For multi-state models, the sampling is often performed in two steps, due to the relatively high number of degrees of freedom involved. First, a large set of possible single configurations is sampled. Second, the sets of configurations in a multi-state model are enumerated, for example by using a genetic algorithm [31,32], a maximum entropy approach [33], or a deterministic method [34].

Stage 4: Analyzing models and information

Input information and output models are analyzed in order to estimate model precision and accuracy, to detect inconsistent information and missing information, as well as to suggest most informative future experiments. There are three possible outcomes of modeling, based on the number of clusters of models and consistency between the models and information. The following discussion applies to single-state models, but similar considerations can also be extended to multi-state models. First, if only a single model (or a cluster of similar models) satisfies all restraints and thus all input information, there is probably sufficient information for determining the structure (with the precision corresponding to the variability within the cluster). Second, if two or more different models are consistent with the restraints, the information is insufficient to define the single 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 model satisfies all input information, the information or its interpretation in terms of the restraints are 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).

If multiple structural states are indicated, care must be taken that the scoring function explicitly allows for this possibility [31–34,45,46]. When a mixture of states is modeled, the number of states needs to be determined. Frequently, Occam’s razor suggests that the smallest number of states sufficient to explain the input information within some threshold is the optimal choice. An example of this approach is the “minimal ensemble” method in molecular modeling based on SAXS data [32]. 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 many widely different states; such cases are indicated by similarity between distributions of structural properties, such as the radius of gyration, of the minimal and large ensembles [31].

Once we obtain a model (single or multi-state) that satisfies the input data, we can analyze it to estimate precision and accuracy. It is impossible to know with certainty the accuracy of the proposed structure without knowing the real native structure. However, accuracy can be estimated based on rules derived from benchmark studies that involve modeling of known structures. For example, there is a strong correlation of the accuracy of an X-ray structure with the resolution of the X-ray dataset and R_{free} [1]. In addition to such broad rules, five

types of analysis that are indicative of model precision and accuracy in specific cases have been proposed [20], as follows (the first three tests are examples of statistical resampling [50]).

Estimating model precision based on variability in the ensemble of good-scoring models

The model ensemble is analyzed in terms of the precision of its features, such as the protein positions and contacts [49,51–53]; the precision is defined by the variability in the ensemble and likely provides the lower bound on its accuracy. Of particular interest are the features that are present in most configurations in the ensemble and have a single maximum in their probability distribution. The spread around the maximum describes how precisely the feature was determined by the input information. A more thorough test is performed by estimation of structural variability in multiple random subsets of the ensemble [52••,54••].

Self-consistency of the experimental data

Inconsistencies in the experimental data or its interpretation are indicated by an ensemble of models containing only frustrated structures that do not satisfy the input data, although such an outcome can also arise from the failure of sampling. If there is a model that satisfies all data, the probability of such a model occurring by chance can be indicated by statistical significance tests; if this probability is low, the model is likely to be correct. In these tests, the labels on the data points are randomized or permuted, followed by re-computing the model; for example, one can assign cross-links to random residue pairs [55].

Validating models by using random subsets of experimental data

The structure can be directly validated against experimental data that was not included in the structure calculation [52••]. This criterion is similar to the crystallographic R_{free} parameter and can be used to assess both the model accuracy and the input data [1]. Alternatively, modeling can be repeated with random subsets of the data. Common statistical techniques for this validation include cross validation and bootstrapping [50].

Reproducibility of the model with simulated data

In this approach, a native structure is assumed, the restraints to be tested are simulated from this structure, the structure is then reconstructed based only on these restraints, and finally the reconstruction is compared to the original assumed structure [49]. Using such simulations, the dependence of model accuracy on the amount, quality, and type of information can be mapped for future prediction of accuracy.

Patterns unlikely to occur by chance

Unlikely patterns emerging from mapping independent and unused data on the structure also increase our confidence in a model, similarly to validation by information not used in modeling. For example, the model of the nuclear pore complex (NPC) revealed an unexpected 16-fold pseudo-symmetry in the arrangement of fold types of the constituent proteins, in addition to the known 8-fold symmetry [49]. The 16-fold pseudo-symmetry validates the model because the fold types were not used in modeling and the 16-fold pseudo

symmetry is unlikely to arise by chance (while it can be reasonably explained by gene duplications in the evolution of the NPC).

Conclusions

Integrative structure determination needs *de facto* standards and tools for assessing the input data and resulting models, following in the footsteps of X-ray crystallography and NMR spectroscopy with established structure validation criteria.

Acknowledgments

We acknowledge support from NIH R01 GM083960 and NIH U54 GM103511 (A.S.).

References

1. Brünger AT. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. Nature. 1992; 355:472–475. [PubMed: 18481394]
2. Read RJ, Adams PD, Arendall WB, Brunger AT, Emsley P, Joosten RP, Kleywegt GJ, Krissinel EB, Lütteke T, Otwinowski Z, et al. A new generation of crystallographic validation tools for the protein data bank. Structure. 2011; 19:1395–1412. [PubMed: 22000512]
3. Montelione GT, Nilges M, Bax A, Güntert P, Herrmann T, Richardson JS, Schwieters CD, Vranken WF, Vuister GW, Wishart DS, et al. Recommendations of the wwPDB NMR Validation Task Force. Structure. 2013; 21:1563–1570. [PubMed: 24010715]
4. Henderson R, Sali A, Baker ML, Carragher B, Devkota B, Downing KH, Egelman EH, Feng Z, Frank J, Grigorieff N, et al. Outcome of the first electron microscopy validation task force meeting. 2012:205–214.
5. Jacques DA, Guss JM, Svergun DI, Trewhella J. Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution. Acta Crystallogr D Biol Crystallogr. 2012; 68:620–626. [PubMed: 22683784]
6. Trewhella J, Hendrickson WA, Kleywegt GJ, Sali A, Sato M, Schwede T, Svergun DI, Tainer JA, Westbrook J, Berman HM. Report of the wwPDB Small-Angle Scattering Task Force: data requirements for biomolecular modeling and the PDB. Structure. 2013; 21:875–881. [PubMed: 23747111]
7. Schwede T, Sali A, Honig B, Levitt M, Berman HM, Jones D, Brenner SE, Burley SK, Das R, Dokholyan NV, et al. Outcome of a workshop on applications of protein models in biomedical research. 2009:151–159.
8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Research. 2000; 28:235–242. [PubMed: 10592235]
9. Saxton WO, Baumeister W. The correlation averaging of a regularly arranged bacterial cell envelope protein. J Microsc. 1982; 127:127–138. [PubMed: 7120365]
10. Harauz G, van Heel M. Exact filters for general geometry three dimensional reconstruction. Proceedings of the IEEE Computer Vision and Pattern Recognition. 1986; 78:146–156.
11. Henderson R, Chen S, Chen JZ, Grigorieff N, Passmore LA, Ciccarelli L, Rubinstein JL, Crowther RA, Stewart PL, Rosenthal PB. Tilt-pair analysis of images from a range of different specimens in single-particle electron cryomicroscopy. Journal of Molecular Biology. 2011; 413:1028–1046. [PubMed: 21939668]
12. Scheres SHW, Chen S. Prevention of overfitting in cryo-EM structure determination. Nat Methods. 2012; 9:853–854. [PubMed: 22842542]
13. Chen S, McMullan G, Faruqi AR, Murshudov GN, Short JM, Scheres SHW, Henderson R. High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. Ultramicroscopy. 2013; 135:24–35. [PubMed: 23872039]

•14. Stagg SM, Noble AJ, Spilman M, Chapman MS. ResLog plots as an empirical metric of the quality of cryo-EM reconstructions. Journal of Structural Biology. 2014; 185:418–426. A method for assessing the accuracy of cryo-EM reconstructions is presented. A plot of inverse resolution vs. the logarithm of the number of single particles (a “ResLog” plot) provides metrics for the reliability of the reconstruction and the overall quality of the dataset and processing. [PubMed: 24384117]

••15. Rambo RP, Tainer JA. Accurate assessment of mass, models and resolution by small-angle scattering. Nature. 2013; 496:477–481. A statistical method based on the Nyquist–Shannon sampling and the noisy-channel coding theorems is used for evaluating structural models against SAS data. [PubMed: 23619693]

16. Rambo RP, Tainer JA. Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. Biopolymers. 2011; 95:559–571. [PubMed: 21509745]

17. Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, Schwede T. The Protein Model Portal--a comprehensive resource for protein structure and model information. Database (Oxford). 2013; 2013:bat031–bat031. [PubMed: 23624946]

18. Kryshtafovych A, Barbato A, Fidelis K, Monastyrskyy B, Schwede T, Tramontano A. Assessment of the assessment: evaluation of the model quality estimates in CASP10. Proteins. 2014; 82 (Suppl 2):112–126. [PubMed: 23780644]

19. Sali A, Glaeser R, Earnest T, Baumeister W. From words to literature in structural proteomics. Nature. 2003; 422:216–225. [PubMed: 12634795]

20. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, et al. Determining the architectures of macromolecular assemblies. Nature. 2007; 450:683–694. [PubMed: 18046405]

21. Russel D, Lasker K, Webb B, Velázquez-Muriel J, Tjioe E, Schneidman-Duhovny D, Peterson B, Sali A. Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. Plos Biol. 2012; 10:e1001244. [PubMed: 22272186]

22. Ward AB, Sali A, Wilson IA. Biochemistry. Integrative structural biology. Science. 2013; 339:913–915. [PubMed: 23430643]

23. Habeck M. Statistical mechanics analysis of sparse data. J Struct Biol. 2011; 173:541–548. [PubMed: 20869444]

24. Borek D, Otwinowski Z. Everything Happens at Once – Deconvolving Systematic Effects in X-ray Data Processing. Advancing Methods for Biomolecular Crystallography. 2013:105–112.

25. Guentert P, Braun W, Billeter M. Automated stereospecific proton NMR assignments and their impact on the precision of protein structure determinations in solution. J Am Chem Soc. 1989; 111:3997–4004.

26. Duhovny, D.; Nussinov, R.; Wolfson, HJ. Efficient Unbound Docking of Rigid Molecules. Proceedings of the 2’nd Workshop on Algorithms in Bioinformatics (WABI); 2002. p. 185-200.

27. Murtola T, Bunker A, Vattulainen I, Deserno M, Karttunen M. Multiscale modeling of emergent materials: biological and soft matter. Phys Chem Chem Phys. 2009; 11:1869–1892. [PubMed: 19279999]

28. Bahaman A, Malmström L, Aebersold R. Xwalk: computing and visualizing distances in cross-linking experiments. Bioinformatics. 2011; 27:2163–2164. [PubMed: 21666267]

••29. Burnley BT, Afonine PV, Adams PD, Gros P. Modelling dynamics in protein crystal structures by ensemble refinement. Elife. 2012; 1:e00311–e00311. An ensemble of structures is used for refinement of high-resolution X-ray diffraction datasets resulting in a better fit to the data than a single structure. [PubMed: 23251785]

30. Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M. Simultaneous determination of protein structure and dynamics. Nature. 2005; 433:128–132. [PubMed: 15650731]

31. Bernado P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI. Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. J Am Chem Soc. 2007; 129:5656–5664. [PubMed: 17411046]

32. Pelikan M, Hura GL, Hammel M. Structure and flexibility within proteins as identified through small angle X-ray scattering. Gen Physiol Biophys. 2009; 28:174–189. [PubMed: 19592714]
33. Różycki B, Kim YC, Hummer G. SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions. Structure/Folding and Design. 2011; 19:109–116.
34. Berlin K, Castañeda CA, Schneidman-Duhovny D, Sali A, Nava-Tudela A, Fushman D. Recovering a representative conformational ensemble from underdetermined macromolecular structural data. J Am Chem Soc. 2013; 135:16595–16609. [PubMed: 24093873]
35. Shatsky M, Hall RJ, Nogales E, Malik J, Brenner SE. Automated multi-model reconstruction from single-particle electron microscopy data. J Struct Biol. 2010; 170:98–108. [PubMed: 20085819]
36. Rieping W, Habeck M, Nilges M. Inferential structure determination. Science. 2005; 309:303–306. [PubMed: 16002620]
37. Nilges M, Bernard A, Bardiaux B, Malliavin T, Habeck M, Rieping W. Accurate NMR structures through minimization of an extended hybrid energy. Structure/Folding and Design. 2008; 16:1305–1312.
38. Scheres SHW. RELION: implementation of a Bayesian approach to cryo-EM structure determination. J Struct Biol. 2012; 180:519–530. [PubMed: 23000701]
••39. Scheres SHW. A Bayesian view on cryo-EM structure determination. J Mol Biol. 2012; 415:406–418. A Bayesian formulation of cryo-EM structure determination is presented, where smoothness in the reconstructed density is imposed through a Gaussian prior in the Fourier domain. The structure and the parameters are determined from the data without user bias. [PubMed: 22100448]
40. Rieping W, Nilges M, Habeck M. ISD: a software package for Bayesian NMR structure calculation. Bioinformatics. 2008; 24:1104–1105. [PubMed: 18310055]
41. Fernández IS, Bai X-C, Hussain T, Kelley AC, Lorsch JR, Ramakrishnan V, Scheres SHW. Molecular architecture of a eukaryotic translational initiation complex. Science. 2013; 342:1240585–1240585. [PubMed: 24200810]
42. Bai X-C, Fernández IS, McMullan G, Scheres SH. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. Elife. 2013; 2:e00461–e00461. [PubMed: 23427024]
43. Sauerwald A, Sandin S, Cristofari G, Scheres SHW, Lingner J, Rhodes D. Structure of active dimeric human telomerase. Nat Struct Mol Biol. 2013; 20:454–460. [PubMed: 23474713]
44. Cossio P, Hummer G. Bayesian analysis of individual electron microscopy images: towards structures of dynamic and heterogeneous biomolecular assemblies. J Struct Biol. 2013; 184:427–437. [PubMed: 24161733]
45. Molnar KS, Bonomi M, Pellarin R, Clinthorne GD, Gonzalez G, Goldberg SD, Sali A, DeGrado WF. Cys-scanning Disulfide crosslinking and Bayesian modeling suggest scissoring motions in the histidine kinase, PhoQ. Structure. (in press).
46. Street TO, Zeng X, Pellarin R, Bonomi M, Sali A, Kelly MJS, Chu F, Agard DA. Elucidating the mechanism of substrate recognition by the bacterial Hsp90 molecular chaperone. J Mol Biol. 2014; 426:2393–404. [PubMed: 24726919]
47. Bonomi M, Muller EGD, Pellarin R, Kim SJ, Russel D, Ramsden R, Sundin BA, Davis TN, Sali A. Protein complex structures from Bayesian modeling of *in vivo*. FRET data.
48. Dong GQ, Fan H, Schneidman-Duhovny D, Webb B, Sali A. Optimized atomic statistical potentials: assessment of protein interfaces and loops. Bioinformatics. 2013; 29:3158–3166. [PubMed: 24078704]
49. Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, et al. The molecular architecture of the nuclear pore complex. Nature. 2007; 450:695–701. [PubMed: 18046406]
50. Efron, B.; Efron, B. The jackknife, the bootstrap and other resampling plans. 1982.
51. Tjong H, Gong K, Chen L, Alber F. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. Genome Res. 2012; 22:1295–1305. [PubMed: 22619363]
••52. Lasker K, Förster F, Bohn S, Walzthoeni T, Villa E, Unverdorben P, Beck F, Aebersold R, Sali A, Baumeister W. Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. Proc Natl Acad Sci US A. 2012; 109:1380–1387. The architecture of the

26S proteosome is determined using a cryo-EM map and cross-links. The final structural ensemble is validated using new data (cross links from a different species, domain contacts, atomic models) and jack-knifing by omitting one restraint at a time.

53. Loquet A, Sgourakis NG, Gupta R, Giller K, Riedel D, Goosmann C, Griesinger C, Kolbe M, Baker D, Becker S, et al. Atomic model of the type III secretion system needle. Nature. 2012; 486:276–279. [PubMed: 22699623]

••54. Murakami K, Elmlund H, Kalisman N, Bushnell DA, Adams CM, Azubel M, Elmlund D, Levi-Kalisman Y, Liu X, Gibbons BJ, et al. Architecture of an RNA polymerase II transcription pre-initiation complex. Science. 2013; 342:1238724–1238724. The architecture of an RNA polymerase II transcription pre-initiation complex is determined using a cryo-EM map and cross-links. The final ensemble of models is validated using bootstrapping. [PubMed: 24072820]

55. Kalisman N, Adams CM, Levitt M. Subunit order of eukaryotic TRiC/CCT chaperonin by cross-linking, mass spectrometry, and combinatorial homology modeling. Proc Natl Acad Sci US A. 2012; 109:2884–2889.

56. Tosi A, Haas C, Herzog F, Gilmozzi A, Berninghausen O, Ungewickell C, Gerhold CB, Lakomek K, Aebersold R, Beckmann R, et al. Structure and subunit topology of the INO80 chromatin remodeler and its nucleosome complex. Cell. 2013; 154:1207–1219. [PubMed: 24034245]

57. Ciferri C, Lander GC, Maiolica A, Herzog F, Aebersold R, Nogales E. Molecular architecture of human polycomb repressive complex 2. Elife. 2012; 1:e00005–e00005. [PubMed: 23110252]

58. Greber BJ, Boehringer D, Leitner A, Bieri P, Voigts-Hoffmann F, Erzberger JP, Leibundgut M, Aebersold R, Ban N. Architecture of the large subunit of the mammalian mitochondrial ribosome. Nature. 2014; 505:515–519. [PubMed: 24362565]

••59. Erzberger JP, Stengel F, Pellarin R, Zhang S, Schaefer T, Ayelett CH, Cimermančič P, Boehringer D, Sali A, Aebersold R, et al. Molecular architecture of the 40S•eIF1•eIF3 translation initiation complex. Cell. (in press). The architecture of eukaryotic initiation factor 3 (eIF3) in complex with 40S ribosomal subunit was determined through integration of chemical cross-linking and crystallographic structures by a Bayesian approach. The model is validated using EM density map.

60. Politis A, Stengel F, Hall Z, Hernández H, Leitner A, Walzthoeni T, Robinson CV, Aebersold R. A mass spectrometry-based hybrid method for structural modeling of protein complexes. Nat Methods. 201410.1038/nmeth.2841

•61. Boura E, Różycki B, Herrick DZ, Chung HS, Vecer J, Eaton WA, Cafiso DS, Hummer G, Hurley JH. Solution structure of the ESCRT-I complex by small-angle X-ray scattering, EPR, and FRET spectroscopy. Proc Natl Acad Sci US A. 2011; 108:9437–9442. A multi-state model consisting of six conformations of ESCRT-1 is used for interpretation of SAXS and EPR data. The model is validated using FRET data.

62. Huang J-R, Warner LR, Sanchez C, Gabel F, Madl T, Mackereth CD, Sattler M, Blackledge M. Transient electrostatic interactions dominate the conformational equilibrium sampled by multi-domain splicing factor U2AF65: A combined NMR and SAXS study. J Am Chem Soc. 201410.1021/ja502030n

63. Deshmukh L, Schwieters CD, Grishaev A, Ghirlando R, Baber JL, Clore GM. Structure and dynamics of full-length HIV-1 capsid protein in solution. J Am Chem Soc. 2013; 135:16133–16147. [PubMed: 24066695]

64. Baù D, Sanyal A, Lajoie BR, Capriotti E, Byron M, Lawrence JB, Dekker J, Marti-Renom MA. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. Nat Struct Mol Biol. 2011; 18:107–114. [PubMed: 21131981]

65. Baù D, Marti-Renom MA. Structure determination of genomic domains by satisfaction of spatial restraints. Chromosome Res. 2011; 19:25–35. [PubMed: 21190133]

••66. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol. 2012; 30:90–98. A structural modeling procedure that computes a population of 3D genome structures from the TCC data is introduced. [PubMed: 22198700]

Box 1

Glossary

Input data	experimental data used to compute a model
Input information	experimental data and any additional information
Data sparseness	a measure of the amount of data relative to the number of degrees of freedom in the model
Data error	the difference between the measured data and its true value, which can be computed given a forward model and the true structure; data error can be random and/or systematic, affecting the precision and the accuracy of the measured data
Data ambiguity	a data point is ambiguous when it cannot be assigned to the specific components of the model
Data incoherence	a dataset is incoherent when it is derived from a compositionally or configurationally heterogeneous sample
Single-state model	a model that specifies a single structural state and value for any other parameter
Multi-state model	a model that specifies two or more co-existing structural states and values for any other parameter
Ensemble of structural models	a set of structural models each one of which is consistent with the data
Ensemble precision	variability among structural models in the ensemble
Error or accuracy of a structural model	the difference between the structural model and the true structure(s)
Representation resolution	a descriptor of the detail in the representation of the structural model (*eg*, atomic models consist of atoms)

Highlights

- Integrative modeling needs standards and tools for assessing models and input data
- Model uncertainty originates from sparse, noisy, ambiguous, or incoherent data
- Model uncertainty also originates from representation, scoring function and sampling
- Some methods for assessing data and models are listed

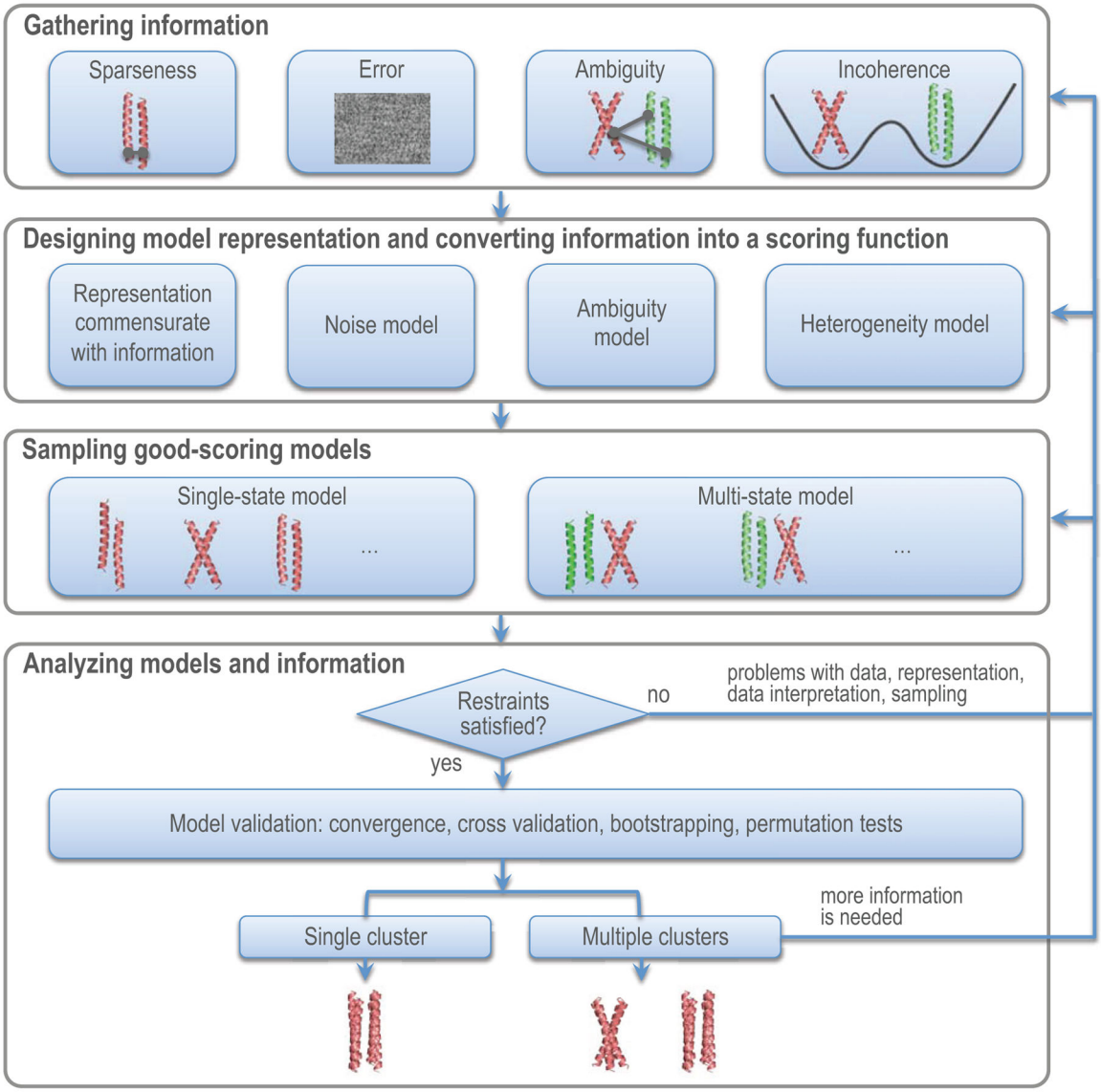

Figure 1.

Uncertainty in integrative structure modeling. The four-stage scheme of integrative structure modeling is used to describe how to approach uncertainty in the data and the models. The collected information is converted into a scoring function that accounts for data error, ambiguity, and incoherence. The model representation should reflect data sparseness. After sampling, if good-scoring models satisfy the restraints, they are further evaluated by structural clustering and data validation tests.

Table 1

Some of the recent structures solved by an integrative approach.

Structure	Experimental information	Method
S. cerevisiae INO80 [56]	Cryo-EM map (17Å resolution), 212 intra-protein and 116 inter-protein cross-links	Manual modeling in Chimera
Polycomb Repressive Complex 2 [57]	Negative stain EM map (21Å resolution) and ~60 intra-protein and inter-protein cross-links	Manual modeling in Chimera
39S large subunit of the porcine mitochondrial ribosome [58]	Cryo-EM map (4.9Å resolution) and ~70 inter-protein cross-links	COOT, O, PHENIX
S. pombe 26S holocomplex [52••]	Cryo-EM map (8.4Å resolution) and 35 cross-links from *S. pombe* and 36 cross links from *S. cerevisiae*	IMP
S. cerevisiae RNA polymerase II transcription pre-initiation complex [54••]	Cryo-EM map (16Å resolution), 157 intra-protein and 109 inter-protein cross-links	Exhaustive enumeration
S. cerevisiae 40S•eIF1•eIF3 translation initiation complex [59••]	965 cross-links, including 126 unique eIF3-eIF3 and 40S•eIF1-eIF3 cross links, negative stain EM map (28Å resolution), crystallographic structures of 40S complex, eIF3 domains	IMP
S. typhimurium Type III secretion system needle [53]	solid-state NMR, cryo-EM (19.5Å resolution)	Rosetta
Methane monooxygenase hydroxylase (MMOH), toluene/o-xylene monooxygenase hydroxylase (ToMOH), and urease [60]	Composition and stoichiometry from native MS, collision cross section from ion mobility–MS and cross-links	IMP
ESCRT-I complex [61•]	SAXS, double electron-electron transfer (DEER), and FRET	EROS
Hsp90 substrate recognition [46]	31 cross-links and NMR spectroscopy	IMP
Splicing factor U2AF65 [62]	Paramagnetic relaxation enhancement (PRE), residue dipolar couplings (RDCs), and SAXS	ASTEROIDS
HIV-1 capsid protein [63]	RDCs and SAXS	Xplor-NIH
500-kilobase (kb) domain of human chromosome 16 [64,65]	Chromosome Conformation Capture Carbon Copy (5C) experiments and excluded volume	IMP
Human genome architecture [66••]	Tethered chromosome conformation capture (TCC) and population-based modeling	IMP