Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

# Uncertainty in integrative structural modeling

Dina Schneidman-Duhovny[1], Riccardo Pellarin[1] and Andrej Sali[1,2]

Integrative structural modeling 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.

**Addresses**
[1] Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA 94158, USA
[2] Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences (QB3), University of California, San Francisco, CA 94158, USA

Corresponding authors: Schneidman-Duhovny, Dina (dina@salilab.org), Sali, Andrej (sali@salilab.org)

**Current Opinion in Structural Biology** 2014, **28**:96–104

This review comes from a themed issue on **Biophysical and molecular biological methods**

Edited by **David P Millar** and **Jill Trewhella**

For a complete overview see the Issue and the Editorial

Available online 28th August 2014

**http://dx.doi.org/10.1016/j.sbi.2014.08.001**



## 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 (e.g. 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 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 (Box 1 and Figure 1). 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

**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.

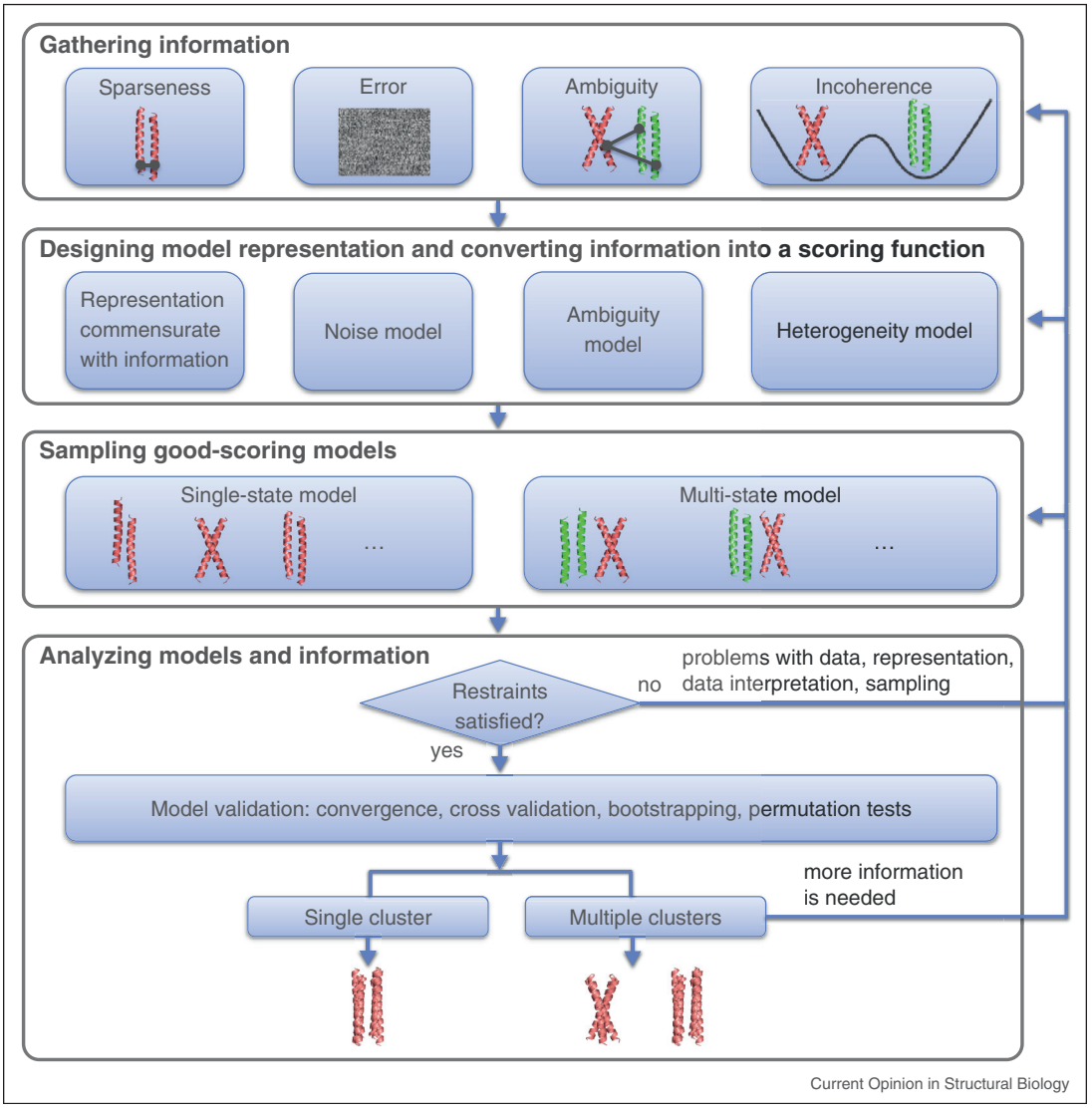
**Ensemble of structural models** — a set of structural models each one of which is consistent with the data.

**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 (e.g. atomic models consist of atoms).

**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.

[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