Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Hybrid methods for macromolecular structure determination: experiment with expectations

Gunnar F Schröder[1,2]

Studies of large and heterogeneous macromolecules often yield low-resolution data that alone does not suffice to build accurate atomic models. Adding information from molecular simulation or other structure prediction methods can lead to models with significantly better quality. Different strategies are discussed to combine experimental data with results from simulation and prediction. This review describes recent approaches for building atomic models with a focus on X-ray diffraction and single-particle cryo-electron microscopy (cryo-EM) data. In addition, both cryo-EM and X-ray diffraction provide information on molecular dynamics. Therefore, the best description of molecular structures is often by an ensemble of models. It furthermore becomes apparent that using raw data for the modeling ensures that all information obtained by the experiment can be fully exploited. It is also important to quantify the errors of both experiment and simulation to correctly weigh their different contributions.

Addresses
[1] Institute of Complex Systems, Structural Biochemistry (ICS-6), Forschungszentrum Jülich, 52425 Jülich, Germany
[2] Physics Department, Heinrich-Heine Universität Düsseldorf, 40225 Düsseldorf, Germany

Corresponding author: Schröder, Gunnar F (gu.schroeder@fz-juelich.de)

Current Opinion in Structural Biology 2015, **31**:20–27

This review comes from a themed issue on **Theory and simulation**

Edited by **Claire Lesieur** and **Klaus Schulten**

For a complete overview see the Issue and the Editorial

Available online 18th March 2015

http://dx.doi.org/10.1016/j.sbi.2015.02.016

Hybrid modeling in structural biology describes the combination of computational modeling with experimental data to determine macromolecular structures (cf. Figure 1). It has become particularly important to determine structures with low-resolution or sparse data, where the data alone would not suffice to build molecular models. Hybrid modeling is often used synonymously with integrative modeling, which emphasizes more the simultaneous use of different types of experimental information in the structure determination process.

Even though hybrid modeling in structural biology is a highly modern topic and a very active field of research, its birth can be traced to the method proposed by Jack and Levitt in 1978 [1] who introduced a hybrid energy function to optimize the energy of a protein at the same time as its fit to X-ray diffraction data. The fit of the protein model to the data has been defined ad hoc as an energy term E_{Data}, which is added with a weight w to the molecular mechanics energy E_{MM} of the protein

$$E_{\mathrm{Hybrid}} = E_{\mathrm{MM}} + wE_{\mathrm{Data}}.$$

E_{Data} in its most basic form simply describes the deviation of experimental observables from those calculated from the model, e.g. $E_{\mathrm{Data}} = \sum_{hkl}(F_{\mathrm{obs}}(hkl) - F_{\mathrm{calc}}(hkl))^2$ for diffraction data.

Minimization of this combined hybrid energy function, E_{Hybrid}, yields a refined structure that fulfills both the restraints imposed by the experimental data, as well as the stereo-chemical restraints, which represent information we have on protein structures in general. Such refinement is instrumental in the interpretation of the data, e.g. in the case of crystallographic data, to improve phase information. Later this hybrid approach made it also possible to determine protein structures using restraints derived from NMR experiments.

These early developments lead 30 years later to the notion that including as much information as possible is the best way of building models that are as accurate as possible [2•]. A tour de force in such exhaustive integrative modeling was the determination of the molecular architecture of the nuclear pore complex [3,4].

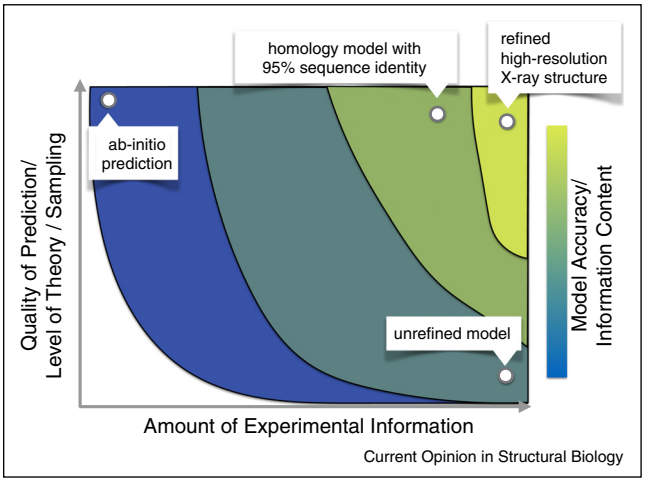
This review focuses on the use of intermediate to low-resolution X-ray diffraction and single-particle cryo-electron microscopy (cryo-EM) data to determine atomic models of protein structures. In particular cryo-EM have made tremendous progress in the past few years, which spurred the development of several new computational model-building techniques.

The classical approach of building a single model that best fits the data is still prevalent even though the uncertainty in the modeling process could be captured more appropriately by generating an entire ensemble of models. However, the determination of model ensembles poses significant challenges, as will be discussed further below.

Refinement of a single model

At intermediate to low-resolution (4–8 Å) the observation-to-parameter ratio is too low to completely

Figure 1



The motivation for hybrid modeling is that the more information is used to build a model the more accurate it will be. Both experiment and simulation should be considered information, both improve the accuracy of a structural model.

determine the atomic coordinates from the data alone. This problem can be solved by either reducing the number of parameters (e.g. by allowing only torsional degrees of freedom) or by adding information. The additional information could come from very different sources.

For example known structures of homologous proteins can be used to guide the refinement and effectively reduce the number of degrees of freedom, since it is known that homologous proteins fold into similar structures. Approaches that exploit this similarity such as the deformable elastic network (DEN) [5,6], jelly-body, or reference model [7–9] restraints have been implemented in crystallographic refinement programs.

Another source of additional information are simulation and structure prediction techniques, which use molecular mechanics force fields. Such force fields bias and confine the sampling of the conformational space to physically realistic and energetically favorable conformations. When interpreting both the experimental observations as well as the added molecular mechanics energy function as general information about the protein structure, the information-to-parameter is increased, which facilitates the structure determination. For example electrostatic interactions had disappeared completely from standard crystallographic refinement procedures, but it has recently been shown to improve the refinement [10]; even at high resolution when accounting for polarizability and anisotropic structure factors [11,12].

More exhaustive sampling of protein conformations using all-atom explicit solvent MD simulations allow for larger conformational changes and can lead to significant improvement of the refinement with increased radius of convergence and better phases [13] as compared to standard crystallographic refinement.

Similarly, the combination of energy-guided remodeling by the program Rosetta with the Autobuild and (now also real-space [14]) refinement tools of the program Phenix shows significant improvement in the refinement [15,16]. This clearly demonstrates that force field/energy functions can provide valuable information that help to build better models or even solve structures that could otherwise not have been solved.

The same strategies have been followed in the refinement of (high-resolution) X-ray protein structures against (lower resolution) cryo-EM density maps, which is also referred to as flexible fitting. Several flexible fitting methods have been developed and they, again, differ in the type of additional information that is used during the structure refinement: Normal mode based approaches (e.g. NORMA [17], NMFF [18]) directly deform the high-resolution structure along the first few normal modes and thereby reduce the number of degrees of freedom. Other methods use information from a reference model (typically the starting high-resolution structure): For example the program DireX [19] employs DEN restraints (but no normal modes) and MDfit [20] uses Go-type structure-based potentials. Another class of methods rely mostly on molecular energy functions (force fields), such as the molecular dynamics based fitting methods MDFF [21] or Tama's approach [22]; although in practice some MD based fitting methods often use restraints to the starting structure (e.g. to maintain secondary structure). MDFF has been applied successfully also to very large systems, for example for building a model for the entire HIV capsid [23]. Since flexible fitting is not the focus of this review the existing programs are far from completely covered and the reader is referred to very good reviews on flexible fitting and modeling with cryo-EM density maps which have been recently published [24–27].

Most of these flexible fitting methods work in real-space, however, crystallographic refinement programs which work in reciprocal space have also been used successfully [28] for this task. In that case the EM density map needs to be converted to structure factors. However, how to best translate errors in the EM reconstructions (e.g. quantified by Fourier shell correlation) to errors in the complex structure factors for use in maximum likelihood target functions still remains to be worked out.

When fitting structures against low-resolution data, over-fitting is a major concern. The standard tool to detect over-fitting is cross-validation, where a portion of the data (the test set), which needs to be independent from the