Determining protein structural ensembles using cryo-electron microscopy
Massimiliano Bonomi and Michele Vendruscolo

Achieving a comprehensive understanding of the behaviour of proteins is greatly facilitated by the knowledge of their structures, thermodynamics and dynamics. All this information can be provided in an effective manner in terms of structural ensembles. A structural ensemble can be obtained by determining the structures, populations and interconversion rates for all the main states that a protein can occupy. To reach this goal, integrative methods that combine experimental and computational approaches provide powerful tools. Here we focus on cryo-electron microscopy, which has become over recent years an invaluable resource to bridge the gap from order to disorder in structural biology. In this review, we provide a perspective of the current challenges and opportunities in determining protein structural ensembles using integrative approaches that can combine cryo-electron microscopy data with other available sources of information, along with an overview of the tools available to the community.

Address
Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom

Corresponding authors: Bonomi, Massimiliano (massimiliano.bonomi@pasteur.fr), Vendruscolo, Michele (mv245@cam.ac.uk)

1 Current address: Structural Bioinformatics Unit, Institut Pasteur, CNRS UMR 3928, 75015 Paris, France.

Current Opinion in Structural Biology 2018, 56:37–45
This review comes from a themed issue on Sequences and topology
Edited by Anna Panchenko and Monika Fuxreiter

https://doi.org/10.1016/j.sbi.2018.10.006
0959-440X/© 2018 Elsevier Ltd. All rights reserved.

Understanding protein behaviour
All proteins are to some extent dynamic as they populate a variety of different states [1,2]. Such behaviour is often required to achieve specific functions [3] and encompasses a wide range of time and length scales, from side-chains motions [4] and loop jiggling [5] to larger conformational changes [6]. In many cases, proteins present entirely disordered regions that can extend to encompass their entire sequences, as in the case of disorder proteins [7]. Proteins in this class constitute perhaps one-third of the human proteome, are highly involved in regulation and signaling processes [8], and play a central role in a wide variety of human disorders [9]. More generally, every macromolecular system exhibits a spectrum of conformational motions that spans a continuum between order and disorder [10].

As a consequence, understanding the behaviour of proteins requires the simultaneous determination of the structures of all the relevant states that these molecules can occupy under certain external conditions, their populations (i.e. their thermodynamics), and the rates of interconversion among these states (i.e. their dynamics). Therefore, the problem that one should strive to solve to shed light into protein behaviour is the determination of a structural ensemble (Figure 1a), as defined by a set of structures (Figure 1b), along with the statistical weights quantifying their relative populations (Figure 1c) and the transition rates measuring dynamics (Figure 1d). The central question thus becomes: operatively, how can one determine these structural ensembles?

Integrative methods for protein structural ensemble determination
In recent years, integrative (or hybrid) methods that combine computational and experimental approaches have been proven very successful in determining the structures of complex biological systems at various resolutions [11,12]. These methods have also been extended to enable the determination of structural ensembles using ensemble-averaged experimental data, such as those provided by nuclear magnetic resonance (NMR) spectroscopy [13] and small angle X-ray scattering (SAXS) [14], or distributions of experimental observations over the ensemble, such as Förster resonance energy transfer (FRET) [15] or double electron-electron resonance (DEER) data [16]. Without entering into the details of these methods, which are covered in many excellent reviews [17**,18*,19–22], integrative ensemble approaches are aimed at properly combining all the information available to overcome the limitations of individual techniques and maximize the accuracy of the resulting structures or structural ensembles. In most cases, a physico-chemical a priori description of the system (e.g. a force field) is augmented by additional conformational restraints aimed at enforcing an agreement of the structural ensemble with the available experimental data [17**]. These terms are encoded into a hybrid energy function, which is used to drive sampling methods like Monte Carlo (MC) or molecular dynamics
(MD) and ultimately to generate structural models consistent with the input information [17**].

Challenges in the implementation of protein structural ensemble determination methods

The procedure outlined above presents three difficult challenges (Figure 2), whose importance should not be underestimated [18*]. The first one is obtaining an accurate hybrid energy function (Figure 2a). Current force fields for MC and MD simulations, despite continuous progress [23,24*], still provide only an approximate prior description of the system, and therefore they should be complemented by highly-informative experimental data. The second challenge is that experimental data, as well as the theoretical models used to predict the data from a conformation and assess the consistency with experimental observations (i.e. the predictors or forward models), are always affected by random and systematic errors [25]. Quantifying such errors is crucial to avoid enforcing a too strong agreement with noisy or incorrect data, and ultimately to properly integrate all pieces of information together into a well-balanced energy function (Figure 2b). The third challenge is that the free energy landscape of a complex system is almost invariably difficult to sample at reasonable computational cost by MC or MD methods, as these simulations can easily remain trapped in local free energy minima. Therefore, the use of enhanced sampling techniques [26], such as umbrella sampling [27], metadynamics [28] or replica exchange [29], is almost always necessary to exhaustively explore the conformational landscape (Figure 2c).

Provided that all these challenges are successfully overcome, integrative methods provide excellent tools to maximize the use of all the information available, to model accurate structural ensembles, and ultimately to obtain exquisite insights into protein behaviour.

The cryo-electron microscopy revolution

Among the different types of experimental data that can be used in integrative modelling, cryo-electron microscopy (cryo-EM) has become an invaluable source of information [30]. Stimulated by the recent developments in both instrumentation and image-processing software [31,32], over the last years cryo-EM has allowed the structures of complexes and individual proteins of extraordinary biological importance to be determined at nearly-atomic resolution, thus establishing itself as a powerful technique in structural biology [33,34*].

Cryo-EM has the potential to provide an exhaustive characterization of the entire conformational landscape...
Major challenges in protein structural ensemble determination. Integrative methods for protein ensemble determination present three fateful challenges. (a) The first challenge is obtaining an accurate energy function to drive MD or MC simulations, which should incorporate experimental and prior information on the system. This energy function could be thought as a compass, which the distressed traveller in the figure is missing, thus failing to reach his destination. (b) The second challenge consists in accurately quantifying all the sources of error and uncertainty, both in the experimental data and in the predictors used in the modelling. Lacking in this quantification is represented by the missing map in the hand of the traveller, again preventing him from arriving to the oasis. (c) The final challenge is related to having enough resources to sample the conformational space of the system, or in other words achieving convergence in integrative simulations. This challenge is symbolized by the lack of water, which is the missing bottle at the belt of the traveller. Accurate structural ensembles, here represented by the oasis, can only be obtained if the three challenges are successfully addressed. A different destiny (illustrated by the condor) is waiting for the distressed traveller, if any of these three elements is missing. This illustration has been prepared by Giulia Vecchi.

The image-processing of the raw cryo-EM data has the desirable outcome of generating a small number of density maps, which represent the system in distinct conformational states [37,38], often at high resolution. These density maps provide a picture of the highly-populated free energy minima in the conformational landscape, but intermediates along the pathways that connect these states might remain elusive. In this regard, integrating cryo-EM data with molecular simulations or other experimental data might help connect the dots. In addition, in many cases part of the conformational space visible in the 2D single-particle images is discarded, and thus lost, or averaged out in the classification and reconstruction process. In this situation, low-resolution areas of high-resolution density maps might conceal highly dynamic parts of the system, provided that these regions result from the averaging of images in different conformations and not from the noise caused, for example, by radiation damage. When focused classification [39,40*] of the highly dynamic regions is not able to resolve individual conformations, integrative ensemble approaches [17**,18*,19–22] can be particularly helpful for providing an exhaustive characterization of the conformational landscape of the system.

Determination of protein structural ensembles with cryo-EM data

We will not focus our attention here on maximum-likelihood and Bayesian methods for 2D particle classification or 3D reconstruction in conformationally heterogeneous systems, nor on approaches that are aimed at determining a single structural model from 2D images or 3D reconstructions, as these methods are already covered in other recent reviews [36,41]. We note in particular that in addition to individual structures, many of these methods [42–54] generate ensembles of structural models that reflect the limited information available on the systems, and thus the fact that different models might be equally consistent with the input data. We refer to these ensembles as uncertainty ensembles [17**]. The characteristic of these ensembles is that they converge more and more to individual structures as the amount of experimental information used to derive them is increased. These uncertainty ensembles are not aimed to reflect the conformational heterogeneity arising from the internal
dynamics of the systems, which is instead described by the thermodynamic ensembles described in Figure 1. This section will provide an overview of the approaches that model thermodynamic ensembles using cryo-EM 2D and 3D data along with other available information and that can deal with systems with a continuous spectrum of dynamics. Particular attention will be given to those methods that are implemented and distributed in open-source, freely-available software.

BioEM [55**] is a Bayesian approach that uses a probabilistic framework to assess the consistency between a structural model and a set of single-particle cryo-EM images. The major advantages of this approach are the use of the raw data (Figure 3a), before any clustering or averaging procedure, and the fact that it accounts for uncertainties in the position and orientation of the system as well as noise in the experimental images. A maximum parsimony approach [17**] can also be used to identify a
minimal ensemble that can collectively explain the data, as demonstrated in the case of the ESCR T I–II supercomplex [55**]. The method is implemented in a high-parallelized and GPU-accelerated open source software [56], freely-available at https://bioem.readthedocs.io.

The EMageFit approach [57] is based on a maximum-likelihood score that quantifies the agreement with 2D class averages (Figure 3b), measured in terms of the cross-correlation between experimental and predicted images, as well as with other experimental data and any prior information available. This method has been used to determine the conformational states of the human TIR-Tf complex [57], in combination with cross-linking/mass-spectrometry data and proximity information, and of the architecture of the yeast spindle pole body [58], in combination with in vivo FRET, yeast two-hybrids, SAXS, and X-ray crystallography data. EMageFit is implemented in the Integrative Modeling Platform (IMP, http://integrativemodeling.org) [59], an open-source, freely-available library for integrative modelling using a wide variety of experimental data.

The Mosaics–EM method [60] utilizes a MC sampling procedure with natural moves, combined with a simulated annealing protocol [61], to refine conformations using 2D class averages (Figure 3b). Similarly to EMageFit, this approach relies on a hybrid-energy function to quantify the agreement of a model with respect to the experimental 2D class averages in terms of cross-correlation between target and model projections. This method was used to resolve the conformational heterogeneity of the M. maripaludis chaperonin system [60], which is known to populate a mixture of open and close states.

The recently proposed metainference approach [62] combines a general Bayesian method to model structural ensembles using noisy data with a maximum entropy approach to take into account the ensemble-averaged nature of experimental data [13,63]. While metainference was initially applied to the characterization of disordered systems using NMR data [64], it has recently been extended to model conformational ensembles using cryo-EM 3D maps (Figure 3c) [65,66**]. The goal of this method is to unravel the conformational heterogeneity hidden in low-resolution areas of high-resolution maps and to properly combine cryo-EM with other experimental data. Therefore, it is particularly useful for characterizing highly flexible regions of a system, when focused classification methods might fail. Metainference for cryo-EM (EMMI) is based on a Bayesian probabilistic framework that accounts for both ensemble-averaging over multiple conformations and the presence of non-uniform noise across the map. Sampling is carried out by MD, possibly in combination with metadynamics [28] to accelerate the exploration of the conformational space [67]. EMMI has been applied to the simultaneous determination of structure and dynamics of the STRA6 integral membrane receptor [66**] and of the ClpP protease [68]. The approach is implemented in the ISDB module [69] of the open-source PLUMED library [70], freely available at www.plumed.org.

Integrating cryo-EM with other methods to improve the quality of protein structural ensembles

Integrating medium- and low-resolution cryo-EM data (and similarly X-ray crystallography data in the 3–5 Å resolution range [71]) with other sources of information enables one to overcome the issue of data sparseness [25] and thus to obtain more accurate single-structure models. Some examples include the combination of cryo-EM with NMR [72], SAXS [73], and cross-linking/mass-spectrometry [74] data. In this regard, Bayesian inference [65,75,76] provides a robust framework to objectively weight each piece of information used in the modelling by automatically quantifying its level of noise or uncertainty.

As the resolution of cryo-EM data is rapidly approaching atomic detail, one could ask in which situations we can still benefit from integrating cryo-EM with other sources of information. One case is when the single-particle images, which typically have a low signal-to-noise ratio, are used, as for example in the BioEM approach [55**]. Another example is when high-resolution 3D maps present areas at lower resolution, due to averaging over multiple conformations during the classification and reconstruction process, as used in the EMMI approach [65,66**]. In both cases, integrating cryo-EM with other experimental data or more accurate prior information (such as explicit descriptions of solvent and membrane environments) can help increase the accuracy of the reconstructed ensemble. One particularly promising combination is with NMR observables, which in most cases, such as chemical shifts and residual dipolar couplings, report average quantities over an ensemble of conformations. These ensemble-averaged data can be optimally integrated with cryo-EM in the metainference approach [62] or using other reweighting techniques [77] to modify the ensemble obtained from cryo-EM data alone [55**]. This integration should be carried out with particular care, as NMR averaging depends on the time scale of interconversion among different states. Therefore, it is essential to estimate kinetic properties on the cryo-EM-derived structural ensemble before properly incorporating ensemble-averaged NMR data.

A representative example of the opportunities offered by integrative approaches is the recent characterization of the structure and dynamics of the ClpP protease by a combined cryo-EM and NMR strategy [68]. In this study, the EMMI approach has been used to generate a
conformational ensemble (Figure 3d) using cryo-EM data at 3.6 Å resolution (Figure 3c). This ensemble highlighted the presence of significant dynamics in the N-terminal gating region of the protease (Figure 3e), which was resolved at lower resolution by cryo-EM (Figure 3c). NMR methyl chemical shifts and paramagnetic relaxation enhancement data were back-calculated from the cryo-EM ensemble and compared to the experimental measurements, showing a good agreement. This example can be regarded as an initial step toward the integration of cryo-EM with NMR data to generate accurate structural ensembles that are revealing of the dynamics, and ultimately function, of a complex biological system.

Finally, MD simulations can complement the characterization of the conformational landscape of a system when multiple cryo-EM maps are first used to model individual structures representing local free-energy minima. By connecting these landmarks using standard or enhanced-sampling MD simulations, one can obtain a more complete overview of the conformational landscape of complex biological systems [78].

Community challenges in the determination, validation and dissemination of protein structural ensembles

The recent developments in the cryo-EM field, along with the new advances in integrative modelling approaches for protein structural ensemble determination, are exposing new challenges that the community should collectively strive to solve.

1) Researchers should be able to readily access data and information about specific experiments, such as the raw single-particle images and details about the post-processing of these data into 3D density maps. This information might be required to use, or further develop, specific integrative approaches [55*,56]. An important initial step in this direction is the Electron Microscopy Public Image Archive (EMPIAR) for raw, 2D electron microscopy images (https://www.ebi.ac.uk/pdbe/emdb/empiar/) [79].

2) Modelling procedures and protocols for structural ensemble determination should be shared within the community in order to encourage reproducibility, which can be further facilitated by using well-established approaches for structural ensemble comparison [80]. Furthermore, the resulting structural ensembles should be easily accessible. In this regard, a pioneering effort is the creation of the PDB-Dev prototype repository for integrative structural models (https://pdb-dev.wwpdb.org/) [81], which hosts multi-scale, multi-state and time-ordered ensembles of macromolecular assemblies along with the data used in their modelling. This information should include structural models, possibly after clustering integrative simulations into a set of discrete states using community-established criteria, states populations and transition rates (Figure 1).

3) Effective structural ensemble validation procedures should be established and progressively enforced when depositing structural ensembles to public repositories. These procedures should assess both the protocols used to generate the structural ensembles and their overall quality. First, exhaustive sampling of the conformational landscape, and therefore convergence of integrative simulations, should be monitored by either reporting the statistical error in the estimated state populations or more sophisticated approaches [82]. Second, structural ensembles should be evaluated in terms of their overall physico-chemical quality, as typically quantified by the Molprobity score [83] calculated by the PHENIX [84] validation suite, as well as the fit of the cryo-EM data, often measured in terms of cross-correlation between experimental and model maps or using the EMRinger [85] approach to assess the precise fitting of an atomic model into a cryo-EM map. These validation tools and the corresponding criteria used to establish when single-structure models are acceptable for deposition should therefore be adapted to deal with protein ensembles.

Additional validations of cryo-EM and integrative structural ensembles can be obtained by assessing the consistency between predictions calculated from the ensembles and independent experimental data not used in the modelling, such as NMR and SAXS data. Such comparison must be done with caution and upon considering: first, the kinetics of interconversion among different structural states and thus how experimental observables are averaged over the ensemble; second, the errors in the predictors of experimental observables as well as in the experimental data; third, the statistical uncertainty in the population of the states determined by integrative simulations; fourth, the fact that cryo-EM and NMR/SAXS data are collected in different conditions, for example vitreous ice and solution at room temperature, implies that some differences between the underlying ensembles can be expected.

More extensive discussions of these challenges can be found in a recent Special Issue of the Journal of Structural Biology (http://challenges.emdataresource.org/?q=node/84).

Conclusions

Cryo-EM has become in recent years increasingly successful in structural biology because of its ability to provide accurate characterizations of the conformational landscapes of complex biological systems. Alongside these developments, we have assisted to the flourishing of integrative modelling approaches to integrate cryo-EM with other sources of experimental and prior information [17**,18*,19–22] These methods further increase the
accuracy of the structural ensembles generated using cryo-EM data. As we have emphasised here, however, this goal can only be achieved if the challenges related to the accuracy of the hybrid energy function used, the treatment of all sources of errors and uncertainty, and the sampling issues of MD and MC simulations are successfully addressed (Figure 2).

Over the coming years we will assist to further technical developments in the cryo-EM field as well as in the data analysis and integrative modelling techniques, which will facilitate the synergistic use of all the available information to characterize structure and dynamics, and ultimately function, of ever more complex and important biological systems. Therefore, it is now the right time for the entire community to discuss how the determination and validation methods, as well as dissemination protocols, developed over the years for single-structure models should be revisited and updated for the era of protein structural ensembles.

Acknowledgements

We are grateful to Alessandro Barducci and Francesco Apriile for their feedback on Figure 1, to Giulia Vecchi for preparing Figure 2, to Zev Ripstein for providing the ClpP cryo-EM data visualized in panels A, B, and C of Figure 3, and to James Fraser for carefully reading the manuscript.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

• of special interest
•• of outstanding interest


A review of the principles and methods to model protein structural ensembles, with focus on the classification between maximum entropy principle and maximum parsimony approaches.


A review of the latest advances in developing methods to enhance sampling of the conformational landscape of biomolecules, more accurate force fields, and approaches to integrate experimental data into molecular simulations.


A systematic study that is very helpful in establishing the state of the art in terms of force fields for disordered proteins.


A metainference-based approach to model protein structural ensembles based on cryo-EM density maps and other sources of information, with application to the integral membrane receptor STRA6.


A collection of methods to use ensemble-averaged, noisy data to model ensemble of conformations, including a reweighting approach to optimize the weights of previously determined ensembles to fit the data and a method that uses data directly as restraints in a replica-averaged scheme.


