Simultaneous Determination of Protein Structure and Dynamics Using Cryo-Electron Microscopy
Massimiliano Bonomi, Riccardo Pellarin, Michele Vendruscolo

To cite this version:
Massimiliano Bonomi, Riccardo Pellarin, Michele Vendruscolo. Simultaneous Determination of Protein Structure and Dynamics Using Cryo-Electron Microscopy. Biophysical Journal, Biophysical Society, 2018, 114 (7), pp.1604-1613. 10.1016/j.bpj.2018.02.028. pasteur-02404467

HAL Id: pasteur-02404467
https://hal-pasteur.archives-ouvertes.fr/pasteur-02404467
Submitted on 11 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Simultaneous Determination of Protein Structure and Dynamics Using Cryo-Electron Microscopy

Massimiliano Bonomi,1,* Riccardo Pellarin,2 and Michele Vendruscolo1,*
1Department of Chemistry, University of Cambridge, Cambridge, United Kingdom and 2Structural Bioinformatics Unit, Institut Pasteur, CNRS UMR 3528, Paris, France

ABSTRACT Cryo-electron microscopy is rapidly emerging as a powerful technique to determine the structures of complex macromolecular systems elusive to other techniques. Because many of these systems are highly dynamical, characterizing their movements is also a crucial step to unravel their biological functions. To achieve this goal, we report an integrative modeling approach to simultaneously determine structure and dynamics of macromolecular systems from cryo-electron microscopy density maps. By quantifying the level of noise in the data and dealing with their ensemble-averaged nature, this approach enables the integration of multiple sources of information to model ensembles of structures and infer their populations. We illustrate the method by characterizing structure and dynamics of the integral membrane receptor STRA6, thus providing insights into the mechanisms by which it interacts with retinol binding protein and translocates retinol across the membrane.

INTRODUCTION

Cryo-electron microscopy (cryo-EM) (1–9) is a powerful structural biology technique that enables the characterization of complex biological systems that are not readily amenable to other traditional techniques, such as x-ray crystallography and NMR spectroscopy. With the continuous progress in the development of both instrumentation and image processing software, cryo-EM is rapidly approaching the resolution of x-ray crystallography (2–9), allowing the structures of complexes or individual proteins of great biological relevance to be determined in their natural environments at nearly atomistic resolution (10–17).

Despite these great advances, a major challenge remains open—the characterization of the dynamics of the systems observed (18,19). This is a crucial problem, because most macromolecular machines populate multiple conformational states, and their functions depend on the interplay between structure and dynamics. In several cases, cryo-EM can detect alternative conformations (3–6,15,16), provided that two-dimensional images of particles with distinct conformations are separately classified (5,6,20). However, if the system has highly dynamic regions or the low resolution of two-dimensional images hinders classification, three-dimensional reconstructions generated by combining multiple class-averages present regions at lower resolution. These regions may correspond to flexible parts of the system, which are averaged out in the class-averaging and reconstruction process, or to particularly noisy portions of the map caused, for example, by radiation damage.

Several modeling approaches are currently used to translate cryo-EM data into structural models (21,22). Following the classification proposed in (22), these approaches include methods for rigid-body fitting (Chimera (23), COAN (24), EMfit (25), MODELLER (26), MultiFit (27), SITUS (28)), flexible fitting (EMFF (29), MDFF (30), MODELLER (26), SITUS (28), MDFIT (31), Flex-EM (32)), homology modeling (Fold-EM (33), ROSETTA (34), MODELLER (26)), de novo modeling (EM-fold (35), SITUS (28), IMP (36), RELION (20), Phenix (37)), and integrative approaches (IMP (36)). All these approaches search for single structural models that minimize the deviation between observed and predicted cryo-EM density maps, usually by incorporating additional restraints into Monte Carlo or molecular dynamics (MD) simulations. Some of these techniques can provide multiple alternative models that individually fit the input map to a certain extent (24,38), in the same way as multiple molecular models are derived from NMR spectroscopy data (39), and routinely deposited in the PDB database. The sets of these models can be considered as uncertainty ensembles (40), because they reflect the limited information available on the systems and thus the fact that different models might be equally consistent with the input data; they do not, however, reflect...
the conformational heterogeneity arising from the internal dynamics of the systems themselves (40–44). In many cases, as for instance in inferential structure determination (45), weights can be associated to the members of these uncertainty ensembles. These weights quantify the consistency with the input information and thus the overall confidence in each individual model (46); they are not, however, statistical weights, because they are not meant to represent the equilibrium populations of different conformations present in a conformationally heterogeneous mixture.

The methods described above can successfully determine single-structure models as well as uncertainty ensembles from cryo-EM data, but they do not provide fully quantitative information about the thermodynamics and dynamics of the systems studied. Such information can only be extracted from thermodynamic ensembles representing sets of conformations, together with their statistical weights, which coexist in a heterogeneous mixture along with their equilibrium populations (40,44,47). To obtain a thermodynamic ensemble, one should search for structural models whose ensemble-averaged observables, rather than individual conformations, fit the input data (40,44,47). Importantly, in the construction of such thermodynamic ensembles, all possible sources of errors should be taken into account when evaluating the fit of the ensemble with the experimental data (46,48).

Here, we report an approach to enable the simultaneous determination of structure and dynamics of proteins and protein complexes by modeling thermodynamic ensembles from cryo-EM density maps. This approach is based on the recently introduced meta-inference (46), a general Bayesian inference method (45) that enables the modeling of heterogeneous systems from noisy, ensemble-averaged experimental data. This method incorporates a Bayesian treatment of cryo-EM data that accounts for variable levels of noise in the maps and enables structural modeling at atomistic resolution (49).

**MATERIALS AND METHODS**

**Overview of the meta-inference approach**

Meta-inference (46) is a Bayesian framework (45) for modeling thermodynamic ensembles by integrating prior information on a system (physical, chemical, or statistical knowledge) with noisy experimental data. The method is designed to deal with conformationally heterogeneous systems, in which experimental observations reflect an ensemble of states rather than a single conformation, and with data affected by known and unknown errors. The meta-inference approach enables: 1) optimally combining and weighting multiple sources of information, 2) modeling thermodynamic ensembles, 3) determining the population of all states in the ensemble, and 4) determining the level of noise in the input data. In doing so, meta-inference allows overcoming the limitations of individual computational and experimental techniques (40). Standard MD simulations, which are hampered by inaccuracies in the empirical physico-chemical description of the system (the force field), are complemented by experimental data, which alone provide only sparse information affected by random and systematic errors.

In meta-inference, the generation of models is guided by the meta-inference energy function, defined as $E_{\text{MI}} = -k_B T \log p_{\text{MI}}$, where $k_B$ is the Boltzmann constant, $T$ is the temperature of the system, and $p_{\text{MI}}$ is the meta-inference posterior probability. The posterior expresses the probability of observing a given set of structural models, and possibly other parameters, in terms of prior information and data likelihood. The former encodes physico-chemical knowledge of the system; the latter quantifies the agreement of the experimental data and incorporates a model of experimental noise.

In the case of Gaussian noise, the meta-inference posterior of observing a set of $N$ conformations $X = \{X_i\}$ given $N_d$ independent experimental data points $D = \{d_i\}$ is

$$p_{\text{MI}}(X, \sigma^{\text{SEM}}, \sigma^B | D) = \prod_{r=1}^{N} p(X_r) \prod_{i=1}^{N_d} \frac{1}{\sqrt{2\pi}(\sigma^B_{i})^2 + (\sigma^{\text{SEM}}_{i})^2} \exp\left\{ -\frac{1}{2} \frac{[d_i - f_i(X)]^2}{(\sigma^B_{i})^2 + (\sigma^{\text{SEM}}_{i})^2} \right\} p(\sigma^B_{i}),$$

where $f_i(X) = (1/N) \sum_{j=1}^{N} f_j(X_j)$ is the average of the predictor (or forward model) $f_j$ for the experimental observable $d_i$ calculated over the set of $N$ conformations, $\sigma^B_{i}$ is a parameter quantifying the noise level in data point $d_i$, and $\sigma^{\text{SEM}}_{i}$ is the statistical error in calculating an ensemble average of $f_j$ using a finite set of conformations. The parameters $p(\sigma^B_{i})$ and $p(X_r)$ are the priors on $\sigma^B_{i}$ and $X_r$, respectively.

A sample of the posterior distribution is typically obtained by running a multiple-replica simulation (50) guided by the associated meta-inference energy function

$$E_{\text{MI}} = E_{\text{MD}} + k_B T \frac{1}{2} \sum_{i,j} \frac{[d_i - f_i(X)]^2}{(\sigma^B_{i})^2 + (\sigma^{\text{SEM}}_{i})^2} + E_a,$$

in which the force field of standard MD simulations, $E_{\text{MD}} = -k_B T \sum_{i=1}^{N} \log p(X_i)$, is modified by 1) a series of (harmonic) data restraints that ensure the agreement of the structural ensemble with the experimental data and 2) a series of error restraints, $E_a = k_B T \sum_{i,j} - \log p(\sigma^B_{i}) + 0.5 \log[(\sigma^B_{i})^2 + (\sigma^{\text{SEM}}_{i})^2]$. In this multiplica simulation scheme, one needs to sample, for each replica, not only the space of possible conformations $X_r$, but also of the parameters $\sigma^B_{i}$ that quantify the level of noise in the data. This is typically achieved by a Gibbs sampling scheme, which combines MD to sample the coordinates space with Monte Carlo for the noise parameters $\sigma^B_{i}$. The values of these noise parameters ultimately determine the intensities of the harmonic data restraints: low noise will result in a strong structural restraint; inconsistent data points and outliers will be automatically labeled as noisy and down-weighted in the construction of the final ensemble.

**The meta-inference approach for cryo-EM density maps**

In the following, we define the meta-inference components, previously introduced in general terms, specifically for the case of cryo-EM data. The
development of these elements builds on the approach proposed in (49). A summary of all the elements at the basis of our approach is reported in Table S1. The method is implemented in the PLUMED-ISDB module (51) of the open-source, freely available PLUMED library (http://www.plumed.org) (52).

Experimental data. Typically, a cryo-EM density is distributed as a map defined on a grid, or set of voxels, in real space. In our approach, we will represent the experimental density map using a Gaussian mixture model (GMM) \( \phi_D \) with \( N_D \) components (data-GMM):

\[
\phi_D(x) = \sum_{i=1}^{N_D} \phi_{D,i}(x) = \sum_{i=1}^{N_D} \omega_{D,i} G(x \mid x_{D,i}, \Sigma_{D,i}),
\]

where \( \omega_{D,i} \) is the (normalized) weight of the \( i \)th component of the data GMM and \( G \) is a normalized Gaussian function centered in \( x_{D,i} \) and with covariance matrix \( \Sigma_{D,i} \). This representation has three advantages. First, it is computationally convenient to use an analytical representation of the input data, rather than a discrete definition on a grid. Second, a GMM can provide a representation of the data in terms of independent bits of information, whereas in the grid representation neighboring voxels should be considered as correlated data points affected by correlated noise. Third and finally, a GMM can be tuned to represent the data at different resolutions, from coarse-grained for initial efficient modeling or for low-resolution maps, to atomistic for refinement of high-resolution maps. To efficiently fit high-resolution maps at near-atomic detail, we used a divide-and-conquer approach (49), which starts from a low-resolution fit using few Gaussians and refines it in subsequent iterations to increase the resolution of the final GMM (Supporting Material).

The forward model. We developed a forward model to simulate a cryo-EM map from a structural model. As for the representation of the experimental map, the forward model \( \phi_M \) is a GMM. Because here we employed high-resolution synthetic and real cryo-EM maps, we represented each heavy atom of the system by one Gaussian function, whose parameters were obtained by fitting the electron atomic scattering factors (53) for a given atomic species (Table S2). In the case of low-resolution maps, a single Gaussian can be used to represent each coarse-grained bead, with the Gaussian width proportional to the size of the bead (49).

The noise model. The deviation between predicted and observed density maps is measured in terms of the overlap \( \psi_{MD,j} \) between the forward model GMM \( \phi_M \) and the \( j \)th component \( \phi_{D,j} \) of the data-GMM,

\[
\psi_{MD,j} = \int dx \phi_M(x) \phi_{D,j}(x).
\]

The overlap \( \psi_{MD,j} \) can be expressed in a computationally convenient analytical form (49), with \( \phi_{D,j}(x) \) as a Gaussian function and \( \phi_M(x) \) as a GMM. In a heterogenous system, the forward model \( \phi_M \) is an average over the \( N \) metainference replicas \( \phi_M^r \), and thus the overlap can be written as

\[
\int dx \left( \frac{1}{N} \sum_{r=1}^{N} \phi_M^r(x) \right) \phi_{D,j}(x) = \frac{1}{N} \sum_{r=1}^{N} \int dx \phi_M^r(x) \phi_{D,j}(x)
\]

\[
= \frac{1}{N} \sum_{r=1}^{N} \psi_{MD,j}^r = \bar{\psi}_{MD,j}.
\]

For each component of the data-GMM, we used a Gaussian noise model with one uncertainty parameter per data point to account for a variable level of noise across the map. The data likelihood for the overlap \( \psi_{DD,i} \) of the \( i \)th component of the data-GMM with the entire data-GMM can then be written as

\[
p\left(\psi_{DD,i} \mid \mathbf{X}, \sigma_{\epsilon,i}, \sigma_{\epsilon,i}^2\right) = \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{1}{2} \frac{\left(\psi_{DD,i} - \bar{\psi}_{MD,i}\right)^2}{\sigma_{\epsilon,i}^2 + \sigma_{\epsilon,i}^2}\right].
\]

Prior information. As structural prior information \( p(X_i) \), we used the AMBER99SB*-ILDN molecular mechanics force field (54) and the GB/S implicit model of water (55). For the uncertainty parameters, we used an uninformative Jeffreys prior (56). To avoid sampling all the uncertainty parameters, we marginalized them before simulating the system (Supporting Material). The level of noise in each component of the data GMM can then be estimated a posteriori using all the structural models produced by the metainference simulations (Supporting Material).

Meta-inference energy function. After defining the noise model as outlined in the previous paragraphs and marginalizing all the noise parameters, the meta-inference energy function for cryo-EM data becomes

\[
E_{MD} = E_{MD} - k_B T \sum_{i,j} \log \left[ \frac{1}{2(\psi_{DD,i} - \bar{\psi}_{MD,i})} \times \text{erf} \left( \frac{\psi_{DD,i} - \bar{\psi}_{MD,i}}{\sqrt{2} \sigma_{\epsilon,i}} \right) \right],
\]

where erf(\(x\)) is the error function.

General details of the metainference simulations

All simulations were performed using as prior the AMBER99SB*-ILDN force field (54) along with the GB/S implicit model of water (55). This is a computationally convenient combination of force fields, although its accuracy has been shown to be modestly inferior compared to AMBER99SB*-ILDN combined with explicit solvents models (57). Starting models were equilibrated at 300 K for 1 ns. A time step of 2 fs was used together with LINCS constraints (58). The van der Waals and Coulomb interactions were cut off at 2.0 nm. Neighbor lists for all long-range interactions were cut off at 2.0 nm and updated every 10 MD steps. Simulations were carried out using a nonperiodic cell in the canonical ensemble at 300 K, enforced by the Bussi-Donadio-Parinello thermostat (59). Configurations were saved every 2 ps for postprocessing. To improve computational efficiency, the cryo-EM restraint was calculated every 2 MD steps (60), using neighbor lists to compute the overlaps between model and data GMMS, with cutoff equal to 0.01 and update frequency of 100 steps. Well-tempered metadynamics (61) was used to accelerate sampling of the metainference ensemble (Supporting Material). All simulations were carried out with GROMACS 4.5.7 (62) and PLUMED (52). Parameters of the GB/S implicit solvent were imported from GROMACS 5.1.4.

GroEL metainference simulations

The crystal structures of apo GroEL (PDB: 1XCK) (63) and GroEL-ADP complex in the relaxed allosteric state (PDB: 4K18) (64) were used to generate a synthetic cryo-EM map, using the following procedure. Chains A were extracted from the two PDBs (Fig. 1 A) and aligned using UCSF Chimera (23). MODELLER v9.17 (26) was used to generate a
A divide-and-conquer approach (49) was used to fit a GMM with 4000 components, which resulted in a cross correlation with the original map equal to 0.97 (Fig. S2). The cytosolic loop (residues 575–597) missing from the apo-GroEL model. The metainference ensemble was simulated using 16 replicas for a total aggregated time of 355 ns per production run. σ^NSE was set to 0.1 · σ^obs. To enhance sampling, we used well-tempered metadynamics with W = 950,000.

**STRA6 metainference simulations**

The cryo-EM map of the complex of zebrafish STRA6 with copurified calmodulin at 3.9 Å resolution (EMD: 8315 (66)) was fit with a GMM using a divide-and-conquer approach (49). The final GMM was composed of 11,585 Gaussians and resulted in a cross correlation with the original map equal to 0.97 (Fig. S2). The cytosolic loop (residues 575–597) missing from the deposited model (PDB: 5SY1) was modeled using the software MODELLER v9.17. The residue numbering scheme was kept as in the deposited model. The resulting comparative model was then equilibrated at 300 K, as previously described. Initial models for two independent production runs (labeled as “RUN I” and “RUN II”) were then randomly extracted from the equilibration run. The metainference ensemble was simulated using 16 replicas for a total aggregated time of 355 ns per production run. σ^NSE was set to 0.1 · σ^obs. To enhance sampling, we used well-tempered metadynamics with W = 1000 kJoul/mol and γ = 150,000.

**Comparison with the STRA6 experimental cryo-EM map**

To evaluate the quality of the fit of our metainference ensemble with the experimental map and compare it with the deposited model, we used the gmconvert utility (65) to calculate synthetic cryo-EM maps from structural models. It is important to note that the algorithm implemented in gmconvert is different from our forward model and from the approach implemented in RELION (20) and used in (66) to refine the deposited model. In this way, gmconvert provides a method to predict a cryo-EM map independent from those used in the generation of our ensemble and in the refinement of the deposited model. We believe that this is a fair procedure to evaluate the agreement with the experimental map. The local cross correlation (CC) with the experimental map (EMD: 8315) of the average map computed on our metainference ensemble and the one computed on the deposited model

**Stereochemistry assessment of the STRA6 ensemble**

To measure the stereochemical quality of the ensemble of STRA6 models generated by the metainference method, we calculated the distribution of the backbone dihedral angles φ and ψ on the conformations sampled in the two independent simulations. To achieve this, we used the program PROCHECK (67), specifically the procheck_nmr collection of codes designed to evaluate the quality of NMR ensembles. This program classifies all residues in all models in four regions of the Ramachandran plot (Fig. S3 A): residues in most favored regions (red), in additionally allowed regions (yellow), in generously allowed regions (light yellow), and in disallowed regions (white). The percentages of residues in each of these regions for the two independent metainference simulations were, in both cases: 87.4, 11.6, 0.5, and 0.5% (Fig. S3 C). These values were comparable to those obtained using the STRA6 deposited model (PDB: 5SY1): 86.4, 13.3, 0.3, and 0.0% (Fig. S3 B).

**FIGURE 1** Validation of the metainference approach on a conformationally heterogeneous ensemble of the chaperonin GroEL. The crystal structure of Apo-GroEL (63) (A, blue) and a comparative model built from the structure of GroEL in complex with ADP (64) (A, red) were used to create synthetic cryo-EM maps at near-atomistic resolution (B). An average map was then computed by mixing contributions from the two models in ratio 1:1 (C). The metainference approach was capable of disentangling the contribution of the two states (D), determining their relative populations in the mixture, and inferring the local level of noise in the map (E). To see this figure in color, go online.
was evaluated in a five-voxel sliding window using the “vop” localcorrelation command in UCSF Chimera (23). The same program was then used to color the metainference maps and deposited-model maps based on the value of the local CC (Figs. 2, B and D and S4, B and D).

**RESULTS AND DISCUSSION**

**Validation with GroEL data**

We first assessed the accuracy of the metainference approach by applying it to the case of the chaperonin GroEL in two different conformations (Fig. 1 A): apo-GroEL, the compact crystal structure of apo GroEL (63), and GroEL-ADP*, a comparative model built from an extended allosteric state adopted by GroEL in complex with ADP (64). We created synthetic cryo-EM maps from these two models (Fig. 1 B), and used them to construct an average cryo-EM map that equally mixed contributions from the two states (Fig. 1 C), which have a backbone RMSD of 3.9 Å (Fig. S5). As expected, the metainference ensemble was characterized by the presence of two distinct, equally populated structural clusters (Fig. 1 D), which corresponded to apo-GroEL and GroEL-ADP*, with backbone RMSD of the centers of the two clusters from the correspondent model equal to 2.0 Å (Fig. S5). Our approach accounts for unknown and variable levels of noise across the experimental map during the modeling of the ensemble. In this case, the only source of error, i.e., of deviation between predicted and observed maps, is the difference between the procedures used to generate the synthetic map and in our modeling approach. The inferred level of relative noise was fairly uniform across the map, with an average value of ~0.003 (Figs. 1 E and S6).

**Structure and dynamics of the STRA6 membrane receptor**

We then applied our metainference approach to the integral membrane receptor STRA6 (66), which mediates the cellular uptake of retinol by extracting it from its carrier...
(retinol binding protein, or RBP) and moving it across the membrane (69). Recently, zebrafish STRA6 was determined at 3.9 Å resolution by single-particle cryo-EM (EMD: 8315) (66). The structural model obtained from the cryo-EM map (PDB: 5SY1) revealed a dimer of STRA6 in complex with the protein calmodulin, and enabled the characterization of the regions involved in RBP binding and retinol translocation across the membrane (66). In this work we focus in particular on the regions of the STRA6 map that were determined at lower resolution (66), including the periphery of the complex and the RBP binding region, to show that they can be interpreted in terms of a combination of conformational dynamics and noise in the data.

The STRA6 metainference ensemble

Starting from the 5SY1 structure (Fig. 2 A), we modeled a thermodynamic ensemble of conformations by integrating cryo-EM data (Fig. S2) with an a priori, physico-chemical knowledge of the system (Materials and Methods). In the generated ensemble (Figs. 2 C and S4 C), STRA6 presents a significant degree of flexibility around the single-structure model, in particular in the N-terminal and C-terminal domains, the lid peptide (LP) and the RBP-binding motif, the cytosolic loop, and the JM helix (Fig. S7 A). These regions correspond to areas at lower resolution in the experimental map (66), as well as in the back-calculated maps (Figs. 2, B and D and S4, B and D). The cytosolic loop, which is not included in the PDB: 5SY1 structure, displays the largest fluctuations (Fig. S7 A, magenta), whereas calmodulin (Fig. S7 B) and the STRA6 trans-membrane (TM) domain (Fig. S7 A, gray) are instead more rigid and deviate less from the PDB: 5SY1 model (Fig. S7, C and D). The latter result is particularly relevant, given the low accuracy of the prior in the TM region (Materials and Methods).

We then measured the agreement of the metainference ensemble with the experimental data by calculating the maps predicted from the ensemble and the single-structure model (Supporting Material). We found that the metainference ensemble provided a better cross correlation with the experimental map (global CC = 0.91) than the single-structure model (global CC = 0.86), especially in the more dynamical regions of STRA6 (Figs. 2, B and D and S4, B and D). We also verified that the improved agreement with the experimental data was not achieved at the expense of the stereochemical quality of the models (Fig. S3).

Noise inference

To quantify the level of noise in the data, we calculated an error density map from the metainference ensemble (Supporting Material) and visualized it on the experimental map (Figs. 2 E and S4 E). The inferred level of relative error was fairly uniform, with an average value of ~0.38 (Fig. S8), except for a few specific regions: the binding sites in-between the two horizontal intramembrane helices (IM), the interior of the outer cleft, and the external region.
surrounding the TM domain (Figs. 2 E and S4 E). It has been suggested (66) that all these regions are occupied by components that were not explicitly modeled, including cholesterol in the binding sites and in the outer cleft, and amphipols in the shell around the TM domain. Consistent with the conclusion that these electron densities could not be explained by the presence of STRA6 alone, the metainference method resulted in the assignment of a high level of noise to these regions.

**Dynamics in the RBP-binding region**

Next, we investigated how the conformational dynamics obtained by the metainference approach can be linked to the biological function of STRA6 and thus to what extent the resulting ensemble can be more informative and predictive than a single-structure model. The observed flexibility in the N-terminal domain might be functional, as this region might act as a sensor for unidentified ligands (66) or to facilitate RBP recognition and recruiting through nonspecific interactions, before binding specifically. We found that the LP region (Fig. 3 A) is characterized by the presence of an equilibrium among different conformations (Fig. 3 B). The state in which the two LPs from the two STRA6 monomers are close together (LP1, Fig. 3 C), as in the PDB model, appear to have a relatively low population. In the most populated state, the two LPs fold back and approach the loop region between TM8 and TM9 (LP2, Fig. 3 D). States in which only one of the two LPs folds back were also present (LP3 and LP4, Fig. 3 B). As these results are consistent with the possibility that LP2, rather than LP1, may be more productive for RBP binding, incorporating further information about the physico-chemical environment in the surroundings of the LP region will improve their prior description and offer a more accurate quantification of their relative stabilities. The existence of the LP2 state, not directly visible in the single-structure model, could explain the fact that inserting a Myc tag at the apex of the TM8-TM9 loop impairs RBP binding (70). This effect would be the result of altering the equilibrium among LP states in favor of either a conformation not productive for binding or one destabilizing the actual binding state, depending upon whether LP2 is the actual binding conformation.

**Translocation of retinol across the membrane**

Concerning the exit mechanism of retinol from STRA6 into the cytosol (Fig. 4, A and B), the single-structure model suggests a lateral pathway, as the alternative pathway from the central dimer interface would require significant conformational changes (66). From our study of the conformational dynamics of STRA6, we identify a potential role of the JM helix in regulating retinol release through either of the two pathways. This fragment populates multiple distinct conformations in our ensemble (Fig. 4 C). In one state (JM1, Fig. 4 D), JM points outwards from STRA6, as in
the single-structure model. In the second (JM2, Fig. 4 E), this peptide interacts with the JML, which is situated below the horizontal IM helices and the putative retinol binding site (Fig. 4 B). In another state (JM3, Fig. 4 F), JM resides in proximity to TM7. This complex equilibrium among states suggests that JM can transiently interact with JML and possibly compete with the formation of the conserved salt bridge D539-R511 between JML and IM. This salt bridge appears to be crucial for consolidating the JML-IM interaction and stabilizing the retinol binding site located in-between the IM helices, as its disruption by mutation in human STRA6 results in Matthew-Wood syndrome (71). By competing with the salt-bridge formation, JM could promote IM and JML mobility, weaken the stability of the retinol binding site, and eventually favor retinol unbinding. Translocation of retinol across the membrane can later occur through either the lateral or central pathways. The latter scenario would require additional conformational changes, which might be facilitated by the transient disruption of the salt bridge and the increased mobility of this region.

This dynamical picture of JM offers interesting perspectives, regardless of the specific retinol exit pathway. It could rationalize why mutations in TM6 and TM7 inhibit retinol uptake (72), as they could shift the equilibrium toward a state (JM3) in which JM is close to TM7 and cannot destabilize the IM-JML interaction to favor retinol unbinding. Furthermore, JM is adjacent in sequence to CamBP0 (66), one of the STRA6 helices that directly interact with calmodulin, suggesting a possible role of calmodulin in altering the equilibrium among JM states, and ultimately regulating retinol uptake. This observation is particularly intriguing, as the role of calmodulin still remains enigmatic, with no direct link to retinol transport being identified so far.

CONCLUSIONS

We have reported a method to determine structure and dynamics of macromolecular systems by modeling thermodynamic ensembles from cryo-EM density maps. The application to the integral membrane receptor STRA6 has illustrated how functional dynamics might remain hidden in areas of the map at lower resolution. The method is capable of revealing such dynamics by disentangling the effect of noise in the maps from conformational heterogeneity. This method is implemented in the PLUMED-ISDB module (51) of the open-source PLUMED library (http://www.plumed.org) (52), allowing the integration of cryo-EM with other ensemble-averaged experimental data, thus readily enabling integrative structural and dynamical biology studies (18,19). The approach can be extended to model multiple ensembles using three-dimensional reconstructions obtained from different two-dimensional class-averages and to thoroughly characterize the conformational landscape, dynamics, and function of complex biological systems.

SUPPORTING MATERIAL

Supporting Materials and Methods, ten figures, and two tables are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(18)30288-1.

AUTHOR CONTRIBUTIONS

M.B. and R.P. performed research. All authors contributed to developing the methodology, analyzing the data, and writing the manuscript.

ACKNOWLEDGMENTS

The metainference simulations were performed on the Sisu supercomputer managed by the CSC - IT Center for Science (Finland) under the DECI-14 project “EMML.”

REFERENCES

1. Henderson, R. 1995. The potential and limitations of neutrons, electrons and x-rays for atomic resolution microscopy of unstained biological molecules. Q. Rev. Biophys. 28:171–193.
2. Kühlbrandt, W. 2014. Biochemistry. The resolution revolution. Science. 343:1443–1444.
3. Bai, X. C., G. McMullan, and S. H. Scheres. 2015. How cryo-EM is revolutionizing structural biology. Trends Biochem. Sci. 40:49–57.
4. Callaway, E. 2015. The revolution will not be crystallized: a new method sweeps through structural biology. Nature. 525:172–174.
5. Nogales, E. 2016. The development of cryo-EM into a mainstream structural biology technique. Nat. Methods. 13:24–27.
6. Glaeser, R. M. 2016. How good can cryo-EM become? Nat. Methods. 13:28–32.
7. Brilot, A. F., J. Z. Chen, ..., N. Grigorieff. 2012. Beam-induced motion of vitrified specimen on holey carbon film. J. Struct. Biol. 177:630–637.
8. Carroni, M., and H. R. Saibil. 2016. Cryo electron microscopy to determine the structure of macromolecular complexes. Methods. 95:78–85.
9. Li, X., P. Mooney, ..., Y. Cheng. 2013. Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. Nat. Methods. 10:584–590.
10. Liao, M., E. Cao, ..., Y. Cheng. 2013. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. Nature. 504:107–112.
11. Bai, X. C., C. Yan, ..., Y. Shi. 2015. An atomic structure of human γ-secretase. Nature. 525:212–217.
12. Campbell, M. G., D. Veesler, ..., B. Carragher. 2015. 2.9 Å resolution reconstruction of the Thermoplasma acidophilum 20S proteasome using cryo-electron microscopy. eLife. 4.
13. Bartesaghi, A., A. Merk, ..., S. Subramaniam. 2015. 2.2 Å resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor. Science. 348:1147–1151.
14. Fernández, I. S., X. C. Bai, ..., S. H. W. Scheres. 2013. Molecular architecture of a eukaryotic translational initiation complex. Science. 342:1240–305.
15. Zhao, J., S. Benlekhbir, and J. L. Rubinstein. 2015. Electron cryo-microscopy observation of rotational states in a eukaryotic V-ATPase. Nature. 521:241–245.
16. Cianfrocco, M. A., G. A. Kassavetis, ..., E. Nogales. 2013. Human TFIID binds to core promoter DNA in a reorganized structural state. Cell. 152:120–131.
17. Fitzpatrick, A. W. P., B. Falcon, ..., S. H. W. Scheres. 2017. Cryo-EM structures of tau filaments from Alzheimer’s disease. Nature. 547:185–190.
31. Ratje, A. H., J. Loerke, 32. Topf, M., K. Lasker, 30. Trabuco, L. G., E. Villa, 27. Lasker, K., M. Topf, 37. Adams, P. D., P. V. Afonine, 39. Brussi, G., M. Bonomi et al.

25. Rossmann, M. G., R. Bernal, and S. V. Pletnev. 2001. Combining electron microscopy data. WIREs Comput. Mol. Sci. 5:62–81.

22. Lopez-Blanco, J. R., and P. Chacon. 2015. Structural modeling from electron microscopy data. WIREs Comput. Mol. Sci. 5:62–81.

21. Schroeder, G. F. 2015. Hybrid methods for macromolecular structure determination: experiment with expectations. Curr. Opin. Struct. Biol. 31:20–27.

23. Pettersen, E. F., T. D. Goddard, 24. Caffrey, J., D. M. Honig. 2000. Force-field free energy perturbation and free energy simulation. Curr. Opin. Struct. Biol. 10:e1001244.

24. Volkmann, N., and D. Hanei. 1999. Quantitative fitting of atomic models into observed densities derived by electron microscopy. J. Struct. Biol. 125:176–184.

25. Rossmann, M. G., R. Bernal, and S. V. Pletnev. 2001. Combining electron microscopy with x-ray crystallographic structures. J. Struct. Biol. 136:190–200.

26. Salì, A., and T. L. Bundell. 1993. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234:779–815.

27. Lasker, K., M. Topf, …, H. J. Wolfson. 2009. Inferential optimization for simultaneous fitting of multiple components into a CryoEM map of their assembly. J. Mol. Biol. 388:180–194.

28. Wriggers, W. 2012. Conventions and workflows for using SITUS. Acta Crystallogr. D Biol. Crystallogr. 68:344–351.

29. Zheng, W. 2011. Accurate flexible fitting of high-resolution protein structures into cryo-electron microscopy maps using coarse-grained pseudo-energy minimization. Biophys. J. 100:478–488.

30. Trabuco, L. G., E. Villa, …, K. Schulten. 2008. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. Structure. 16:673–683.

31. Ratej, A. H., J. Loerke, …, C. M. Spahn. 2010. Head swivel on the ribosome facilitates translocation by means of intra-subunit tRNA hybrid sites. Nature. 468:713–716.

32. Topf, M., K. Lasker, …, A. Salì. 2008. Protein structure fitting and refinement guided by cryo-EM density. Structure. 16:295–307.

33. Saha, M., and M. C. Morais. 2012. FOLD-EM: automated fold recognition in medium- and low-resolution (4–15 Å) electron density maps. Bioinformatics. 28:3265–3273.

34. DiMaio, F., M. D. Tyka, …, D. Baker. 2009. Refinement of protein structures into low-resolution density maps using ROSETTA. J. Mol. Biol. 392:181–190.

35. Lindert, S., N. Alexander, …, J. Meiler. 2012. EM-fold: de novo atomic-detail protein structure determination from medium-resolution density maps. Structure. 20:464–478.

36. Russel, D., K. Lasker, …, A. Salì. 2012. Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. PLoS Biol. 10:e1001244.

37. Adams, P. D., P. V. Afonine, …, P. H. Zwart. 2011. The Phenix software for automated determination of macromolecular structures. Methods. 55:94–106.

38. Singaroy, A., I. Teo, …, K. Schulten. 2016. Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps. eLife. 5:e16105.

39. Brünger, A. T., P. D. Adams, …, G. L. Warren. 1998. Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. 54:905–921.

40. Bonomi, M., G. T. Heller, …, M. Vendruscolo. 2017. Principles of protein structural ensemble determination. Curr. Opin. Struct. Biol. 42:106–116.

41. Soromena, P. D., Povesan, …, M. Vendruscolo. 2017. Simultaneous quantification of protein order and disorder. Nat. Chem. Biol. 13:339–342.

42. Allison, J. R. 2017. Using simulation to interpret experimental data in terms of protein conformational ensembles. Curr. Opin. Struct. Biol. 43:79–87.

43. Fisher, C. K., and C. M. Stultz. 2011. Constructing ensembles for intrinsically disordered proteins. Curr. Opin. Struct. Biol. 21:426–431.

44. Gaalswyk, K., M. I. Muniyat, and J. MacCallum. 2017. The emerging role of physical modeling in the future of structure determination. bioRxiv. https://doi.org/10.1101/228247.

45. Rieping, W., M. Habeck, and M. Nilges. 2005. Inferential structure determination. Science. 309:303–306.

46. Bonomi, M., C. Camilloni, …, M. Vendruscolo. 2016. Metainference: a Bayesian inference method for heterogeneous systems. Sci. Adv. 2:e1501177.

47. Lindorff-Larsen, K., R. B. Best, …, M. Vendruscolo. 2005. Simultaneous determination of protein structure and dynamics. Nature. 433:128–132.

48. Schneidman-Duhovny, D., R. Pellarin, and A. Salì. 2014. Uncertainty in integrative structural modeling. Curr. Opin. Struct. Biol. 28:96–104.

49. Hanot, S., M. Bonomi, …, R. Pellarin. 2017. Bayesian multi-scale modeling of macromolecular structures based on cryo-electron microscopy density maps. bioRxiv. https://doi.org/10.1101/113951.

50. Cavalli, A., C. Camilloni, and M. Vendruscolo. 2013. Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. J. Chem. Phys. 138:094112.

51. Bonomi, M., and C. Camilloni. 2017. Integrative structural and dynamical biology with PLUMED-ISDB. Bioinformatics. 33:3999–4000.

52. Tribello, G. A., M. Bonomi, …, G. Bussi. 2014. PLUMED 2: new feathers for an old bird. Comput. Phys. Commun. 185:604–613.

53. Prince, E. 2004. International Tables for Crystallography, Vol. C. Wiley, Hoboken, NJ.

54. Best, R. B., and G. Hummer. 2009. Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. J. Phys. Chem. B. 113:9004–9015.

55. Still, W. C., A. Tempczyk, …, T. Hendrickson. 1990. Semianalytical treatment of solvation for molecular mechanics and dynamics. J. Am. Chem. Soc. 112:6127–6129.

56. Sivia, D. S., and J. Skilling. 2006. Data Analysis: A Bayesian Tutorial. Oxford University Press, Oxford, New York.

57. Beauchamp, K. A., Y. S. Lin, …, V. S. Pande. 2012. Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. J. Chem. Theory Comput. 8:1409–1414.

58. Hess, B. 2008. P-LINCS: a parallel linear constraint solver for molecular simulation. J. Chem. Theory Comput. 4:116–122.

59. Bussi, G., D. Donadio, and M. Parrinello. 2007. Canonical sampling through velocity rescaling. J. Chem. Phys. 126:014101.

60. Ferrarotti, M. J., S. Bottaro, …, G. Bussi. 2015. Accurate multiple time step in biased molecular simulations. J. Chem. Theory Comput. 11:139–146.

61. Barducci, A., G. Bussi, and M. Parrinello. 2008. Well-tempered meta-dynamics: a smoothly converging and tunable free-energy method. Phys. Rev. Lett. 100:020603.

62. Hess, B., C. Kutzner, …, E. Lindahl. 2008. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. J. Chem. Theory Comput. 4:435–447.

63. Bartolucci, C., D. Lamba, …, H. Heumann. 2005. Crystal structure of wild-type chaperonin GroEL. J. Mol. Biol. 354:940–951.
64. Fei, X., D. Yang, ..., G. H. Lorimer. 2013. Crystal structure of a GroEL-ADP complex in the relaxed allostERIC state at 2.7 Å resolution. *Proc. Natl. Acad. Sci. USA.* 110:E2958–E2966.

65. Kawabata, T. 2008. Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a Gaussian mixture model. *Biophys. J.* 95:4643–4658.

66. Chen, Y., O. B. Clarke, ..., F. Mancia. 2016. Structure of the STRA6 receptor for retinol uptake. *Science.* 353:6302.

67. Laskowski, R. A., M. W. Macarthur, ..., J. M. Thornton. 1993. Procheck—a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* 26:283–291.

68. Daura, X., K. Gademann, ..., A. E. Mark. 1999. Peptide folding: when simulation meets experiment. *Angew. Chem. Int. Ed.* 38:236–240.

69. Kawaguchi, R., J. Yu, ..., H. Sun. 2007. A membrane receptor for retinol binding protein mediates cellular uptake of vitamin A. *Science.* 315:820–825.

70. Kawaguchi, R., J. Yu, ..., H. Sun. 2008. Mapping the membrane topology and extracellular ligand binding domains of the retinol binding protein receptor. *Biochemistry.* 47:5387–5395.

71. van Esch, H., A. Jansen, ..., J. P. Fryns. 2007. Encephalopathy and bilateral cataract in a boy with an interstitial deletion of Xp22 comprising the CDKL5 and NHS genes. *Am. J. Med. Genet. A.* 143:364–369.

72. Zhong, M., R. Kawaguchi, ..., H. Sun. 2013. Vitamin A transport and the transmembrane pore in the cell-surface receptor for plasma retinol binding protein. *PLoS One.* 8:e73838.