Real-Space Refinement with DireX: From Global Fitting to Side-Chain Improvements

Zhe Wang,[1] Gunnar F. Schröder[1,2]

[1] *Institute of Complex Systems (ICS-6), Forschungszentrum Jülich, 52425 Jülich, Germany*

[2] *Department of Physics, Heinrich-Heine-University Düsseldorf, 40225 Düsseldorf, Germany*

Received 6 November 2011; revised 22 January 2012; accepted 27 January 2012
Published online 9 March 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/bip.22046

ABSTRACT:

Single-particle cryo-electron microscopy (cryo-EM) has become an important tool to determine the structure of large biomolecules and assemblies thereof. However, the achievable resolution varies considerably over a wide range of about 3.5–20 Å. The interpretation of these intermediate- to low-resolution density maps in terms of atomic models is a big challenge and an area of active research. Here, we present our real-space structure refinement program DireX, which was developed primarily for cryo-EM-derived density maps. The basic principle and its main features are described. DireX employs Deformable Elastic Network (DEN) restraints to reduce overfitting by decreasing the effective number of degrees of freedom used in the refinement. Missing or reduced density due to flexible parts of the protein can lead to artifacts in the structure refinement, which is addressed through the concept of restrained grouped occupancy refinement. Furthermore, we describe the performance of DireX in the 2010 Cryo-EM Modeling Challenge, where we chose six density maps of four different proteins provided by the Modeling Challenge exemplifying typical refinement results at a large resolution range from 3 to 23 Å. © 2012 Wiley Periodicals, Inc. Biopolymers 97: 687–697, 2012.

Additional Supporting Information may be found in the online version of this article.

Correspondence to: Gunnar F. Schröder; e-mail: gu.schroeder@fz-juelich.de

Keywords: *Cryo-EM; DireX; low resolution; flexible fitting; real-space refinement*

This article was originally published online as an accepted preprint. The "Published Online" date corresponds to the preprint version. You can request a copy of the preprint by emailing the Biopolymers editorial office at biopolymers@wiley.com

INTRODUCTION

Single-particle cryo-electron microscopy is a powerful method to study the structure of large biomolecules and macromolecular assemblies. This technique has attracted considerable attention since the achievable resolution is steadily increasing over the past years to even better than 4 Å in some cases.[1–3] Depending on the resolution, the density maps derived from cryo-EM experiments can reveal great structural details both on individual biomolecules as well as on interactions between molecules in assemblies. However, the interpretation of these density maps is usually hampered by significant noise and, thus, by a limited resolution.

In general, it is of course desirable to build an atomic model to represent the information contained in the cryo-EM density. Computational procedures that can automatically build atomic models from the density have been developed mainly for X-ray crystallography but are in principle applicable to the interpretation of density maps obtained by cryo-EM as well. These procedures allow building almost complete models routinely at resolutions better than 3 Å,[4,5] it however becomes a great challenge at resolutions lower than 3 Å.[6,7] And at resolutions beyond about 5 Å, the density can usually only be interpreted if a higher-resolution structure or at least a fragment thereof has been determined to higher resolution by other means, such as X-ray crystallo-

Table 1 Overview of the Six Cases that were Selected from the Cryo-EM Modeling Challenge

Name	Map Resolution (Å)	Starting Model	Occupancy Refinement	Energy Minimization	Density Map Cross-correlation	
					Start	End
GroEL-ES	23.5	1AON	no	yes	0.931	0.939
GroEL-ES	7.7	1PCQ	no	no	0.826	0.852
Mm-cpn	4.3	Homology model based on 1Q3Q	no	yes	0.423	0.651
GroEL	4	1SS8	no	yes	0.624	0.660
Rotavirus VP6	3.8	1QHD	no	yes	0.212	0.281
Aquaporin-0	3	2B6P	yes	yes	0.529	0.560

The last two columns show the cross-correlation coefficient of the target density map with the density map compute from the atomic model.

graphy. We consider here only this latter case, where a higher-resolution structure (either a crystal structure or a homology model) is known already. The task is then to optimize or refine this model against the (medium- or low-resolution) cryo-EM density. This task is in principle highly similar to real-space refinement approaches that have been developed for fitting atomic models into electron density maps.[8–10] These techniques are also used extensively in interactive modeling programs such as, e.g., O[11] and Coot.[12] The main difference of density maps determined by Cryo-EM as compared to X-ray crystallography is that the density maps do not describe electron density but nucleic density and that the resolution is usually significantly lower.

Currently, a number of approaches and programs exist to fit high-resolution structures into density maps.[13,14] Several methods have been developed to perform rigid-body fitting of atomic structures into low-resolution maps.[15–19] However, when the resolution of the density improves to better than about 10 Å, deviations of the density from the known X-ray structure can become apparent, which means that the density map contains more information than just the position and orientation of the molecule. When allowing flexibility of the model during the refinement, overfitting can be a serious problem and could lead to overinterpretation of low-resolution density maps.[20] Therefore, additional retraints have to be used to reduce the effective number of degrees of freedom, thereby decreasing the unfavorable parameter to observable ratio.

Such flexible fitting of high-resolution structures into low-resolution cryo-EM maps has been realized in different ways using different approaches to introduce flexibility: NMFF[21] and other programs[22,23] select degrees of freedom for the refinement from the set of low-frequency elastic normal modes, which usually capture a large part of the collective motion of a protein. Other approaches define restraints either manually or in an automated way. The Situs package[24] provides an efficient and robust method for the localization of protein subunits in low-resolution density maps. Real space refinement and molecular dynamics simulations have been combined to fit structures into density maps.[9,25–28] The Flex-EM method uses a combination of Monte-Carlo search, conjugate gradient minimization and simulated annealing molecular dynamics.[29] Furthermore, the combination of comparative modeling and structure refinement was used to improve the sequence alignment and obtain better homology models.[30]

Here, we use the deformable elastic network (DEN)[31,32] method, which is implemented in the program DireX[31] to fit models into density maps, allowing large deformations in the atomic model. The DEN defines harmonic distance restraints between randomly chosen atom pairs within the high-resolution reference structure. During the fitting, the DEN restraints are allowed to change their equilibrium distances and are able to balance the forces from the reference model and the density map by adapting the minimum of the DEN potential.

For the Cryo-EM Modeling Challenge, six density maps from four proteins were chosen as targets for flexible fitting, which was performed with DireX. In the following, the computational approach and the structure refinement results are described in more detail.

METHODS

Selection of Targets

We selected six out of the 13 cases presented by the Cryo-EM Modeling Challenge 2010, to demonstrate the performance of DireX. The selection covers a wide range of resolutions from 3 to 23 Å for four different proteins Aquaporin-0, Rotavirus VP6, and the chaperonins GroEL (with and without the cofactor GroES) and Mm-cpn. An overview of the six cases is shown in Table 1. We decided not to submit a model for the open state of Mm-cpn at a resolution of 8 Å, as the provided target model (PDB ID 3IYF[3]) was generated with DireX and would therefore be highly similar. All starting models were docked as rigid-bodies into the corresponding density maps using the "Fit in Map" feature of the program Chimera.[33]

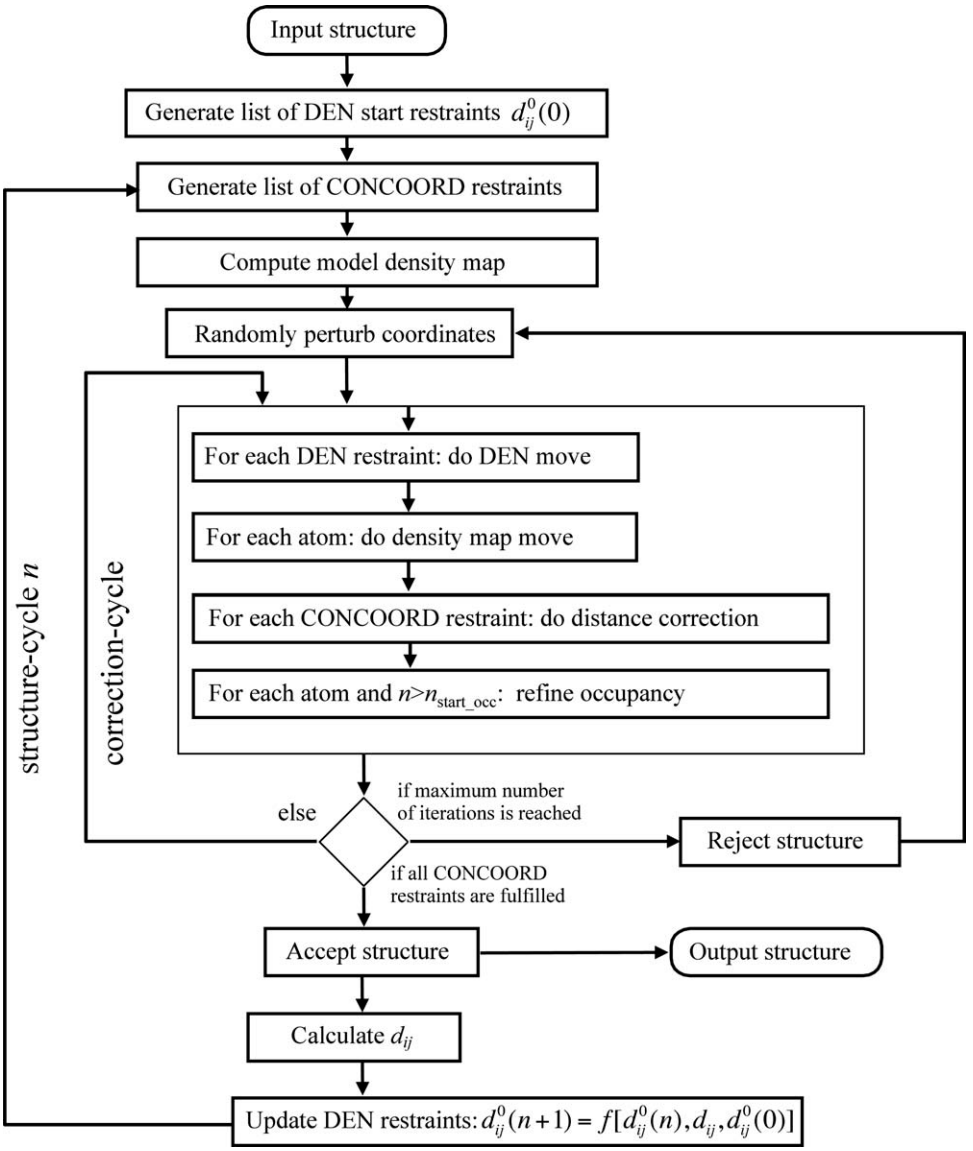

FIGURE 1 Diagram showing the workflow of the program DireX. The correction cycle implements the CONCOORD algorithm which is interspersed with the application of forces derived from the density map, from Deformable Elastic Network (DEN) restraints, and with the refinement of occupancy values.

DireX Software

To fit the starting high-resolution protein structures into the target density maps, our program DireX is applied. DireX uses an efficient conformational sampling algorithm to generate move steps, while strongly restraining chemical bond lengths, angles and planarity and preventing atom overlaps. In addition, the density map exerts forces onto the atoms that drive the model into the density map, such that the overlap of the target density with a density computed from the model is maximized. To account for the low observable-to-parameter ratio, deformable elastic network (DEN) restraints are applied, which add global restraints to a reference model but at the same time allow for refining those degrees of freedom for which the density map provides information, as is explained below. DireX is typically used to generate a structural ensemble that is biased by both the reference model as well as by a density map. Because of the efficient sampling algorithm a large number of structures can be obtained that all fit similarly well to the density map. DireX is free for academic users and can be downloaded from https://simtk.org/home/direx. In the following we describe those features of DireX that are relevant for the results presented here.

Figure 1 gives an overview of the workflow in DireX. During the course of the refinement procedure, three mains steps are performed to move the atomic coordinates and to eventually fit the structure into the density maps: (1) a conformational sampling algorithm exerts a diffusional force and generates a random walk, (2) a stochastic gradient of an electron density map moves the model into

the density map, and (3) the deformable elastic network restraints balance the influence from a reference model. These three forces are described in the next three sections separately.

Conformational Sampling Algorithm

The conformational sampling algorithm is based on the CONCOORD algorithm.[34] In the first CONCOORD step, a large number of distance restraints, which are represented as distance intervals, are generated from the input (starting) structure. These CONCOORD restraints fall into two groups: (1) bonded restraints, which maintain correct stereochemistry such as bond lengths and angles and the planarity of certain groups; and (2) nonbonded restraints which prevent atoms from overlapping and which define an upper limit for the maximally allowed coordinate change per step. The bonded restraints are computed once from the starting coordinates and are kept constant afterwards. The list of nonbonded restraints is rebuilt at every structure generation cycle. The number of CONCOORD restraints is typically about ten times larger than the number of atoms. In the second CONCOORD step, the atomic coordinates are perturbed using random numbers drawn from a Gaussian distribution. The width of this Gaussian distribution determines the effective diffusion coefficient that results from the conformational sampling algorithm. In the third CONCOORD step, the correction cycle, those atom pairs that are not within the CONCOORD distance interval are moved, disregarding all other restraints, along their interatomic vector to a target distance that is picked randomly from within the allowed interval. The list of all CONCOORD restraints is traversed in random order. The correction cycle is repeated until all CONCOORD restraints are fulfilled or until a maximum number of cycles, typically 500, is reached. A new structure obeying all CONCOORD restraints is generated typically within 5 to 50 correction cycles and is accepted as the starting point for the next structure generation cycle. Once a new structure has been accepted, new CONCOORD restraints are computed from this structure in the next iteration. The iteration over structure generation cycles results in a trajectory of structures, which, in the absence of any other forces, leads to a random walk in conformational space.

Forces Derived from the Density

In addition to the conformational sampling, forces are applied to the atoms that drive the structure into the density map. These forces are obtained by comparing the target density map, $\rho_{\mathrm{exp}}(\vec{x})$, with a density map, $\rho_{\mathrm{model}}(\vec{x})$, computed from the current coordinates of the model. At the beginning of each structure generation cycle, the current model density map is calculated by convoluting the atomic model with a kernel function that is the Fourier transform of a hollow sphere, as described in Ref. [35]. This is done before the coordinate perturbation step. Both maps, $\rho_{\mathrm{exp}}(\vec{x})$ and $\rho_{\mathrm{model}}(\vec{x})$, are normalized to have a mean value of zero and a standard deviation of one, yielding $\tilde{\rho}_{\mathrm{exp}}(\vec{x})$ and $\tilde{\rho}_{\mathrm{model}}(\vec{x})$. The difference density is computed by $\tilde{\rho}_{\mathrm{diff}}(\vec{x}) = \tilde{\rho}_{\mathrm{exp}}(\vec{x}) - \tilde{\rho}_{\mathrm{model}}(\vec{x})$. The atoms are then moved into regions of high difference density, where the model does not produce sufficient density and out of regions with low (negative) difference density, where the model produces too much density. This could in principle be achieved by computing directly a gradient, however, as experimental density maps usually contain a significant amount of noise and to make the refinement robust against this noise, we employ a stochastic gradient algorithm for the computation of the forces. For this, atoms are moved during each correction cycle by adding a vector

$$\vec{g}_i = v(s_c)\frac{1}{12}\sum_{j=1}^{12}\rho_{\mathrm{diff}}\left(\vec{r}_j\right)\frac{\left(\vec{r}_j - \vec{x}_i\right)}{\left|\vec{r}_j - \vec{x}_i\right|}$$

to each atom, where $\vec{r}_j$ are random positions taken from an isotropic Gaussian distribution with a width of 1 Å around the atom position $\vec{x}_i$. The scaling factor $v(s_c)$ depends on the correction cycle step s_c, and linearly decreases from 1 to 0 within the first 40 steps. This is done to fade out the density forces during the CONCOORD correction cycle, which enables the structure to eventually converge to the CONCOORD restraints.

Deformable Elastic Network

At low resolution, the number of experimental observables is usually smaller than the number of parameters (atomic coordinates). Therefore, potential over-fitting becomes a major issue. The general strategy of the DEN method is to refine only those degrees of freedom that are defined by the data and to use prior structural information for those not defined by the data. The DEN potential is defined by

$$E_{DEN}(n) = v(s_c)w_{DEN}\sum\left(d_{ij}(n) - d_{ij}^0(n)\right)^2$$

where $d_{ij}^0(0)$ is the distance between atom i and j at structure cycle number n, $d_{ij}^0(n)$ is the corresponding equilibrium distance, which will be changed after each structure cycle, w_{DEN} is the force constant, and $v(s_c)$ is the same scaling factor as above. The DEN restraints are defined by randomly choosing atom pairs that are within a distance interval of typically 3–15 Å. Atoms within different peptide chains are usually excluded. In general, the total number of DEN restraints is chosen to be two times the number of atoms.

DEN restraints are applied in random order during each correction cycle, which moves each pair of atoms closer to $d_{ij}^0(n)$, by a step proportional to $d_{ij}(n) - d_{ij}^0(n)$. After obtaining a structure that satisfies all CONCOORD restraints, the equilibrium distances $d_{ij}^0(n)$ of the DEN restraints are updated with the function:

$$d_{ij}^0(n+1) = d_{ij}^0(n) + \kappa\left[\gamma\left(d_{ij}(n) - d_{ij}^0(n)\right) + (1-\gamma)\left(d_{ij}^0(0) - d_{ij}^0(n)\right)\right]$$

$d_{ij}^0(n)$ and $d_{ij}(n)$ are defined as above and κ determines the network adaptation speed and is set to be smaller than 1, typically 0.1. The parameter γ balances two forces: an adaptation force $\kappa[\gamma d_{ij}(n)]$ which makes the DEN follow the structural change if a DEN restraint is distorted by forces from the density map, and a restoring force $\kappa[(1-\gamma)d_{ij}^0(0)]$ which allows the equilibrium distance of the elastic network to move back to its reference value $d_{ij}^0(0)$ (in the absence of forces from the density map). The parameter γ therefore controls how the influence of the density map and the reference model is balanced. For small γ-values the structure stays closer to the reference model and for larger γ-values the structure can fit the data better and moves farther away from the reference model. In general, higher-resolution data allow for larger γ-values, i.e., for larger structural deformations as the data contain more structural

information. However, the strength of the DEN potential, w_{DEN}, and the γ-parameter have to be optimized for each case.

Secondary Structure Information

Secondary structure elements are usually more stable than loop regions during conformational changes and they are also more conserved between homologous proteins. To account for this, DEN restraints outside secondary structure elements can be made weaker to increase the flexibility in loop regions. This affects all DEN restraints within loop regions, between secondary structure elements and loop regions, and between different helices that are connected by a loop region. A factor λ_{SS} (a value between 0 and 1) is multiplied to the strength of these DEN restraints. The factor λ_{SS} is typically chosen to be 0.5. The secondary structure assignment is done once at the beginning with the program DSSP[36] and is not updated during the refinement.

Side-Chains

At resolutions where side-chain densities become visible (better than about 4.5 Å), the starting model might have side-chain rotamers that need to be modified to fit the observed density. In that case it can be necessary to make the side-chains more flexible than the main-chain structure. For this, the DEN restraints that connect atom pairs within side-chains or that connect a main-chain atom with a side-chain atom are scaled by a factor λ_{SC} (a value between 0 and 1).

Occupancy Refinement

The conformational heterogeneity and structural flexibility of a protein can lead to missing or reduced observed density for a certain region. Sometimes entire protein domains can have significantly lower density values than the rest of the protein. At higher resolutions there could be some side-chain densities visible, while others are missing. Ignoring this effect can lead to artifacts during the fitting as the excess model density will be forced into the too small observed density volume, which results in a shifted position and orientation as well as a distortion of the fitted structure. Here we account for this effect by introducing the occupancy, Ω_i, as an additional parameter for each atom i. The occupancy is a value between 0 and 1 and scales the contribution of each atom to the computation of the model density. The occupancy values are updated during the occupancy refinement according to

$$\Omega_i(n+1) = \Omega_i(n) - \kappa_{\mathrm{occ}} \frac{\tilde{\rho}_{\mathrm{model}}(\vec{x}_i) - \tilde{\rho}_{\mathrm{exp}}(\vec{x}_i) - \langle \tilde{\rho}_{\mathrm{diff}} \rangle}{\tilde{\rho}_{\mathrm{model}}(\vec{x}_i)},$$

where $\Omega_i(n+1)$ is the updated occupancy value of atom i at the occupancy refinement step n, $\langle \tilde{\rho}_{\mathrm{diff}} \rangle$ is the average value of the difference density, and κ_{occ} is a damping factor that is typically chosen to be 0.01. The damping slows down the occupancy refinement, such that the positional refinement has sufficient time to adapt to the updated occupancy values. The occupancy refinement has a similar effect as the B-factor refinement in reciprocal-space refinement of crystallographic structures against diffraction data.

Introducing additional parameters for the refinement adds significant risk of over-fitting the data. For this reason, the occupancy parameters can be grouped such that the same Ω_i -value is used for all atoms within one protein residue. In addition, the occupancy values between neighboring residues can be restrained, which further reduces the effective number of free occupancy parameters. The occupancy refinement is usually performed as the last step in the whole refinement procedure.

Energy Minimization

The CONCOORD algorithm defines restraints in terms of allowed distance intervals. The smaller these distance intervals are, the longer it takes for the algorithm to converge to a new structure. DireX typically uses relatively large intervals for the bonded interactions, for example the width of the allowed distance interval for bond lengths is 0.12 Å. This leads to small deviations in the local geometry. The distance intervals could be chosen smaller, leading to better local geometry, however, at the cost of slower convergence. To correct these small deviations in bond geometry, it is usually sufficient to minimize the structure with the program CNS[37,38] for 100 steps.

RESULTS

The Cryo-EM Modeling Challenge included 13 modeling problems in total from which we selected 6 as targets for refinement with DireX. The runtime of DireX depends on the number of atoms of the model and was 7 min for the smallest system (Aquaporin, 1785 atoms) and about 4 hours for the largest system (GroEL-GroES, 58674). For each of these six cases we performed refinements with different DEN parameters using different weights to balance the influence of density map and reference model. In particular, three parameters were systematically changed: 1) γ, which controls the deformability of the elastic network, 2) λ_{SS}, which scales the strength of those DEN restraints that are involved in loop regions, 3) λ_{SC}, which scales the strength of DEN restraints of side-chains. All combinations of four different values (0.0, 0.4, 0.8, 1.0) were tested for the three parameters, yielding 64 refined models in total.

Some structures fit the density better while others stayed closer to the reference model. The question is which of the parameters yields the best model? How do the different forces need to be balanced to end up with a structure that is closest to the true structure? The answer could be given by a cross-validation procedure. However, since the implementation of a cross-validation approach in DireX was still in test phase during the Cryo-EM Modeling Challenge, we decided for a simpler, practical approach to choose the optimal refinement parameters. Since the DEN restraints are global and randomly chosen and there are no additional restraints from a molecular mechanics force field, the Ramachandran statistics is a relatively good indicator for how much the local structure is distorted during the refinement. We therefore chose the fitting parameters to trade off between a good

Ramachandran score, which is the fraction of residues in the allowed region of the Ramachandran plot as measured by Molprobity[39], and a high correlation of the target density map with the density map computed from the model. It should be noted that the fit to the density map can always be improved when weakening the restraints or increasing the γ-value, however risking severe artifacts from overfitting. Here, the choice of the optimum DEN parameters was done manually and was, thus, rather subjective. We decided to not allow the Ramachandran score to drop by more than 10% of the starting value. Supporting Information Figure S1 shows examples for the GroEL-GroES 7.7Å and the Rotavirus cases. With the rigorous cross-validation technique mentioned above, this parameter choice would now be much better defined.

In general, the choice of the parameters depends, among other factors, critically on the resolution and the difference between starting model and correct solution. At higher resolution, larger γ-values (higher deformability of the elastic network) and weaker restraints can be justified, while at lower resolution stronger restraints and smaller γ-values need to be used. However, if the model is already close to the correct structure, even strong restraints and a low γ-value can yield a good fit to the density map. Increasing side-chain flexibility (lower λ_{SC}-values) can only be justified, if the density shows clear side-chain details, typically at resolutions better than 4.5 Å.

Supporting Information Table S1 shows the optimal values determined for the each case. In the following we will discuss the individual cases in order of increasing resolution.

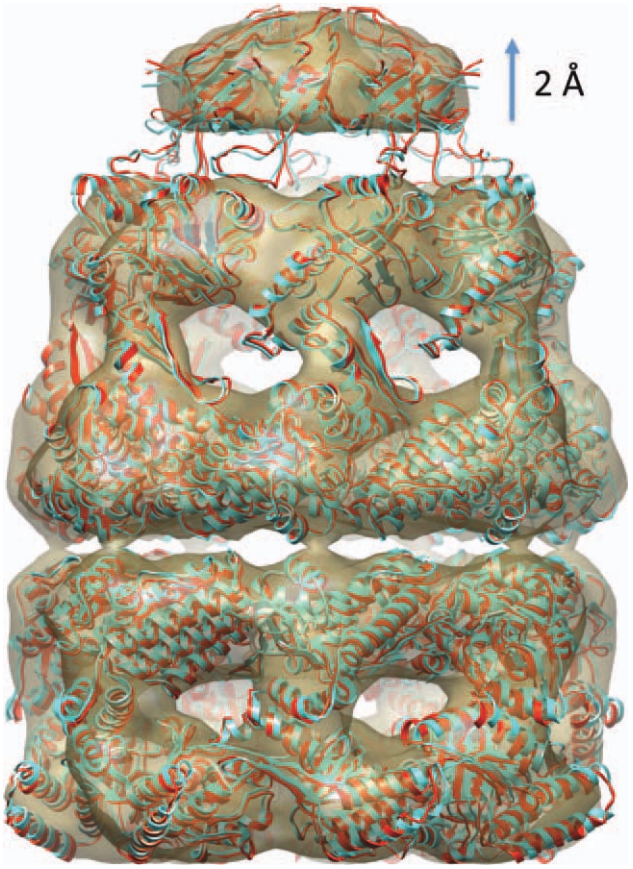

FIGURE 2 Comparison of starting (cyan) and refined (red) model of GroEL at a resolution of 23.5 Å. The starting model is the crystal structure 1AON. The GroES subunits were shifted upward by about 2 Å upon refinement. Strong DEN restraints kept the individual subunits almost rigid during the refinement.

GroEL-GroES (23.5 Å)

For the density map of GroEL-GroES with a resolution of 23.5 Å,[40] the lowest resolution of all cases, we chose to start the refinement from the crystal structure with PDB ID 1AON,[41] which was determined to a resolution of 3.0 Å. The starting model was docked, as in all other cases described below, as a rigid-body into the density map using the program Chimera, defining the starting point for subsequent refinement. The refinement included 1000 steps with DireX followed by 100 steps of energy minimization with CNS. Strong DEN restraints were used which kept the subunits almost rigid. The average Cα-RMSD (all RMSD values are computed after least-square alignment) between the subunits in the starting and the refined model was 0.9 Å, 0.6 Å, and 0.9 Å, for the lower (*trans*-) ring, the upper (*cis*-) ring, and GroES, respectively, showing that only little conformational change was allowed within one subunit. However, no restraints were applied between the subunits, which resulted in small shifts of subunits with respect to each other. The largest shift was observed for GroES, which shifted upwards (*z*-direction) by about 2 Å compared with the starting model (1AON). The starting model (cyan) and the refined model (red) are superimposed in Figure 2 together with the density map (ochre). The Cα-RMSD between the entire starting and refined model is 1.7 Å. The Ramachandran analysis yielded 70% of residues in the allowed region for the DireX refined model, while the CNS energy minimization improved this value to 86%, compared with 79% for the starting model. No symmetry restraints were used, which resulted in small differences between the identical subunits of 0.32 Å Cα-RMSD on average.

GroEL-GroES (7.7 Å)

For the 7.7 Å GroEL-GroES density map,[42] we chose the crystal structure 1PCQ[43] (determined to 2.8 Å), as the starting model for the refinement. The DEN restraints involving

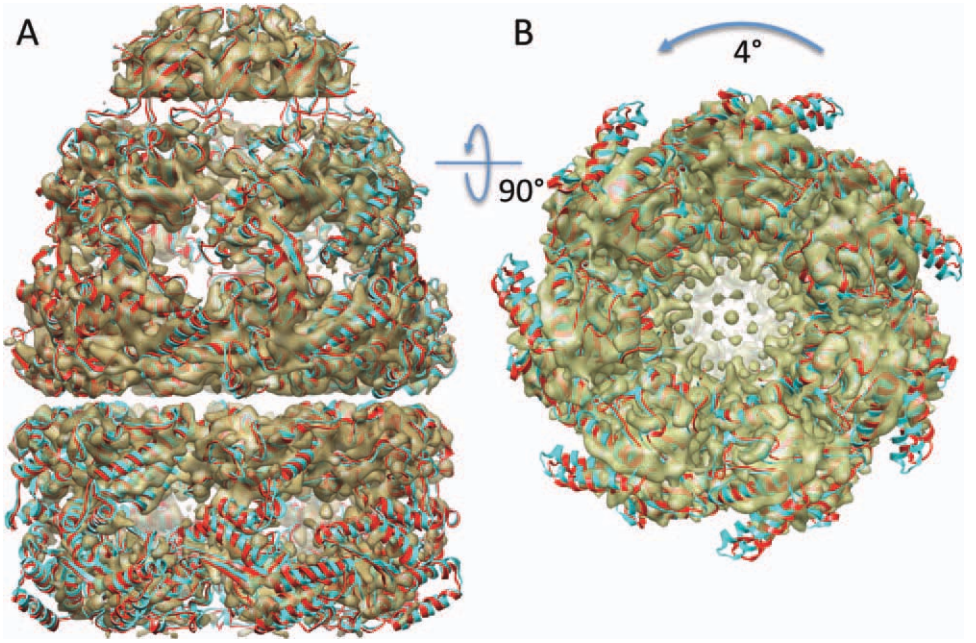

FIGURE 3 Comparison of starting (cyan) and refined (red) model of GroEL at a resolution of 7.7 Å. The crystal structure 1PCQ was used as the starting model. A: The overall structural change upon refinement is small with an Cα-RMSD of only 1.66 Å. B: The bottom view reveals that the largest difference is a rotation of the lower (*trans*-) ring with respect to the upper (*cis*-) ring.

loop regions were weakened by a factor of 0.8 compared with restraints within α-helices or β-sheets, to allow for more flexibility in the loop regions. The starting model fitted already well to the density such that the refinement with 1000 steps moved the structure only to a Cα-RMSD of 1.66 Å. Figure 3 shows the starting model and the final refined structure superimposed. The models are superimposed using only the subunits in the *cis*-ring. The largest structural change is a rotation of the *trans*-ring with respect to the *cis*-ring by about 4° as is shown in the bottom view (Figure 3B). Again, no symmetry restraints were used, which resulted in small differences between the identical subunits of 0.26 Å Cα-RMSD on average.

All models submitted to the Cryo-EM Modeling Challenge 2010 are publicly accessible through the website http://ncmi.bcm.edu/challenge/. The GroEL-GroES-7.7Å case is the only case where we did not perform a final energy minimization with CNS to give the interested reader the chance to assess a model generated directly by DireX.

Mm-cpn (4.3 Å)

As a starting model for the 4.3 Å Mm-cpn density map,[3] we used a homology model based on the thermosome structure (PDB ID 1Q3Q[44]) as described earlier.[3] The DEN restraints were limited to the backbone atoms in α-helices and β-sheets, which means that restraints involving loop regions were completely removed. Likewise, no DEN restraints were used for side-chain atoms. We performed 500 steps of refinement with DireX followed by 500 steps of energy minimization with CNS. The conformational change was relatively large, with an all-atom RMSD of 5.4 Å between starting and refined structure. Figure 4 shows a superposition of only one subunit of these two models. The complete system used in the refinement, however, comprised all 16 (identical) subunits (61,952 atoms), which are not shown for clarity. Figure 4A shows a detailed view of clear differences in the backbone trace and a much better fit to the density for the refined structure.

GroEL (4 Å)

For the 4 Å GroEL density map[45] fitting, we chose the PDB structure 1SS8[46] with a resolution of 2.7 Å. As the 1SS8 contains just a single ring structure, while the density map contained two rings.

The single ring 1SS8 model was duplicated and both rings combined by docking both rings into the density map using the program Chimera. Here the DEN restraints for the side-chains were slightly weakened by a factor of 0.8. However, there are only few side-chains visible in the density map, which did not allow for more flexibility of the side-chain restraints. In total, 1000 steps of DireX refinement were performed. The obtained model was then energy minimized with CNS for 100 steps, yielding a refined model with an all-

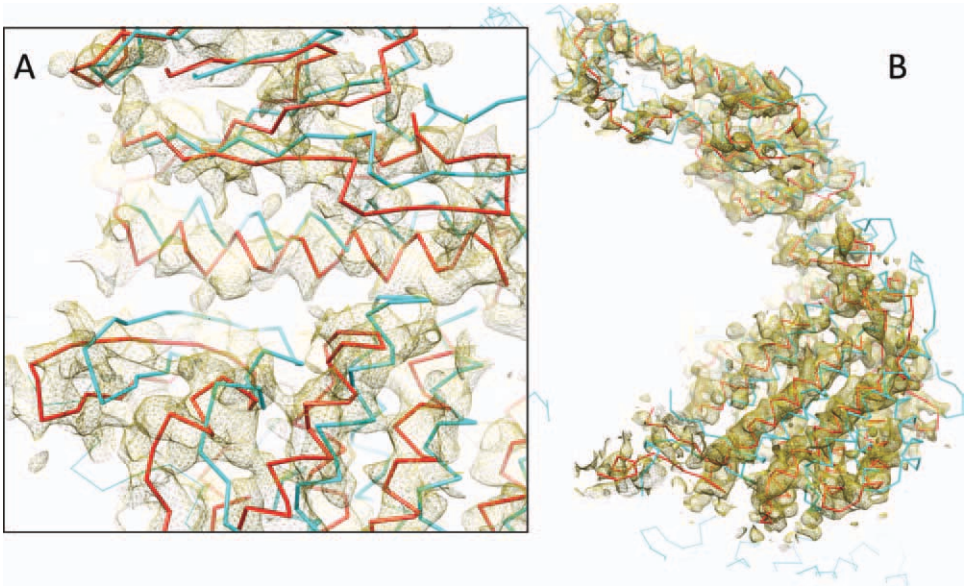

FIGURE 4 Closed state of the chaperonin Mm-cpn at a resolution of 4.3 Å. Shown is the starting model (cyan) and the model refined with DireX (red) superimposed on the density map. Only one out of all 16 subunits is shown for clarity. The conformational change during the refinement was relatively large with an RMSD of 5.4 Å.

atom RMSD of 3.0 Å to the starting model. Figure 5 shows both structures superimposed. The most prominent motion is an outward tilt of the apical domains, which leads to a small opening of the subunits with respect to the starting model. Figures 5B and 5C show detailed views on how secondary structure elements moved into the corresponding density and are thereby improving the fit. The fraction of residues in the allowed regions of the Ramachandran plot decreased from an initial 97% down to 82% for the refined model. The average Cα-RMSD between the symmetry related subunits is 0.90 Å.

Rotavirus VP6 (3.8 Å)

We chose the given target structure 1QHD[47] as the starting structure for fitting against the Rotavirus density map[1]. Calcium ions, chloride ions and zinc ions were removed from the PDB model, and three copies of the single chain from 1QHD were docked into the density map using Chimera to form the full triplet structure of VP6. This starting structure fits already very well to the density. After 1000 steps of refinement with DireX and 100 steps of energy minimization with CNS, the value of the Cα-RMSD to the starting model was only 0.67 Å. Although the overall structural change was tiny, many of the side-chains that were initially out of the density moved into the density during the fitting procedure. This is reflected in the significantly larger all-atom RMSD of 3.0 Å, compared with the Cα-RMSD of 0.67 Å. Examples of this improvement are shown in Figure 6. The side-chains shown in Figures 6B and 6C moved from the starting conformation (cyan) into the density in the final structure (red).

Aquaporin-0 (3 Å)

For the Aquaporin density map, which has been obtained by electron crystallography, we chose a crystal structure of an open state (PDB ID 2B6P,[48] 2.4 Å) as the starting structure. The difference of this structure to the given target structure (PDB ID 3M9I[49]) is small with a Cα-RMSD of only 0.47 Å. Five residues at the N-terminus (residue 2–6) and 37 residues at the C-terminus (residue 227–263) were removed from the starting structure since they were not visible in the density map and were also not present in the target structure. The side-chains were left unrestrained and the strength of DEN restraints that involved atoms in loop regions were scaled with a factor 0.5, to increase the flexibility of the loops. After 1000 steps of refinement with DireX and 100 steps of energy minimization with CNS, the refined model shows an all-atom RMSD of 0.84 Å and a Cα-RMSD of 0.43 Å. The refined model, shown in Figure 7, is a bit closer to the target structure (PDB ID 3M9I) than the starting model. The GDT-TS score[50] measures the similarity between two structures and is a value between 0 and 100% for identical structures. The GDT-TS score of the refined model increased from 91.6 to 92.2% and the RMSD decreased marginally from 0.47 to 0.42 Å. The Ramachandran score decreased slightly from 91 to 88%.

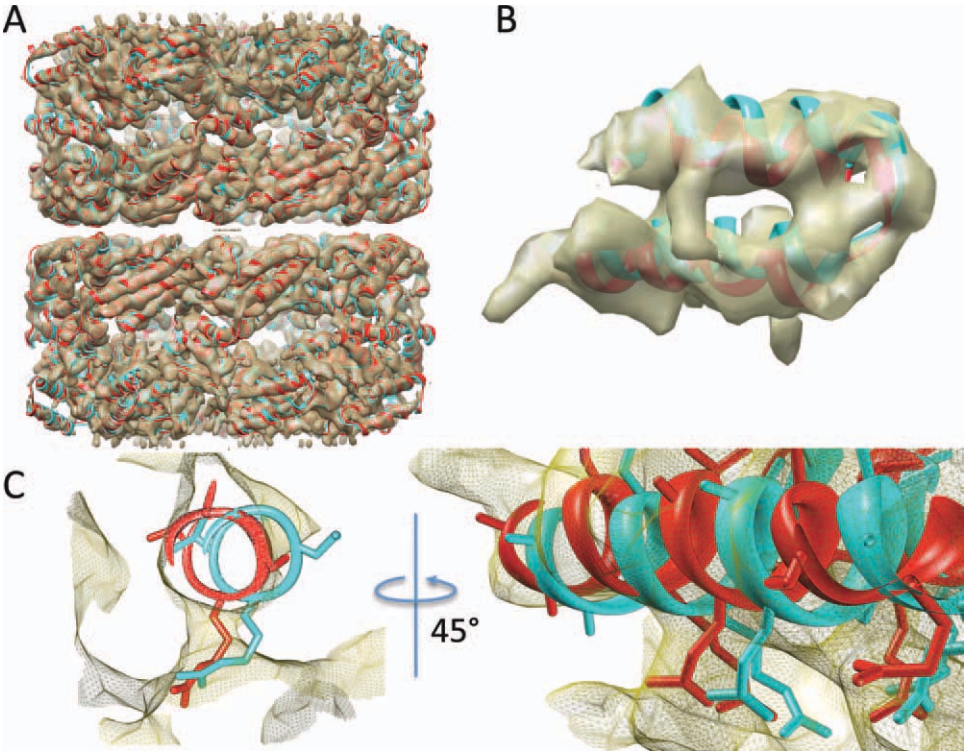

FIGURE 5 Refinement results for GroEL at a resolution of 4 Å. A: Shows a side view of GroEL with the starting (cyan) and the refined model (red) superimposed. B,C: show detailed views on how the fit of α-helices were improved by the refinement. Only few side-chains were visible in the density, such that DEN restraints for side-chains were chosen to be almost as strong as for the backbone structure.

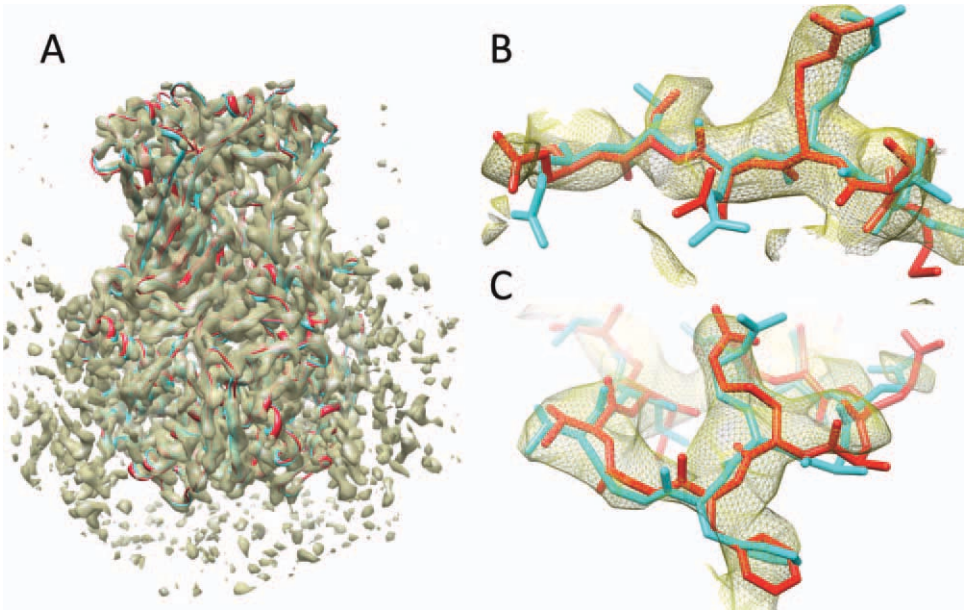

FIGURE 6 Rotavirus VP6 at a resolution of 3.8 Å. A: Showing a side-view of the VP6 trimer refined to the density. The main differences between the starting (cyan) and the refined model (red) were improvements of the side-chain fits to the density, as shown in (B) and (C).

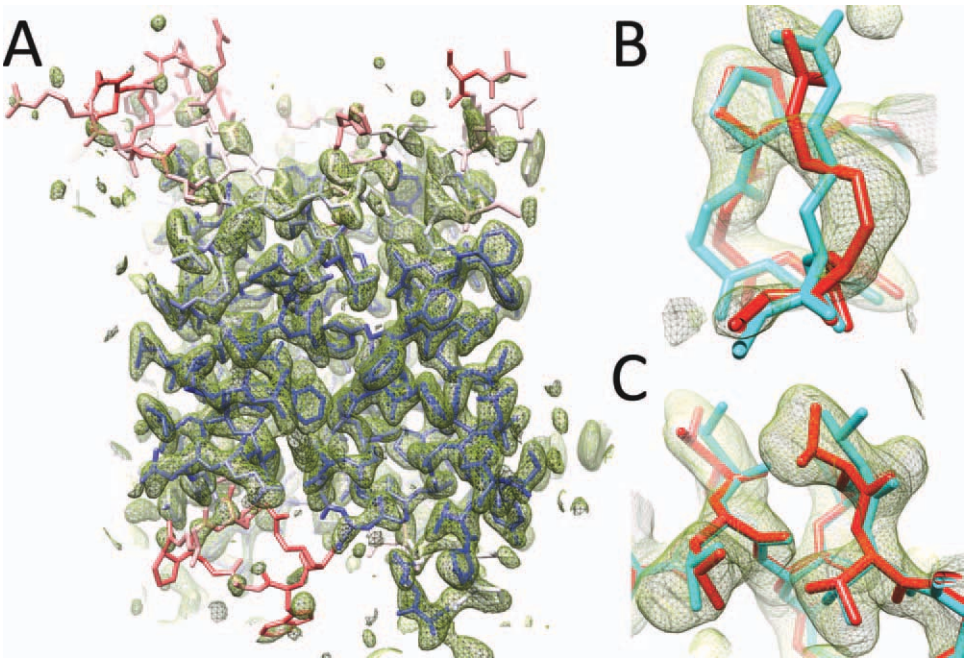

FIGURE 7 Aquaporin-0 at a resolution of 3 Å. A: Shows the refined model superimposed on the target density map. The density for the solvent exposed protein residues (upper and lower side) were significantly lower than for residues within the membrane core region. An occupancy refinement was performed to avoid that residues without sufficient density would drift into high-density regions, where they would cause distortions. The atomic model is color-coded by the occupancy values. One occupancy value was used per residue. At this relatively high resolution, the side-chains were left completely unrestrained. B,C: show the resulting improvements of side-chain positions.

The density for the solvent exposed regions of the protein is weaker than for the membrane embedded part. This can lead to distortions of the structure, since residues for which there is insufficient density will move into higher density regions. To reduce such artifacts from missing or reduced density, we performed an occupancy refinement using one occupancy value per residue and restraints between the occupancy values in neighboring residues (see Methods) to avoid introducing too many additional parameters. The occupancy values are used in DireX to scale the amplitude of each atom in the calculation of the model density map. The refined occupancy values for protein residues range from 0.41 to 1.00 in this case. Figure 7A shows the backbone of the refined model color-coded by the occupancy value, where blue means large and red low occupancy. The DireX parameters used for the occupancy refinement are shown in the Supporting Information Methods. Figures 7B and 7C show details of the improvement of the side-chains by comparing the starting model (cyan) with the refined model (red).

DISCUSSION

The aim of the Cryo-EM Modeling Challenge was to provide a number of prototypic cryo-EM data sets to allow for comparing different approaches for the interpretation of cryo-EM density maps. We applied our program DireX to a variety of cases from this Modeling Challenge including the lowest (23.5 Å) and highest (3 Å) resolution density maps to demonstrate its wide range of applicability. Starting with the lowest resolution density map (GroEL, 23.5 Å), we refined a strongly restrained model of GroEL nearly as a rigid-body and identified shifts of individual subunits. For the highest resolution density maps we improved the fit of secondary structure elements, backbone structures and side-chain positions.

DireX employs deformable elastic network (DEN) restraints, which can be adapted to the particular case and most importantly to the resolution and the level of detail that is provided by the data. The DEN restraints combine information from the density map and from a reference model in that those degrees of freedom for which the density provides information are refined and those degrees of freedom that are not defined by the data are instead defined by the reference model. This balancing between reference model and data is controlled automatically. However, the criteria to choose the optimal properties of the DEN restraints, such as strength, deformability, and distance cutoffs is still somewhat arbitrary and subjective. A cross-validation approach, which

is standard in X-ray crystallographic refinement, is able to determine these optimal parameters and will be available with the next release of DireX; this is expected to provide a more objective approach to choosing restraints and will enable to detect overfitting more easily.

The authors thank Wah Chiu and Steve Ludtke for organizing the Cryo-EM Modeling Challenge, which is a great opportunity for the whole community to learn from each other and stimulates new approaches for the interpretation of Cryo-EM data.

REFERENCES

1. Zhang, X.; Settembre, E.; Xu, C.; Dormitzer, P. R.; Bellamy, R.; Harrison, S. C.; Grigorieff, N. Proc Natl Acad Sci USA 2008, 105, 1867–1872.
2. Zhou, Z. H. Curr Opin Struct Biol 2008, 18, 218–228.
3. Zhang, J.; Baker, M. L.; Schröder, G. F.; Douglas, N. R.; Reissmann, S.; Jakana, J.; Dougherty, M.; Fu, C. J.; Levitt, M.; Ludtke, S. J.; Frydman, J.; Chiu, W. Nature 2010, 463, 379–384.
4. Terwilliger, T. C. Acta Cryst D 2003, D59, 38–44.
5. Langer, G.; Cohen, S. X.; Lamzin, V. S.; Perrakis, A. Nat Protocols 2008, 3, 1171–1179.
6. Baker, M. L.; Abeysinghe, S. S.; Schuh, S.; Coleman, R. A.; Abrams, A.; Marsh, M. P.; Hryc, C. F.; Ruths, T.; Chiu, W.; Ju, T. J Struct Biol 2011, 174, 360–373.
7. Baker, M. L.; Ju, T.; Chiu, W. Structure 2007, 15, 7–19.
8. Diamond, R. Acta Cryst A 1971, 27, 436–452.
9. Korostelev, A.; Bertram, R.; Chapman, M. S. Acta Cryst D 2002, 58, 761–767.
10. Chen, Z.; Blanc, E.; Chapman, M. S. Acta Cryst D 1999, 55, 464–468.
11. Jones, T. A.; Zou, J.-Y.; Cowan, S. W.; Kjeldgaard, M. Acta Cryst A 1991, 47, 110–119.
12. Emsley, P.; Lohkamp, B.; Scott, W. G.; Cowtan, K. Acta Cryst D 2010, 66, 486–501.
13. Fabiola, F.; Chapman, M. S. Structure 2005, 13, 389–400.
14. Rossmann, M. G.; Morais, M. C.; Leiman, P. G.; Zhang, W. Structure 2005, 13, 355–362.
15. Volkmann, N.; Hanein, D. J Struct Biol 1999, 125, 176–184.
16. Rossmann, M. G. Acta Cryst D 2000, D65, 1341–1349.
17. Chaco, P.; Wriggers, W. J Mol Biol 2002, 317, 375–384.
18. Roseman, A. M. Acta Cryst D 2000, 56, 1332–1340.
19. Lasker, K.; Topf, M.; Sali, A.; Wolfson, H. J. J Mol Biol 2009, 388, 180–194.
20. Brünger, A. T. Nature 1992, 355, 472–475.
21. Tama, F.; Miyashita, O.; Brooks, C. L. J Mol Biol 2004, 337, 985–99.
22. Suhre, K.; Navaza, J.; Sanejouand, Y.-H. Acta Cryst D 2006, D62, 1098–1100.
23. Delarue, M.; Dumas, P. Proc Natl Acad Sci USA 2004, 101, 6957–6962.
24. Wriggers, W.; Milligan, R. A.; Mccammon, J. A. J Struct Biol 1999, 125, 185–195.
25. Trabuco, L. G.; Villa, E.; Mitra, K.; Frank, J.; Schulten, K. Structure 2008, 16, 673–683.
26. Orzechowski, M.; Tama, F. Biophys J 2008, 95, 5692–5705.
27. Gao, H.; Frank, J. Structure 2005, 13, 401–406.
28. Zheng, W. Biophys J 2011, 100, 478–488.
29. Topf, M.; Lasker, K.; Webb, B.; Wolfson, H.; Chiu, W.; Sali, A. Structure 2008, 16, 295–307.
30. Topf, M.; Baker, M. L.; Marti-Renom, M. A.; Chiu, W.; Sali, A. J Mol Biol 2006, 357, 1655–1668.
31. Schröder, G. F.; Brunger, A. T.; Levitt, M. Structure 2007, 15, 1630–1641.
32. Schröder, G. F.; Levitt, M.; Brunger, A. T. Nature 2010, 464, 1218–1222.
33. Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. J Comp Chem 2004, 25, 1605–1612.
34. de Groot, B. L.; van Aalten, D. M.; Scheek, R. M.; Amadei, A.; Vriend, G.; Berendsen, H. J. Proteins 1997, 29, 240–251.
35. Chapman, M. S. Acta Cryst D 1995, 51, 69–80.
36. Kabsch, W.; Sander, C. Biopolymers 1983, 22, 2577–2637.
37. Brunger, A. T.; Adams, P. D.; Clore, G. M.; DeLano, W. L.; Gros, P.; Grosse-Kunstleve, R. W.; Jiang, J. S.; Kuszewski, J.; Nilges, M.; Pannu, N. S.; Read, R. J.; Rice, L. M.; Simonson, T.; Warren, G. L. Acta Cryst D 1998, 54, 905–921.
38. Brunger, A. T. Nat Protocols 2007, 2, 2728–2733.
39. Chen, V. B.; Arendall, W. B.; Headd, J. J.; Keedy, D. A.; Immormino, R. M.; Kapral, G. J.; Murray, L. W.; Richardson, J. S.; Richardson, D. C. Acta Cryst D 2010, D66, 12–21.
40. Ranson, N. A.; Farr, G. W.; Roseman, A. M. B. G.; Fenton, W. A.; Horwich, A. L.; Saibil, H. R. Cell 2001, 107, 869–879.
41. Xu, Z.; Horwich, A. L.; Sigler, P. B. Nature 1997, 388, 741–750.
42. Ranson, N. A.; Clare, D. K.; Farr, G. W.; Houldershaw, D.; Horwich, A. L.; Saibil, H. R. Nat Struct Mol Biol 2006, 13, 147–152.
43. Chaudhry, C.; Farr, G. W.; Todd, M. J.; Rye, H. S. Brunger, A. T.; Adams, P. D.; Horwich, A. L.; Sigler, P. B. EMBO J 2003, 22, 4877–4887.
44. Shomura, Y.; Yoshida, T.; Iizuka, R.; Maruyama, T.; Yohda, M.; Miki, K. J Mol Biol 2004, 335, 1265–1278.
45. Ludtke, S. J.; Baker, M. L.; Chen, D.-H.; Song, J.-L.; Chuang, D. T.; Chiu, W. Structure 2008, 16, 441–448.
46. Chaudhry, C.; Horwich, A. L.; Brunger, A. T.; Adams, P. D. J Mol Biol 2004, 342, 229–45.
47. Mathieu, M.; Petitpas, I.; Navaza, J.; Lepault, J.; Kohli, E.; Pothier, P.; Prasad, B. V.; Cohen, J.; Rey, F. A. EMBO J 2001, 20, 1485–1497.
48. Gonen, T.; Cheng, Y.; Sliz, P.; Hiroaki, Y.; Fujiyoshi, Y.; Harrison, S. C.; Walz, T. Nature 2005, 438, 633–638.
49. Hite, R. K.; Li, Z.; Walz, T. EMBO J 2010, 29, 1652–1658.
50. Zemla, A. Nucleic Acids Res 2003, 31, 3370–3374.

Reviewing Editor: Steven J. Ludtke