Structure
Article

Protein Structure Fitting and Refinement Guided by Cryo-EM Density

Maya Topf,[1,*] Keren Lasker,[2,3] Ben Webb,[2] Haim Wolfson,[3] Wah Chiu,[4] and Andrej Sali[2,*]
[1]School of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, United Kingdom
[2]Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, Byers Hall, 1700 4th Street, University of California, San Francisco, San Francisco, CA 94158-2330, USA
[3]School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel
[4]National Center for Macromolecular Imaging, Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA
*Correspondence: m.topf@cryst.bbk.ac.uk (M.T.), sali@salilab.org (A.S.)
DOI 10.1016/j.str.2007.11.016

SUMMARY

For many macromolecular assemblies, both a cryo-electron microscopy map and atomic structures of its component proteins are available. Here we describe a method for fitting and refining a component structure within its map at intermediate resolution (<15 Å). The atomic positions are optimized with respect to a scoring function that includes the cross-correlation coefficient between the structure and the map as well as stereochemical and nonbonded interaction terms. A heuristic optimization that relies on a Monte Carlo search, a conjugate-gradients minimization, and simulated annealing molecular dynamics is applied to a series of subdivisions of the structure into progressively smaller rigid bodies. The method was tested on 15 proteins of known structure with 13 simulated maps and 3 experimentally determined maps. At ~10 Å resolution, Cα rmsd between the initial and final structures was reduced on average by ~53%. The method is automated and can refine both experimental and predicted atomic structures.

INTRODUCTION

High-resolution structures of macromolecular assemblies, such as ribosomes, viruses, ion channels, and chaperones, are needed for studying their function and evolution (Sali et al., 2003). Although a number of assembly structures have been determined by X-ray crystallography and NMR spectroscopy, thousands of complexes remain to be structurally defined. Thus, improved methods are needed for structure characterization of assemblies at near-atomic resolution, providing approximate positions of the main chain and side chains.

Single-particle cryo-electron microscopy (cryo-EM) has already proven useful for determining macromolecular assembly structures at resolutions lower than approximately 5 Å (Jiang and Ludtke, 2005; Chiu et al., 2005). With very small sample amounts, it can determine the single-particle structures of assemblies with molecular weights larger than approximately 150 kDa. A particularly important advantage of cryo-EM is its ability to visualize different functional states (Saibil, 2000; Mitra and Frank, 2006). However, cryo-EM is often hampered by its relatively low resolution, which does not yield direct determination of atomic structures.

Fortunately, atomic-resolution structures of the isolated assembly components (e.g., domains, proteins, and complexes of a subset of all proteins in the assembly) are often available from crystallography, NMR spectroscopy, or comparative protein structure modeling (Eswar et al., 2007). By fitting the structures of these components into a cryo-EM density map of the whole assembly, a more detailed picture of the intact assembly can be provided (Rossmann et al., 2005; Topf and Sali, 2005). This task can be performed by a manual adjustment of the components in the map using interactive visual tools (Goddard et al., 2007). However, a better alternative is to use an automated computational method to decrease the level of subjectivity as well as increase the accuracy and efficiency (Chiu et al., 2005; Fabiola and Chapman, 2005).

Most such methods attempt to find an optimal position and orientation of a rigid component in the density map by optimizing a quality-of-fit measure (rigid fitting), such as the crosscorrelation coefficient between the component and the map. However, the atomic structure of the isolated component is often not the same as that in the assembly. The variations can originate from the different conditions under which the isolated component and assembly structures are determined and from errors in the experimental methods (Alber et al., 2004). Common conformational differences are shear and hinge movements of domains and secondary structure elements, as well as loop distortions and movements. Furthermore, when an experimentally determined structure of the component is unavailable, the use of protein structure prediction methods (Baker and Sali, 2001) can also introduce additional errors, such as the misassignment of secondary structure elements to incorrect sequence regions, which will cause their shifts in space in comparative modeling.

To address the problem of fitting an inaccurate component structure into a cryo-EM map, the conformation of the component needs to be optimized simultaneously with its position and orientation in the cryo-EM map (flexible fitting). Several such methods have been developed. The Situs package relies on a reduced representation of the component structure and the density map to deform the structure while fitting the map (Wriggers et al., 1999). NMFF-EM and other programs (Tama

et al., 2002, 2004; Ming et al., 2002; Suhre et al., 2006) use normal mode analysis (Brooks and Karplus, 1983) to follow the dynamics of the components in the context of a cryo-EM map. RSRef performs real-space refinement to simultaneously optimize the stereochemistry and fit of the structure into the density map (Fabiola and Chapman, 2005; Chen and Champman, 2001). Our Mod-EM and Moulder-EM methods consider the flexibility of the component structures via the fitting of alternative comparative models based on different sequence-structure alignments and different loop conformations (Topf et al., 2005, 2006). A similar use of a cryo-EM map as a filter was applied to ab initio models (Baker et al., 2006). The S-flexfit method exploits the structural variability of protein domains within a given superfamily (Velazquez-Muriel et al., 2006).

The input for almost all flexible fitting methods is the initial structure of the component rigidly fitted into the approximate position and orientation in the density map. This task is often performed by a separate rigid-body fitting program. Most of these methods also require a one-to-one correspondence between the fitted component and the density map (i.e., the map has to be segmented or masked around the region of interest). Furthermore, except for RSRef, current methods do not explicitly take into account the stereochemistry and nonbonded interactions of proteins during deformation and fitting into the map. Instead, they employ a final step of energy minimization to "fix" potential nonphysical geometries introduced during deformation and fitting.

Here we present a method (Flex-EM) that integrates rigid and flexible fitting of a component structure into the cryo-EM density map of their assembly. The component structure can originate from either an experiment (e.g., in a different chemical state) or a modeling calculation. The method combines the identification of the position and orientation of the component in the larger map with the refinement of its atomic conformation (Theory). We tested the method on a benchmark of 13 protein structures consisting of one or two domains in a nonnative conformation; these structures are fitted and refined in the context of their native density maps simulated at 4–14 Å resolution. We also tested it on two structures with experimentally determined cryo-EM maps at this resolution range (Results). Finally, we discuss our approach and its implications for refining structures and models of assembly components using cryo-EM density maps (Discussion).

RESULTS

Theory

The goal is to refine an atomic structure of a protein, given an initial structural model and a cryo-EM-derived density map. The refined structure needs to fit optimally into the density map as well as satisfy the general rules of protein structures. We express this task as an optimization problem. Thus, we need to specify (1) the representation of the protein structure; (2) the scoring function; and (3) the optimization protocol.

Representation of the Protein Structure

The input to our protocol includes an atomic structure of a protein (probe, P) and a density map at intermediate resolution (<15 Å) (Figure 1). The density map is represented by intensities at points i on a cubic grid ($\rho^{EM}{}_i$). The spacing between the grid points is equal to the sampling of the input density map (Å/voxel). The probe is defined by its N atomic coordinates and corresponding atomic numbers in real space, using the same coordinate system as for the grid. In addition, the probe density of atom j at position $\vec{r}_i$ is

$$\rho^{P}_{i,j} = \frac{Z_j}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{\vec{r}_j - \vec{r}_i}{\sigma}\right)^2\right], \quad (1)$$

where $\vec{r}_j$ is the position of atom j, Z_j is its atomic number, and σ is 0.425 times the resolution of the density map. This value was calculated based on the full width at half maximum criterion (i.e., the resolution equals $2(\vec{r}_j - \vec{r}_i)$ when $\rho^{P}_{i,j}$ is equal to half of its maximum value).

A major problem that needs to be overcome is the large size of the search space. To reduce the number of degrees of freedom, the structure is partitioned into L rigid bodies, b_l. A rigid body $b_l(\vec{r}, \vec{Z}, \vec{m}) \in B$ can be any set of atoms, including a single atom, a secondary structure element, a domain, or the whole protein; $\vec{r}$, $\vec{Z}$, and $\vec{m}$ represent the coordinates, atomic numbers, and atomic masses in this set, respectively. B is a set of rigid bodies that covers the whole probe structure (P) such that each atom is a member of exactly one rigid body. The two extreme cases correspond to either the entire structure or each atom being a rigid body. It is up to the user to define these rigid bodies.

Scoring Function

The scoring function for a given probe structure P is

$$E = w_1 \cdot E^{CCF}(P) + w_2 \cdot E^{SC}(P) + w_3 \cdot E^{NB}(P), \quad (2)$$

where $E^{CCF}(P)$ quantifies the fit between the probe density, ρ^{P}, and the density map, ρ^{EM}; $E^{SC}(P)$ quantifies the stereochemistry of the model; and $E^{NB}(P)$ quantifies the nonbonded atom-atom contacts. The weights w_1, w_2, and w_3 determine the relative importance of the corresponding terms.

$E^{CCF}(P)$ is defined as the negative sum of crosscorrelation coefficients (CCFs) between the density map and the rigid bodies. For a rigid body b_l, CCF is

$$\mathrm{CCF}(b_l) = \frac{\sum_{i \in Vox(b_l)} \rho^{EM}_i \left(\sum_{j=1}^{N} \rho^{P}_{i,j}\right)}{\sqrt{\sum_{i \in Vox(b_l)} (\rho^{EM}_i)^2 \sum_{i \in Vox(b_l)} \left(\sum_{j=1}^{N} \rho^{P}_{i,j}\right)^2}}, \quad (3)$$

where $Vox(b_l)$ represents all the voxels in the density grid that are within two times the resolution of the map from any of the atoms of rigid-body b_l; and where the total density of P at grid point i is $\sum_{j=1}^{N} \rho^{P}_{i,j}$.

The gradient of the crosscorrelation term is

$$\vec{F}^{CCF}(b_l) = -\sum_{j \in Atom(b_l)} \frac{\partial \mathrm{CCF}(b_l)}{\partial \vec{r}_j}, \quad (4)$$

where

$$\frac{\partial \mathrm{CCF}(b_l)}{\partial \vec{r}_j} = A\frac{Z_j}{\sqrt{2\pi}\sigma} \sum_{i \in Vox(b_l)} \rho^{EM}_i \left(\frac{\vec{r}_j - \vec{r}_i}{\sigma}\right) \exp\left[-\frac{1}{2}\left(\frac{\vec{r}_j - \vec{r}_i}{\sigma}\right)^2\right]. \quad (5)$$

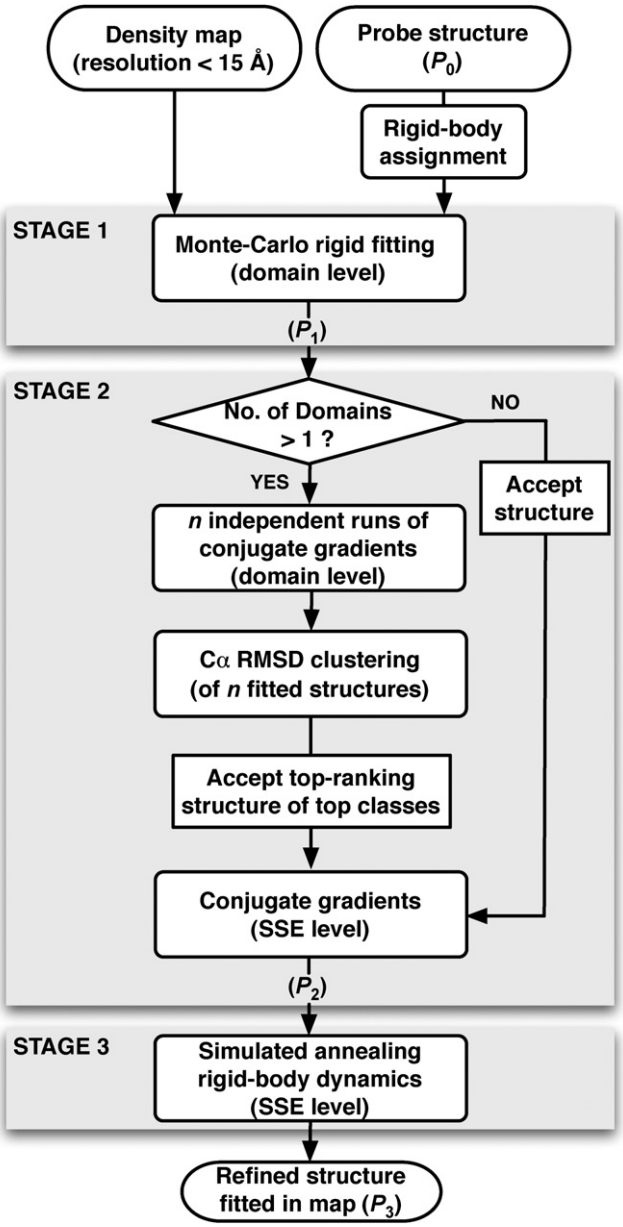

Figure 1. The Flex-EM Protocol for Fitting and Refining an Atomic Structure within Its Cryo-EM Density Map

The inputs to the protocol are an atomic structure and a density map at intermediate resolution (<15 Å). The protocol includes three stages: rigid fitting by an MC optimization (stage 1, MC); refinement by a CG minimization (stage 2, CG); and simulated annealing rigid-body MD (stage 3, MD). For multiple-domain structures, the CG stage is performed n times at the domain level (typically $n = 20$). For single-domain structures, the CG stage is applied only once (i.e., at the SSE level).

j is the atom index and i is the EM density voxel index, both of which are associated with b_l. For computational efficiency, A is considered a constant. It is equal to 10,000 divided by the denominator in Equation 3 for the starting conformation and position. The factor 10,000 was chosen empirically to balance the magnitude of the fitting term relative to the other two terms in the scoring function.

$E^{SC}(P)$ and $E^{NB}(P)$ represent the general conformational preferences of proteins and thus ensure that an optimized structure is physically realistic. $E^{SC}(P)$ restrains the stereochemistry. It is a sum of the harmonic terms of all the chemical bonds, bond angles, dihedral angles, and improper dihedral angles that involve atoms from more than one rigid body. The mean values and force constants were obtained from the CHARMM22 molecular mechanics force field (MacKerell et al., 1998). $E^{SC}(P)$ also includes the two-dimensional (Φ,Ψ) dihedral-angle restraints based on the Ramachandran plot (Fiser et al., 2000). $E^{NB}(P)$ restrains the nonbonded atom-atom contacts. It is a sum of the harmonic lower bounds of all nonbonded atom pairs from different rigid bodies; the lower bound is the sum of the two atomic van der Waals radii, (MacKerell et al., 1998) and the force constant is 400 kcal/mol/Å^2.

The rigid-body gradients of $E^{SC}(P)$ and $E^{NB}(P)$ with respect to the Cartesian coordinates ($\vec{F}_j^{SC}$ and $\vec{F}_j^{NB}$, respectively) are the sums of the gradients for the individual atoms in the rigid body. The gradient of the scoring function is the sum of all three gradient types with the corresponding weights (Equation 2).

Optimization Protocol

The optimization of the scoring function positions, orients, and refines the initial structure so that it satisfies the conformational preferences and fits the density map. We apply a heuristic hierarchical optimization protocol that includes both rigid-body fitting and conformational refinement (Figure 1). The protocol consists of three stages. In the first stage, only the crosscorrelation with the cryo-EM map is optimized by rigid fitting of the whole structure or its domains, using a Metropolis Monte-Carlo (MC) method. The conformational refinement is performed in the second and third stages, with a conjugate-gradients (CG) minimization and a simulated annealing rigid-body molecular dynamics (MD) protocol, respectively. During the refinement, the coordinates of the rigid bodies into which the structure is dissected are displaced in the direction that maximizes their crosscorrelation with the cryo-EM density map and minimizes the violations of the stereochemical and nonbonded terms (Figure 1). As the optimization progresses and the value of the scoring function decreases, we divide the structure into progressively smaller rigid bodies. The rigid bodies can be manually assigned by the user at any stage of the optimization. Here, to make our benchmark automated, the structure is first optimized at the domain level (i.e., the rigid bodies correspond to the domains and the individual atoms that connect the domains), followed by the SSE level (i.e., the rigid bodies correspond to the secondary structure elements in the initial structure and the individual atoms that connect them) (Figure 1).

Stage 1, MC: Rigid Fitting with an MC Method

In the first stage of optimization, the user begins by deciding whether to fit the whole initial probe structure (P_0) or any of its domains independently (with the linkers absent). The corresponding rigid bodies are then placed randomly (or in a specified position) in the density map. Next, the rigid-body positions and orientations are optimized independently in 200 steps of a Metropolis MC optimization protocol using Mod-EM (the *density.grid_search* method in MODELLER 9.0) (Topf et al., 2005). The scoring function at this stage includes only the $E^{CCF}(P_0)$ term; it does not include the stereochemical and nonbonded terms

($w_1 = 1$, $w_2 = 0$, $w_3 = 0$). Therefore, if multiple domains are fitted independently, the resulting structure P_1 can have clashing atoms between the domains. Finally, the linkers that connect the domains are added as follows. Each linker in the initial structure (P_0) is cut at its middle residue. Each of the P_0 domains attached to a half-linker is then superposed onto the corresponding domain in P_1, to obtain the complete P_1 structure including all domains and linkers. Although the linkers in P_1 are generally grossly distorted at their midpoint, they are refined in the next stage.

Stage 2, CG: Conformational Refinement with CG Minimization

In the second stage, we perform a CG refinement of P_1. If the probe structure contains more than one domain, the refinement is performed first at the domain level. A set of 20 random initial structures is obtained from P_1 by rotating and translating each rigid body by a random value ranging from 0° to 30° and from −10 Å to 10 Å, respectively; the user has an option not to randomize the position of a specific rigid body. A CG minimization (Shanno and Phua, 1980) of each of the randomized initial structures is then performed in six iterations, with each iteration progressively increasing the three weights for the individual terms in the scoring function from 0 or small values to 1. Each iteration terminates after 200 CG steps or when the maximum atomic shift is less than 0.01 Å. Next, solutions are clustered in an iterative manner as follows. The structure with the lowest score seeds the first cluster. All the structures with Cα root-mean-square deviation (rmsd) less than 3.5 Å from the seed structure are included in the cluster. The seeding procedure is repeated for the remaining structures until all structures are clustered. The best-scoring structure in each of the top five clusters is then optimized at the SSE level, using the six-iteration CG protocol described above. The structure with the best value of the scoring function is P_2. If the probe structure is composed of a single domain, the six-iteration CG protocol is applied to P_1 only once at the SSE level, to get the refined structure P_2.

Stage 3, MD: Refinement with Simulated Annealing Rigid-Body MD

In the third stage, we optimize P_2 by refining positions and orientations of its rigid bodies with a simulated annealing rigid-body MD protocol (Goldstein, 1980; Brooks et al., 1988). We use the same rigid-body definition as in the final level of stage 2 (i.e., the SSE level) and the same scoring-function weights ($w_1 = w_2 = w_3 = 1$). The state of each rigid body is specified by the position of the center of mass, $\vec{r}_{com}$, and an orientation quaternion, q. At each step, the forces on each atom are summed to give a total force on the center of mass and a torque on the body. $\vec{r}_{com}$ is then updated using a standard Verlet integrator and q is updated by converting the torque to a quaternion angular acceleration. Three cycles of 5600 simulated annealing MD steps are performed (gradually increasing the temperature from 0K to 1000K and decreasing it back to 0K). The optimization is terminated if the change in CCF is < 0.001. Finally, to "relax" the structure, we perform 200 CG steps with $w_1 = w_2 = w_3 = 1$ and 200 CG steps with $w_1 = 0$ and $w_2 = w_3 = 1$, resulting in P_3.

Applicability of the Method

Flex-EM is an automated method for refining experimentally determined atomic structures that undergo conformational changes in the context of the assembly cryo-EM map as well as comparative models that suffer from modeling errors. The assignment of the rigid bodies is given as an input to the program at each step of the optimization protocol. For a 200 residue target sequence and a density map at ~10 Å resolution, the typical running times are less than 1 min for the MC stage on one CPU, less than 4 hr for the CG stage on 20 CPUs, and less than 12 hr for the MD stage on one CPU. The Flex-EM software and the benchmark (below) are available at http://salilab.org/Flex-EM/.

Benchmark

To test the fitting and refinement protocol, we created a benchmark of 13 proteins in a nonnative conformation (P_0) and a density map (ρ^{EM}) calculated from the corresponding native structure. To make the test more realistic, we generated the nonnative conformations with the aid of comparative protein structure modeling (Eswar et al., 2007). This procedure resulted in variations in domain orientations, loop conformations, positions of secondary structure elements, as well as distortions and shifts of secondary structure elements. Ten of the 13 proteins were selected from the molecular motions database (Flores et al., 2006) that stores experimentally determined structures of macromolecules in two distinct conformations. For each of the ten proteins, one conformation was defined as the "target." Using the DBAli database (Marti-Renom et al., 2007), a structure of a protein that is closer to the other conformation was defined as the "template." The remaining three target-template pairs were selected from known sets of homologs containing domains that interact through different surfaces (Han et al., 2006) and DBAli. Next, we calculated sequence-structure alignments and built corresponding models using the *align* method and *automodel* class in MODELLER 9.0 (Eswar et al., 2007; Sali and Blundell, 1993), respectively. The resulting 13 comparative models were used in the benchmarking as the initial probe structures in the nonnative conformation (P_0). The native structures were used to calculate the corresponding "native" density maps (ρ^{EM}) at the resolution of 10 Å with grid spacing of 1 Å/voxel. To minimize bias, the maps were not produced with our program, but with *pdb2vol* in Situs (Wriggers et al., 1999), which uses a different Gaussian smoothing technique. Furthermore, the use of comparative modeling for building the initial benchmark structures introduced test cases that are more challenging for refinement than experimentally determined atomic structures (which tend to be less distorted). By construction, the native structure has the best value of CCF among all structures produced during the optimization process.

In summary, the benchmark consists of eight single-domain and five two-domain protein structures in a nonnative conformation (probes), generated based on homologous structures sharing between 26% and 52% sequence identity (Tables 1 and 2). The average number of residues per structure is 211. The benchmark contains representatives from all major fold classes (i.e., α, β, α + β, and α/β). Domain assignment was based on the domain definition in the Pfam database (Bateman et al., 2004). Secondary structure elements were determined based on the initial probe structures with DSSP (Kabsch and Sander, 1983).

Table 1. Single-Domain Protein Structures Fitted and Refined within Their Native Density Maps at 10 Å Resolution

Probe[a] (PDB ID Code, Range)	Template[a] (PDB ID Code, Range)	Fold Type	% Sequence Identity[a]	$-E^{CCF}$[b]		E^{SC}[b]		E^{NB}[b]		OS[c] (Å, °)	Cα Rmsd[d] (Å)				NO3.5[d] (%)		
				P_0[e]	P_3[e]	P_0	P_3	P_0	P_3	P_3	P_0	P_3	P_{Best}[e]	P_{Min}[e]	P_0	P_3	P_{Best}
1akeA, 1–213	1dvrB, 5–217	α/β	46	0.938	0.969	803	796	1.5	3.4	0.4, 3.1	4.5	2.2	2.2	0.9	75	91	93
1c1xA, 8–345	1gtmA, 31–407	α/β	30	0.951	0.977	1482	1187	5.8	7.9	0.0, 3.3	6.6	4.6	4.5	1.4	53	66	67
1cll, 4–146	2ggmB, 25–168	α	52	0.875	0.958	531	490	1.1	3.4	0.2, 3.6	5.0	3.1	3.0	0.6	51	74	78
1g5yD, 231–442	3erdA, 310–534	α	30	0.933	0.945	768	677	4.4	2.2	0.4, 6.1	5.4	5.1	4.9	1.5	56	67	69
1jxmA, 531–711	1ex7A, 531–711	α/β	33	0.890	0.967	837	754	1.6	4.8	0.4, 5.7	5.4	3.3	3.3	0.9	43	80	82
1ozoA, 1–84	1a03B, 2–83	α	42	0.940	0.966	250	242	1.0	2.9	0.3, 6.4	4.7	4.1	4.0	1.3	56	71	70
1uwoA, 1–90	1k9pA, 3–89	α	41	0.948	0.960	243	308	0.4	23.2	0.3, 5.8	4.7	4.0	4.0	1.3	69	70	73
1cczA, 1–170	1hnf, 1–170	β	37	0.934	0.973	948	892	323.0	11.9	0.2, 8.3	5.2	5.1	4.9	1.2	64	66	74
Average			39	0.926	0.964	733	668	42.7	7.5	0.3, 5.3	5.2	3.9	3.8	1.1	58	73	76

[a] Probe is the structure being refined. The initial coordinates of each probe structure were generated based on the "template" structure using comparative modeling with MODELLER 9.0 (Sali and Blundell, 1993). The target-template sequence identity is calculated from their sequence alignment.
[b] $-E^{CCF}$, E^{SC}, and E^{NB} are the three terms of the scoring function with equal weights ($w_1 = w_2 = w_3 = 1$; Equation 2): the crosscorrelation coefficient (CCF) between a probe structure and the native density map (which is multiplied by 10,000 during refinement; Equations 3 and 5, respectively), the stereochemical restraints, and the nonbonded interactions restraints.
[c] OS and DOS are the orientation and domain-orientation scores, respectively. For the multidomain proteins, the OS score was calculated for the N-terminal domain only.
[d] Cα rmsd is the root-mean-square deviation between the Cα atoms of the probe structure and their corresponding atoms in the native structure, and NO3.5 and NO5.0 are the percentages of Cα atoms in a probe structure that are positioned within 3.5 and 5.0 Å, respectively, from their corresponding atoms in the native structure. These scores are calculated upon superposition of the initial or a refined structure onto the corresponding native structure using a rigid-body least-squares minimization.
[e] P_0, P_3, P_{Best}, and P_{Min} refer, respectively, to the initial structure (before the refinement); the final structure (following the MC, CG, and MD refinement protocol); the structure with the best score (for Cα rmsd, NO3.5, and NO5.0) found in the simulation; and the best-possible structure (based on MinRmsd for fitting and refining a model with secondary structure elements as rigid bodies). The corresponding best-possible NO3.5 is always equal or higher than 97% (99% on average) and NO5.0 is 100%.

Measures of Model Accuracy

Model accuracy is measured through two types of scores: (1) a rigid-body shift and rotation of a fitted component relative to its correct position in the density (i.e., the orientation score [OS] and the domain-orientation score [DOS]); and (2) a distortion of the conformation of the probe structure relative to the native structure (i.e., the Cα rmsd and native overlap [NO]).

Orientation Score

The OS quantifies the difference between the orientation and position of a given rigid body fitted in the density, and the orientation of the equivalent rigid body in the native structure, which by construction is positioned correctly in the map. To calculate the score, we first translate the center of mass of the rigid body onto the center of mass of the equivalent rigid body in the native structure. The first component of the OS score, *dist* (Å), is then defined as the magnitude of the corresponding translation vector. We then rotate the rigid body in the refined structure to optimally superpose it onto the equivalent rigid body in the native structure (using the *superpose* method of MODELLER 9.0). The second component of OS, *ang* (°), is then defined as the angle of rotation.

Domain-Orientation Score

The DOS is similar to the OS, except that it is used for multidomain proteins. It quantifies the difference between the relative orientations and positions of two rigid-body domains in the refined structure and the two equivalent rigid bodies in the native structure. First, the two compared structures are brought into the same frame of reference by superposing the first pair of equivalent domains. Next, the *dist* and *ang* scores are calculated for the second rigid body using the same procedure as for OS.

Rmsd and NO

Cα rmsd is calculated between the Cα atoms of a structure (i.e., the initial structure or a structure being refined) and the corresponding atoms in the native structure. NO3.5 and NO5.0 of the refined structure are the percentage of its Cα atoms that are within 3.5 and 5.0 Å of the corresponding atoms in the native structure, respectively. Both scores are calculated upon superposition of the refined structure onto the corresponding native structure using a rigid-body least-squares minimization, as implemented in the *superpose* method of MODELLER 9.0.

We also calculated a "minimal" Cα rmsd (MinRmsd) for each structure, corresponding to the best-possible model, given that

Table 2. Two-Domain Protein Structures Fitted and Refined within Their Native Density Maps at 10 Å Resolution

Probe[a] (PDB ID Code, Range)	Template[a] (PDB ID Code, Range)	Fold Type	% Sequence Identity[a]	$-E^{CCF}$[b]		E^{SCb}		E^{NBb}		OS (Å, °)		DOS[c] (Å, °)		Cα Rmsd[d] (Å)				NO3.5[d] (%)				NO5.0[d] (%)			
				P_0[e]	P_3[e]	P_0	P_3	P_0	P_3	P_3		P_0	P_3	P_0	P_3	P_{Best}[e]	P_{Min}[e]	P_0	P_3	P_{Best}		P_0	P_3	P_{Best}[e]	
1ffgAB, 2–226	1u0sAY, 2–260	α/β, α + β	26	0.959	0.970	896	796	13.2	5.3	0.6, 15.0		4.8, 123.3	5.0, 6.4	9.1	3.6	3.5	1.1	35	72	76		60	88	89	
1iknA, 19–273	1svc, 43–335	β	46	0.913	0.961	1300	1110	1.9	9.7	0.8, 1.0		12.8, 100.7	2.5, 46.8	10.4	4.4	3.9[f]	0.8	57	57	65		62	78	80	
1a45A, 1–172	1blb, 1–172	β	35	0.771	0.973	996	965	3.8	3.3	1.3, 161.2		37.8, 168.9	39.3, 141.0	28.9	12.2	11.5	0.9	47	24	52		50	41	55	
1ckmA, 60–319	1p16, 44–366	α + β, β	32	0.903	0.960	1311	1414	3.7	6.1	0.7, 7.9		10.7, 41.6	4.0, 9.8	8.3	6.6	6.5	1.2	42	62	70		62	74	81	
1hrdC, 22–446	1hwzA, 7–491	α/β, α/β	28	0.925	0.961	2313	2151	13.2	17.5	1.6, 8.0		6.7, 23.9	3.4, 10.7	8.2	4.9	4.9	0.9	49	67	67		65	83	84	
Average			33	0.894	0.965	1363	1287	1.0	8.4	1.0, 38.6		14.6, 91.7	10.9, 42.9	13.0	6.3	6.1	1.0	46	57	66		59	73	77	

[a] Probe is the structure being refined. The initial coordinates of each probe structure were generated based on the "template" structure using comparative modeling with MODELLER 9.0 (Sali and Blundell, 1993). The target-template sequence identity is calculated from their sequence alignment.

[b] $-E^{CCF}$, E^{SC}, and E^{NB} are the three terms of the scoring function with equal weights ($w_1 = w_2 = w_3 = 1$; Equation 2): the crosscorrelation coefficient (CCF) between a probe structure and the native density map (which is multiplied by 10,000 during refinement; Equations 3 and 5, respectively), the stereochemical restraints, and the nonbonded interactions restraints.

[c] OS and DOS are the orientation and domain-orientation scores, respectively. For the multidomain proteins, the OS score was calculated for the N-terminal domain only.

[d] Cα rmsd is the root-mean-square deviation between the Cα atoms of the probe structure and their corresponding atoms in the native structure, and NO3.5 and NO5.0 are the percentages of Cα atoms in a probe structure that are positioned within 3.5 and 5.0 Å, respectively, from their corresponding atoms in the native structure. These scores are calculated upon superposition of the initial or a refined structure onto the corresponding native structure using a rigid-body least-squares minimization.

[e] P_0, P_3, P_{Best}, and P_{Min} refer, respectively, to the initial structure (before the refinement); the final structure (following the MC, CG, and MD refinement protocol); the structure with the best score (for Cα rmsd, NO3.5, and NO5.0) found in the simulation; and the best-possible structure (based on MinRmsd for fitting and refining a model with secondary structure elements as rigid bodies). The corresponding best-possible NO3.5 is always equal or higher than 97% (99% on average) and NO5.0 is always 100%.

[f] This score is for a structure found in the second cluster obtained at the CG stage.

the secondary structure elements are treated as rigid bodies. The best-possible model has all loop atoms overlapping perfectly with the equivalent native positions and each rigid body in the initial probe structure (an α helix or a β strand, as determined by DSSP) superposed independently onto the corresponding region in the native structure.

Accuracy of the Refined Structures

Single-Domain Proteins

The optimization protocol was able to accurately fit and refine all eight single-domain benchmark proteins: the average OS was [0.3 Å, ~5.3°] and both Cα rmsd and NO3.5 were improved relative to their initial values (Table 1). This improvement is correlated with the increase in CCF. The values of the stereochemical and nonbonded terms ($E^{SC}[P]$ and $E^{NB}[P]$, respectively) were either reduced or increased by less than a factor of 3 (E^{NB} in 7 out of 8 structures and E^{SC} in 8 out of 8). The average Cα rmsd was reduced from 5.2 to 3.9 Å. Given the average MinRmsd of 1.1 Å, the average Cα rmsd corresponds to ~32% (i.e., [5.2 − 3.9]/[5.2 − 1.1]) of the maximum possible improvement. The average NO3.5 improved from 58% to 73%, which is 37% of the maximum possible improvement (the average maximum possible NO3.5 is 99%). According to the Ramachandran plots of the final structures (calculated using MOLprobity; Lovell et al., 2003), the average percentage of residues in the allowed (Φ,Ψ) dihedral-angle regions is 98.9% (e.g., see Figure S1 in the Supplemental Data available with this article online).

Two-Domain Proteins

For all five final structures (P_3) of the two-domain proteins, the Cα rmsd was better than for the initial structures, correlating with the increase in CCF (Table 2). For these proteins, the values of $E^{SC}(P)$ and $E^{NB}(P)$ were either reduced or increased by less than a factor of 2 (E^{SC} in 5 out of 5 structures and E^{NB} in 4 out of 5). The average Cα rmsd was reduced significantly, from 13.0 to 6.3 Å, which is ~56% of the maximum possible improvement (given that the average MinRmsd is 1.0 Å). The average number of residues in the allowed regions of the Ramachandran plot was above 98% in all structures (e.g., Figure S1). The average NO3.5 and NO5.0 increased from 46% to 57% and from 59% to 73%, respectively.

A closer look at the different scores of the five structures reveals that although the Cα rmsd has improved significantly for all proteins, NO3.5 improved for three out of the five proteins (Protein Data Bank [PDB] ID codes: 1ffgAB, 1ckmA, and 1hrdC) and NO5.0 for four out of five (1ffgAB, 1iknA, 1ckmA, and 1hrdC). The latter result is also reflected in OS and DOS (Table 2). The final values of both scores for 1ffgAB, 1ckmA, and 1hrdC were better than [5 Å, 15°], respectively. For 1iknA, the DOS *ang* score could be reduced further (i.e., although the orientation between the domains in the final structure is more accurate than in the initial structure, it is still far from the orientation in the native structure). For 1a45A, however, both final OS *ang* and DOS scores were high, showing that the protein was not fitted correctly in the map.

Sample Model Optimization

For all initial probe structures except 1a45A (for which the two domains were fitted separately), the first stage of the optimization protocol (rigid fitting by MC) was able to identify the approximate position and orientation in the 10 Å resolution native density

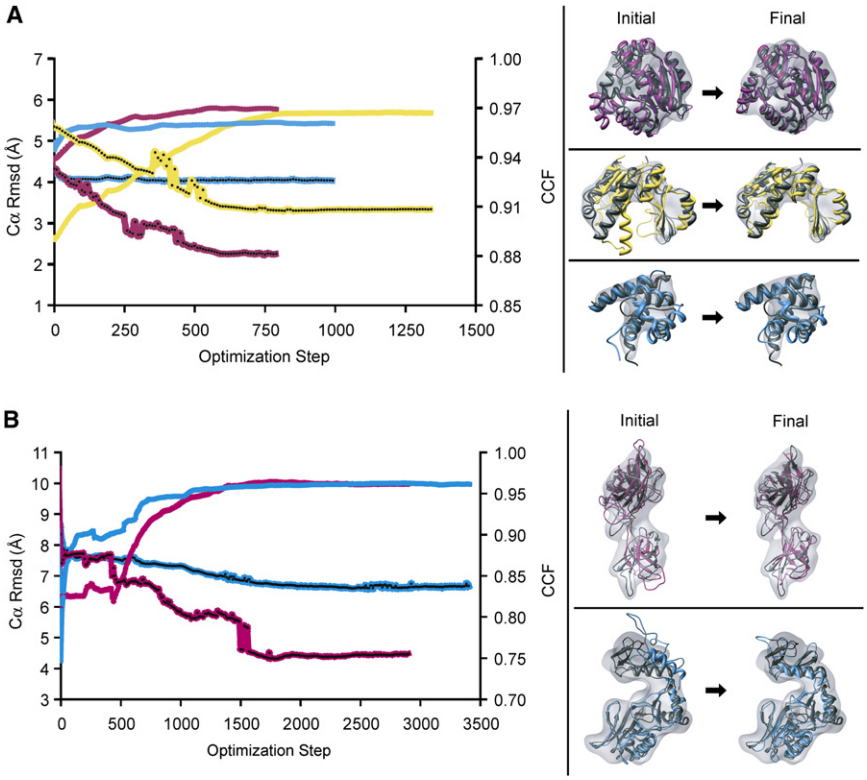

Figure 2. Examples for Combined Fitting and Structure Refinement

(A) CCF (plain lines) and Cα rmsd (lines with black dots) of three nonnative single-domain structures during rigid-body MD refinement within their native 10 Å resolution density maps (MD stage, $P_2 \rightarrow P_3$; see Figure 1): PDB ID codes: 1jxmA (yellow), 1akeA (magenta), and 1uwoA (light blue). The scores were recorded every ten steps.

(B) CCF (plain lines) and Cα rmsd (lines with dots) of two nonnative two-domain structures during the CG and rigid-body MD refinement within their native 10 Å resolution density maps (CG and MD stages, $P_1 \rightarrow P_2 \rightarrow P_3$; see Figure 1): PDB ID codes: 1iknA (magenta) and 1ckmA (light blue). The scores were recorded every ten steps. P_2 starts at 1310 steps for 1iknA and at 1020 steps for 1ckmA.

maps (as reflected in the OS score; Tables 1 and 2). Therefore, we focus on stages 2 and 3 (CG and MD) using five sample proteins, each of which represents a different refinement scenario (Figure 2). Generally, most improvement for the two-domain proteins was achieved in the domain-level CG minimization (CG stage), whereas the SSE-level MD was most beneficial for the single-domain proteins (MD stage). This result is a consequence of the differences between the initial and native structures being dominated by domain and secondary structure element repacking for the two- and single-domain proteins, respectively.

For the three sample single-domain proteins, the improvement in the accuracy of the structure during the MD stage (P_2) was highly correlated with the improvement in CCF (Figure 2A) (i.e., Cα rmsd of the probe structure decreased with the increase in CCF and reached a minimum when CCF reached a maximum). For 1akeA, Cα rmsd was reduced from 4.5 to 2.2 Å, pushing the final structure close to the native structure (MinRmsd is 0.9 Å). For 1jxmA, Cα rmsd also decreased significantly, from 5.4 to 3.3 Å, with a MinRmsd of 0.9 Å (Movie S1). However, there is room for further improvement, especially in the loop regions and to a smaller degree in the orientations of secondary structure elements. For 1uwoA, the refinement process improved CCF only slightly, primarily as a result of an incorrect assignment of some of the rigid bodies, caused by a misplacement of secondary structure elements in the probe structure with respect to the native structure: helix 43–46 corresponds to a loop in the native structure, and helices 50–54 and 56–62 correspond to helix 51–59; these mistakes are because of errors in the comparative modeling of 1uwoA based on the 1k9pA template. As a result of these errors, Cα rmsd was reduced only from 4.7 to 4.0 Å and NO3.5 improved only slightly, from 69% to 70%.

The sample two-domain proteins illustrate the impact of correct and incorrect secondary structure element assignments. For 1iknA, the decrease in Cα rmsd was highly correlated with the increase in CCF throughout the CG and MD stages (P_1 and P_2; Figure 2B; Movie S2). Cα rmsd improved from 10.4 Å for the initial structure to 4.4 Å for the final structure, with a MinRmsd of 0.8 Å. In contrast, Cα rmsd of 1ckmA slightly increased during the MD stage (P_2), despite the small increase in CCF. This contrasting result can be attributed to two different problems with secondary structure element assignments. First, there were missing secondary structure elements in the refined structure (e.g., its loop 196–209 corresponds to an α helix in the native structure and loop 255–266 to a β sheet). In these regions, the atoms were fitted individually, which turned out to be too large a burden for the optimizer to handle correctly. Second, assignment of the secondary structure elements to incorrect segments of the sequence led to the opposite situation in which the atoms in loops that should have been fitted individually were in fact fitted as rigid helices and strands (e.g., helices 178–180 and 182–186 in the structure being refined are loops in the native structure). As a result of the secondary structure element misassignments, Cα rmsd improved only marginally from 8.3 to 6.6 Å, with MinRmsd of 1.2 Å.

The Effect of Map Resolution on Model Accuracy

For four proteins in the benchmark, we tested Flex-EM with "native" density maps simulated at a range of resolutions from 4 to 14 Å (Figure 3). In all cases, both Cα rmsd and NO3.5 of the final structures were better than those for the initial structures, at all tested resolutions, from 4 to 14 Å. For all four tests, NO3.5 of the final structure at 4 Å resolution was higher than 87% and Cα rmsd was lower than 2.5 Å. Furthermore, for 1jxmA, 1akeA, and 1cll, the results suggest a strong correlation between the accuracy of the final structure and the map resolution (Pearson correlation coefficient [R^2] > 0.9). However, for 1uwoA, the correlation is weak as a result of the incorrect rigid-body assignment (as described above).

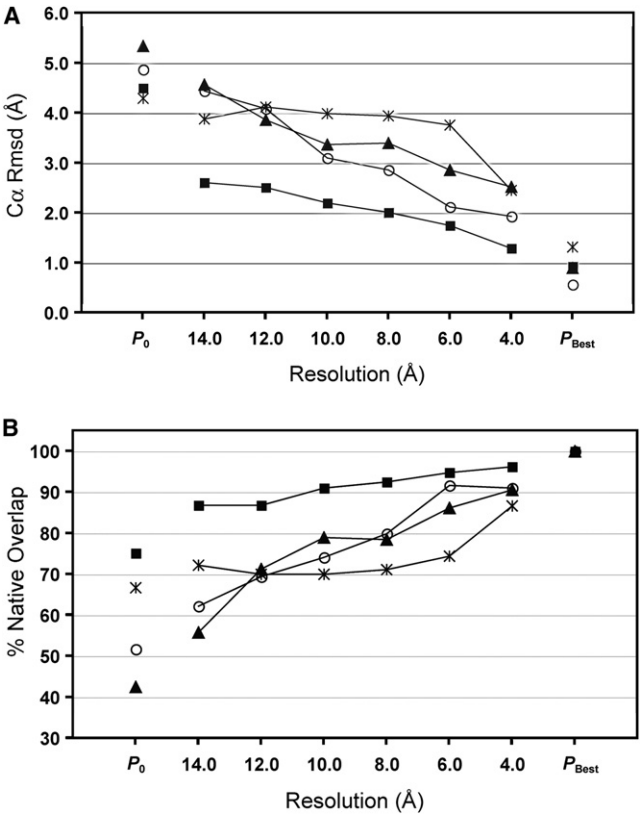

Figure 3. Accuracy of the Final Structures of Four Single-Domain Proteins

PDB ID codes 1akeA (square), 1cll (circle), 1jxmA (triangle), and 1uwoA (asterisk), refined in their corresponding density maps at different resolutions (ranging between 4 and 14 Å).

(A) The Cα rmsd of the initial structure (P_0), final structure (at each resolution), and the best-possible structures (P_{best}) from the native structures.

(B) The native overlap within 3.5 Å (NO3.5).

P_0 and P_{Best} refer to the initial structure (prior to the refinement) and the structure based on which MinRmsd was calculated (Results), respectively.

Modeling Conformational Changes Using Experimentally Determined Cryo-EM Maps

To test the refinement of atomic structures using maps with noise not captured in the simulated maps, we applied Flex-EM to two multidomain proteins with experimentally determined maps: a monomer of the bacterial chaperonin complex GroEL and the bacterial elongation factor EF-Tu. For GroEL, the initial structure was a comparative model based on 63% sequence identity to a monomer in the *bound* state (GroEL-GroES-ADP) of a homologous archaebacterial complex, Thermosome (PDB ID code: 1we3B; Shimamura et al., 2004). Density maps for Flex-EM were segmented with Chimera (Goddard et al., 2007) from cryo-EM maps of the double-ring GroEL complex in the *unbound* state at 11.5 Å (EMDB code 1080; Ludtke et al., 2001) and 6.0 Å (EMDB code: 1081; Ludtke et al., 2004). For EF-Tu, the initial structure was a comparative model based on 55% sequence identity to a mitochondrial homolog complexed with GDP (PDB ID code: 1d2eA; Andersen et al., 2000). The density map was segmented from the 9.0 Å resolution cryo-EM map of *Escherichia coli* 70S ribosome complexed with tRNA-EF-Tu-GDP-kirromycin (EMDB code 1055) (Valle et al., 2003). Both the GroEL monomer and EF-Tu were assigned three domains each, based on SCOP (Murzin et al., 1995). The secondary structure elements in the initial structures were determined with DSSP (Kabsch and Sander, 1983). Each of the initial structures was then fitted and refined in the corresponding density (GroEL in the 11.5 and 6.0 Å resolution maps and EF-Tu in the 9.0 Å resolution map).

In the first stage (MC), we fitted the equatorial domain (I) of GroEL jointly with the small intermediate domain (II), and the apical domain was treated as a separate rigid body (III). In the case of EF-Tu, domain I was fitted individually and domains II and III jointly. Next, we refined each of the structures in the following order: domain-level CG (three domains), SSE-level CG, and SSE-level MD. To evaluate the accuracy of the refined structures at each stage of the protocol, we compared them to the corresponding known structures: an unbound GroEL (2.8 Å resolution) (PDB ID code: 1oelD; Braig et al., 1995) and an *E. coli* EF-Tu-GDP-kirromycin (3.4 Å resolution) (PDB ID code: 1ob2).

In all three cases, the refined structures were significantly more accurately positioned and modeled than the initial structure (Figure 4; Table 3). The largest improvement in the accuracy of the structures occurred during the MC or CG stage, in correlation with the increase in CCF. However, the largest improvement in the stereochemical and nonbonded terms occurred during the CG stage. For GroEL, Cα rmsd was reduced from 16.2 to 3.8 and 1.9 Å in the 11.5 and 6.0 Å maps, respectively. For EF-Tu, Cα rmsd was reduced from 28.6 to 4.0 Å. In the 6.0 Å map of GroEL, the final NO3.5 was 96% (i.e., almost all atoms in the structure were within 3.5 Å from the crystal structure).

DISCUSSION

Method

We present a method for flexible fitting of atomic structures of assembly components into the cryo-EM density map of the whole assembly (Flex-EM). The method, which is applicable to both experimental structures and models, outputs the position and orientation of the component in the density map (MC stage) as well as its refined coordinates (CG and MD stages) (Figure 1). The optimization is applied to the component rigid bodies, specified by the user, and is driven by the quality of their fit into the density (CCF) as well as stereochemistry and nonbonded interactions. The method is fully automated while also allowing user intervention in the fitting process, including assigning and refining rigid bodies. For example, some components may be fixed at certain positions while others are refined in their context.

Conceptually, Flex-EM is similar to RSRef, a real-space refinement method that was originally developed for X-ray crystallography (Chapman, 1995) and has recently been adopted to cryo-EM and applied to maps at resolutions better than 20 Å (Chen and Champman, 2001; Chen et al., 2003). RSRef uses torsion-angle MD to improve the fit of an atomic model to a density map by optimizing a scoring function that also includes the stereochemical and nonbonded interaction terms. However, there are significant differences between the two methods. RSRef was designed to refine an atomic model within a cryo-EM density map once it is already fitted in the approximate position in the map. Flex-EM, in contrast, performs both the initial approximate fitting of the model in the map and its further

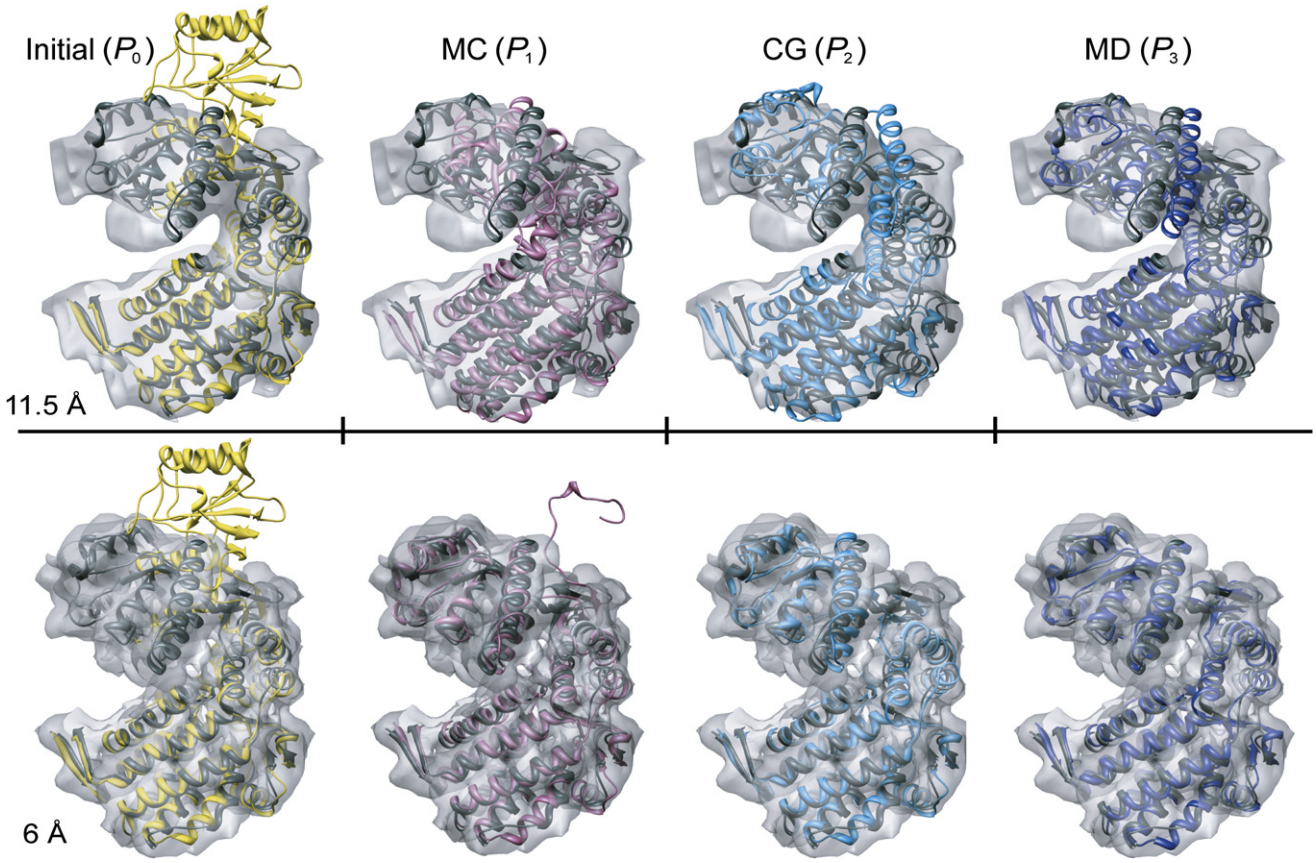

Figure 4. Fitting and Refining a GroEL Monomer

A comparative model of a monomer of the bacterial GroEL in the GroES-ADP-bound conformation (based on the archaeal homolog, PDB ID code: 1we3B) fitted and refined within the segmented experimental cryo-EM maps of the unbound conformation, determined at 11.5 Å (top) and 6.0 Å (bottom) resolution. The known native structure (PDB ID code: 1oelD) is shown as a reference in gray. The input comparative model is shown in yellow (P_0). The structures shown in pink, light blue, and dark blue are the final structures resulting from the MC stage (P_1), CG stage (P_2), and MD stage (P_3) of the optimization protocol, respectively. The initial Cα rmsd (left) is 16.2 Å from the native structure, and the final Cα rmsd's (right) are 3.8 and 1.9 Å in the 11.5 and 6.0 Å resolution maps, respectively. The figures were generated with Chimera (Pettersen et al., 2004).

refinement. During the refinement, Flex-EM, like RSRef, improves the positions and orientations of the domains. Flex-EM has also been shown here to improve the positions and orientations of secondary structure elements within the domains (Table 1). The ability to treat the secondary structure elements of the structure being refined as individual rigid bodies even at ~14 Å resolution is partly because of the inclusion of the two-dimensional (Φ,Ψ) dihedral-angle term in the scoring function. This term allows better modeling of the loops or linkers connecting the rigid bodies, whether domains or secondary structure elements, resulting in high-quality Ramachandran plots for the refined models (Figure S1). Lastly, our method is based on satisfaction of spatial restraints in real space. Therefore, it can be combined relatively easily with additional restraints that provide information about the configuration and conformation of the assembly components, such as footprinting, chemical crosslinking, and various bioinformatics analyses (Russell et al., 2004).

Previously, normal mode analysis relying on the scoring function corresponding to CCF was successfully applied to explore deformations of the structure in the search for an optimal solution (Tama et al., 2004; Ma, 2005). The approach relies on the assumption that a few of the lowest-frequency modes are sufficient to represent the changes needed to refine the initial structure. Although this assumption is often warranted, it does not always apply. For example, a ligand can "stretch" the protein in ways that involve higher-frequency modes (Petrone and Pande, 2006). In addition, high-frequency deformations that occur because of modeling errors may not be corrected using this method. The advantage of the current method implemented in Flex-EM is that it can in principle produce any kind of molecular deformations (including high-frequency), such as shear and hinge movements of whole domains, subdomains, and secondary structure elements as well as loop distortions and movements.

Accuracy of the Modeled Structures

Almost all existing cryo-EM flexible fitting methods can lead to artificial distortions in the refined atomic structure. One of the advantages of our method is the ability to optimize the fit of the structure within the density map while maintaining correct stereochemistry. This goal is achieved during the refinement stages of the optimization (CG and MD) by optimizing a scoring function

Table 3. Nonnative Structures of GroEL Monomer and EF-Tu Fitted and Refined in Their Experimental Cryo-EM Maps

Protein Name, Resolution (Å), Domain Definition	Probe[a]	$-E^{CCF}$[b]	E^{SC}[b]	E^{NB}[b]	OS[c] (Å, °)	DOS1[c] (Å, °)	DOS2[c] (Å, °)	Cα Rmsd[d] (Å)	NO3.5[d] (%)	NO5.0[d] (%)
GroEL, 11.5, I (3–138, 409–527), II (190–373), III (139–189, 374–408)	P_0 (Initial)	0.771	1,759	2.6	2.0, 7.8	7.2, 30.5	12.3, 106.7	16.2	49	53
	P_1 (MC)	0.824	192,786	896.0	2.0, 7.8	7.2, 30.5	12.6, 63.1	13.0	49	53
	P_2 (CG)	0.848	175,950	22.7	4.3, 5.6	7.1, 26.2	3.8, 19.4	6.0	50	80
	P_3 (MD)	0.894	1,517	3.3	2.3, 2.6	6.7, 26.3	2.3, 12.1	3.8	61	86
GroEL, 6.0, I (3–138, 409–527), II (190–373), III (139–189, 374–408)	P_0 (Initial)	0.554	1,759	2.6	0.3, 2.5	7.2, 30.5	12.3, 106.7	16.2	49	53
	P_1 (MC)	0.673	500,194	2.6	0.3, 2.5	7.2, 30.5	5.5, 19.9	7.4	75	84
	P_2 (CG)	0.710	1,896	22.6	0.4, 2.3	2.6, 17.4	4.3, 13.5	2.2	90	96
	P_3 (MD)	0.745	1,496	5.9	0.5, 2.1	1.6, 8.6	1.7, 4.3	1.9	96	98
EF-Tu, 9.0, I (1–202), II (203–300), III (301–393)	P_0 (Initial)	0.659	1,629	8.0	1.1, 7.8	24.4, 91.1	0.8, 12.4	28.6	47	48
	P_1 (MC)	0.851	20,393	5472.5	0.5, 2.8	3.4, 14.4	0.8, 12.4	4.4	84	93
	P_2 (CG)	0.856	1,712	13.2	1.3, 7.8	6.2, 7.5	3.1, 9.8	4.1	84	94
	P_3 (MD)	0.867	2,172	9.6	1.7, 5.0	5.4, 5.2	4.5, 8.0	4.0	86	94

[a] Probe is the structure being refined. P_0, P_1, P_2, and P_3 refer to the initial structure and the structures resulting from the MC, CG, and MD stages of optimization protocol, respectively.
[b] $-E^{CCF}$, E^{SC}, and E^{NB} are the three terms of the scoring function with equal weights ($w_1 = w_2 = w_3 = 1$; Equation 2): the crosscorrelation coefficient (CCF) between a probe structure and the native density map (which is multiplied by 10,000 during refinement; Equations 3 and 5, respectively), the stereochemical restraints, and the nonbonded interactions restraints.
[c] OS, DOS1, and DOS2 are the orientation and two domain-orientation scores, respectively (the OS score was calculated for domain I, and DOS1 and DOS2 were calculated for domains I-II and II-III, respectively).
[d] Cα rmsd is the root-mean-square deviation between the Cα atoms of a probe structure and their corresponding atoms in the native structure, and NO3.5 and NO5.0 are the percentages of Cα atoms in a probe structure that are positioned within 3.5 and 5.0 Å, respectively, from their corresponding atoms in the native structure. These scores are calculated upon superposition of the initial or a refined structure onto the corresponding native structure using a rigid-body least-squares minimization.

that is driven by a CCF term but that also includes stereochemical and nonbonded interaction terms.

We demonstrate the ability of the method to improve the accuracy of structures using a benchmark of nonnative structures and their corresponding native density maps at 10 Å resolution. Although the method does not allow us to predict to what extent a given structure can be refined, none of the initial structures became worse in terms of Cα rmsd as a result of its refinement (Tables 1 and 2). On average, Cα rmsd improved from 5.2 to 3.9 Å for the single-domain proteins and from 13.0 to 6.3 Å for the two-domain proteins. NO3.5 increased from 58% to 73% and from 46% to 57%, respectively. Furthermore, the values of the stereochemical and nonbonded terms were either reduced or remained comparable to those in the initial structure (Tables 1 and 2). The final average number of residues in the allowed (Φ,Ψ) regions of the Ramachandran plot was higher than 97.5% for all final structures (Figure S1), indicating that none of the final structures is distorted.

Although the improvement in the accuracy of the benchmark structures was generally high for both single- and two-domain proteins (Tables 1 and 2), the improvement was higher for the single-domain proteins (as reflected in the NO3.5 score). This result can be explained by the nature of the benchmark, considering that the refinement stages of the optimization are primarily dependent on the rigid-fitting stage. For the single-domain proteins, the initial rigid fitting at the domain level was very accurate (as reflected in the OS score), enabling successful refinement at the SSE level (as reflected in the NO3.5 score). For the two-domain proteins, the SSE-level refinement was successful only when the domains were sufficiently accurately positioned in the map by the CG domain-level optimization (as reflected in the DOS, NO3.5, and NO5.0 scores). If, however, there was initially only partial improvement in the domain positions, the subsequent SSE-level refinement failed owing to the inability of the sampling to benefit from the map.

A further indicator of the accuracy of the method was provided by testing it at a range of resolutions between 4 and 14 Å (Figure 4). The method was shown to improve the structures significantly at 6 and 4 Å resolution, decreasing Cα rmsd below 2.5 Å at 4 Å resolution (for 4 of the 4 test cases) and below 3.0 Å at 6 Å resolution (for 3 of the 4 test cases). In addition, for all four proteins, both Cα rmsd and NO3.5 of the final structures were better than those of the initial structures, even at 14 Å resolution, indicating that the method is certainly useful for the refinement of secondary structure elements and loops even at resolutions where secondary structure elements cannot be identified directly from the density. The method might be helpful even at resolutions worse than 14 Å, as long as the rigid bodies are large enough (i.e., not smaller than domains), a possibility that we will test in the future.

To demonstrate the ability of the method to refine atomic structures in more realistic cases, we applied it to experimentally determined cryo-EM maps of GroEL at 6.0 and 11.5 Å resolution (Table 3; Figure 4) and of EF-Tu at 9.0 Å resolution. Despite the noise in these maps that is not present in the simulated maps, the results were similar to the benchmark average. For GroEL, Cα rmsd was reduced from 16.2 to 1.9 and 3.8 Å using the 6.0 and 11.5 Å maps, respectively; for EF-Tu, the improvement

was from 28.6 to 4.0 Å using the 9.0 Å map. In addition, when the initial rigid fitting is accurate and Cα rmsd is thus significantly decreased in the MC stage (GroEL, 11.5 Å and EF-Tu, 9 Å cases), further refinement significantly reduces the distortions in bond distances and angles in the linkers connecting the domains (Table 3). These realistic test cases demonstrate that the method can significantly improve structures with ~12 Å resolution maps, as well as approach atomic resolution with 6 Å resolution maps.

Scoring Function and Sampling

In all tested cases (including GroEL and EF-Tu), the change in CCF is highly correlated with the change in Cα rmsd and NO (Figure 2), and the native structure has the highest score (CCF = 1). Thus, the scoring function of Flex-EM appears to be sufficiently accurate for the current degree of sampling at the tested map resolutions. Correspondingly, the main shortcoming of the current method is its relatively limited sampling. (If the sampling becomes more thorough in the future, the scoring might also become limiting in terms of achieving higher accuracy.) There are two underlying reasons for the limited sampling, as follows.

First, on the way to the approximately "correct" solution there are many structures that have a similar score, especially if they have similar shapes. For example, for 1iknA, the most accurate structure, (which was found in the second cluster at the CG stage of the optimization), did not have the highest CCF, even following the MD refinement (Table 2). Another example is 1a45A, for which CCF of the final structure was 0.973, even though Cα rmsd was only 12.2 Å, owing to a misorientation of one of the domains by [39 Å, 141°]. This domain is a globular β sandwich fold for which CCF of the final orientation was similar to CCF of the correct orientation (in the crystal structure). To overcome this problem, we need to add other types of information to the scoring function, such as statistical potentials (Shen and Sali, 2006) and geometric complementarity between domains (K.L., M.T., A.S., and H.W., unpublished data).

Second, an incorrect definition of the rigid bodies can "trap" the structure in a local minimum. A good example of this problem is the single-domain protein 1uwoA, where a helix that corresponds to a loop in the native structure was defined as a rigid body, preventing the refinement of the structure toward more accurate conformations. To tackle this problem, we need to assign the rigid bodies more accurately. Possible rigid-body assignments might rely on structural variation within the family of the structures related to the component or within a group of independently calculated models of the component. They might also be obtained using graph theory, neural networks, and other approaches based on energetic interactions within and between the proteins (Alexandrov et al., 2005; Flores and Gerstein, 2007).

Implications for Comparative Protein Structure Modeling

Experimentally determined atomic-resolution structures of molecular components are frequently not available, and most cryo-EM maps are generally still insufficient for atomic structure determination on their own. In such cases, it might be possible to identify a structure (template) that is homologous to the component (target) based on its amino acid sequence, and construct a useful model using comparative modeling (Eswar et al., 2007). Currently, ~1.3 million of the ~4.5 million known protein sequences (Bairoch et al., 2005) have at least one domain that can be modeled based on its similarity to one or more of the ~47,000 known protein structures (Pieper et al., 2006). However, sequence-structure alignments between the target and the template are a major source of errors in comparative models, especially in models of sequences that are only remotely related to their templates (i.e., at less than 30% sequence identity, which includes most detectably related protein sequences) (Eswar et al., 2007). Other errors include distortions and shifts of the backbone and side chains.

We recently showed that CCF between a comparative model and the corresponding density map is highly correlated with the accuracy of the model (Topf et al., 2005). We then built upon this correlation by adopting our Moulder genetic algorithm protocol that reduces alignment errors through the iteration over alignment, model building, and model assessment (John and Sali, 2003). For the application to EM (Moulder-EM), the iteration is guided by a fitness function corresponding to a combination of CCF and statistical potentials (Topf et al., 2006). The method was able to reduce by ~19% the Cα rmsd of 20 comparative models (which were based on less than 30% sequence identity to their homologs) using their 10 Å resolution native density maps. As expected, the improvement in the accuracy of the models was due mainly to a reduction in alignment errors and partly due to better loop modeling. However, errors in comparative modeling that occur due to target-template differences in the correctly aligned regions, (such as those in the relative positions of secondary structure elements and domains), could not be addressed. As with alignment errors, these types of errors can occur even when the sequence identity is high (i.e., higher than 30%), but become more significant at lower sequence identity (Eswar et al., 2007).

The Flex-EM method can address errors that occur as a result of target-template differences, because it relies on structure refinement that is guided by the restraints provided by the native density map. Indeed, the benchmark demonstrated that most of the improvement in accuracy was achieved by minimizing errors in the initial nonnative structures that resulted from target-template differences in the correctly aligned regions (the benchmark structures were based on an average sequence identity of 37%, resulting in accurate alignments; data not shown). A potential future direction is to combine Moulder-EM with Flex-EM to obtain an iterative procedure that can simultaneously address alignment errors and target-template differences in the correctly aligned regions.

Conclusion

We presented a method for fitting and refining atomic protein structures in a density map of their assembly at intermediate resolution. The inclusion of the stereochemical and nonbonded interaction terms during the refinement process enables a more realistic sampling of the conformational space. The method is likely to yield insights into the mechanisms of proteins within macromolecular assemblies for which the structure can often only be obtained at low to intermediate resolutions by cryo-EM techniques.

Supplemental Data

Supplemental Data include one figure and two movies and can be found with this article online at http://www.structure.org/cgi/content/full/16/2/295/DC1/.

ACKNOWLEDGMENTS

We thank Frank Alber, Friedrich Forster, Fred Davis, and Paula Petrone for very helpful discussions. M.T. is supported by an MRC Career Development Award. K.L. is supported in part by a fellowship from the Edmond J. Safra Bioinformatics Program at Tel-Aviv University. H.W. acknowledges support by the Binational U.S.-Israel Science Foundation, Israel Science Foundation (281/05), NIAID, and the Hermann Minkowski-Minerva Center for Geometry at TAU. W.C. is supported by the NIH (P41RR02250). A.S. is supported by the Sandler Family Supporting Foundation, NIH (R01 GM54762, U54 GM074945, P41 RR02250), Hewlett-Packard, NetApps, IBM, and Intel. W.C. and A.S. are supported jointly by NIH (PN2 EY016525) and NSF (EIA-032645 and 1IIS-0705474).

Received: July 26, 2007
Revised: November 20, 2007
Accepted: November 26, 2007
Published: February 12, 2008

REFERENCES

Alber, F., Eswar, N., and Sali, A. (2004). Structure determination of macromolecular complexes by experiment and computation. In Practical Bioinformatics, Volume 15, J. Bujnicki, ed. (Berlin and Heidelberg: Springer-Verlag), pp. 73–96.

Alexandrov, V., Lehnert, U., Echols, N., Milburn, D., Engelman, D., and Gerstein, M. (2005). Normal modes for predicting protein motions: a comprehensive database assessment and associated web tool. Protein Sci. *14*, 633–643.

Andersen, G.R., Thirup, S., Spremulli, L.L., and Nyborg, J. (2000). High resolution crystal structure of bovine mitochondrial EF-Tu in complex with GDP. J. Mol. Biol. *297*, 421–436.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2005). The Universal Protein Resource (UniProt). Nucleic Acids Res. *33*, D154–D159.

Baker, D., and Sali, A. (2001). Protein structure prediction and structural genomics. Science *294*, 93–96.

Baker, M.L., Jiang, W., Wedemeyer, W.J., Rixon, F.J., Baker, D., and Chiu, W. (2006). Ab initio modeling of the herpesvirus VP26 core domain assessed by cryoEM density. PLoS Comput. Biol. *2*, e146.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., et al. (2004). The Pfam protein families database. Nucleic Acids Res. *32*, D138–D141.

Braig, K., Adams, P.D., and Brunger, A.T. (1995). Conformational variability in the refined structure of the chaperonin GroEL at 2.8 Å resolution. Nat. Struct. Biol. *2*, 1083–1094.

Brooks, B., and Karplus, M. (1983). Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc. Natl. Acad. Sci. USA *80*, 6571–6575.

Brooks, C.L., Karplus, M., and Pettitt, B.M. (1988). Proteins: A Theoretical Perspective of Dynamics, Structure, and Thermodynamics (New York and Chichester: Wiley).

Chapman, M.S. (1995). Restrained real-space macromolecular atomic refinement using a new resolution-dependent electron-density function. Acta Crystallogr. A *51*, 69–80.

Chen, J.Z., Furst, J., Chapman, M.S., and Grigorieff, N. (2003). Low-resolution structure refinement in electron microscopy. J. Struct. Biol. *144*, 144–151.

Chen, Z., and Champman, M.S. (2001). Conformational disorder of proteins assessed by real-space molecular dynamics refinement. Biophys. J. *80*, 1466–1472.

Chiu, W., Baker, M.L., Jiang, W., Dougherty, M., and Schmid, M.F. (2005). Electron cryomicroscopy of biological machines at subnanometer resolution. Structure *13*, 363–372.

Eswar, N., Webb, B., Marti-Renom, M.A., Madhusudhan, M.S., Eramian, D., Shen, M.Y., Pieper, U., and Sali, A. (2007). Comparative protein structure modeling with MODELLER. Curr. Protoc. Protein Sci. *50*, 2.9.1–2.9.31.

Fabiola, F., and Chapman, M.S. (2005). Fitting of high-resolution structures into electron microscopy reconstruction images. Structure *13*, 389–400.

Fiser, A., Do, R.K., and Sali, A. (2000). Modeling of loops in protein structures. Protein Sci. *9*, 1753–1773.

Flores, S., Echols, N., Milburn, D., Hespenheide, B., Keating, K., Lu, J., Wells, S., Yu, E.Z., Thorpe, M., and Gerstein, M. (2006). The database of macromolecular motions: new features added at the decade mark. Nucleic Acids Res. *34*, D296–D301.

Flores, S.C., and Gerstein, M.B. (2007). FlexOracle: predicting flexible hinges by identification of stable domains. BMC Bioinformatics *8*, 215.

Goddard, T.D., Huang, C.C., and Ferrin, T.E. (2007). Visualizing density maps with UCSF Chimera. J. Struct. Biol. *157*, 281–287.

Goldstein, H. (1980). Classical Mechanics, Second Edition (Reading, MA: Addison-Wesley).

Han, J.H., Kerrison, N., Chothia, C., and Teichmann, S.A. (2006). Divergence of interdomain geometry in two-domain proteins. Structure *14*, 935–945.

Jiang, W., and Ludtke, S.J. (2005). Electron cryomicroscopy of single particles at subnanometer resolution. Curr. Opin. Struct. Biol. *15*, 571–577.

John, B., and Sali, A. (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res. *31*, 3982–3992.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637.

Lovell, S.C., Davis, I.W., Adrendall, W.B., de Bakker, P.I.W., Word, J.M., Prisant, M.G., Richardson, J.S., and Richardson, D.C. (2003). Structure validation by C α geometry: ϕ,ψ and C β deviation. Proteins Struct. Funct. Genet. *50*, 437–450.

Ludtke, S.J., Jakana, J., Song, J.L., Chuang, D.T., and Chiu, W. (2001). A 11.5 Å single particle reconstruction of GroEL using EMAN. J. Mol. Biol. *314*, 253–262.

Ludtke, S.J., Chen, D.H., Song, J.L., Chuang, D.T., and Chiu, W. (2004). Seeing GroEL at 6 Å resolution by single particle electron cryomicroscopy. Structure *12*, 1129–1136.

Ma, J. (2005). Usefulness and limitations of normal mode analysis in modeling dynamics of biomolecular complexes. Structure *13*, 373–380.

MacKerell, A.D., Jr., Bashford, D., Bellott, M., Dunbrack, R.L., Jr., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B *102*, 3586–3616.

Marti-Renom, M.A., Pieper, U., Madhusudhan, M.S., Rossi, A., Eswar, N., Davis, F.P., Al-Shahrour, F., Dopazo, J., and Sali, A. (2007). DBAli tools: mining the protein structure space. Nucleic Acids Res. *35*, W393–W397.

Ming, D., Kong, Y., Wakil, S.J., Brink, J., and Ma, J. (2002). Domain movements in human fatty acid synthase by quantized elastic deformational model. Proc. Natl. Acad. Sci. USA *99*, 7895–7899.

Mitra, K., and Frank, J. (2006). Ribosome dynamics: insights from atomic structure modeling into cryo-electron microscopy maps. Annu. Rev. Biophys. Biomol. Struct. *35*, 299–317.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. *247*, 536–540.

Petrone, P., and Pande, V.S. (2006). Can conformational change be described by only a few normal modes? Biophys. J. *90*, 1583–1593.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.

Pieper, U., Eswar, N., Davis, F.P., Braberg, H., Madhusudhan, M.S., Rossi, A., Marti-Renom, M., Karchin, R., Webb, B.M., Eramian, D., et al. (2006). MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res. *34*, D291–D295.

Rossmann, M.G., Morais, M.C., Leiman, P.G., and Zhang, W. (2005). Combining X-ray crystallography and electron microscopy. Structure *13*, 355–362.

Russell, R.B., Alber, F., Aloy, P., Davis, F.P., Korkin, D., Pichaud, M., Topf, M., and Sali, A. (2004). A structural perspective on protein-protein interactions. Curr. Opin. Struct. Biol. *14*, 313–324.

Saibil, H.R. (2000). Conformational changes studied by cryo-electron microscopy. Nat. Struct. Biol. *7*, 711–714.

Sali, A., and Blundell, T.L. (1993). Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. *234*, 779–815.

Sali, A., Glaeser, R., Earnest, T., and Baumeister, W. (2003). From words to literature in structural proteomics. Nature *422*, 216–225.

Shanno, D.F., and Phua, K.H. (1980). Remark on algorithm 500. Minimization of unconstrained multivariate functions. Trans. Math. Softw. *6*, 618–622.

Shen, M.Y., and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. Protein Sci. *15*, 2507–2524.

Shimamura, T., Koike-Takeshita, A., Yokoyama, K., Masui, R., Murai, N., Yoshida, M., Taguchi, H., and Iwata, S. (2004). Crystal structure of the native chaperonin complex from *Thermus thermophilus* revealed unexpected asymmetry at the *cis*-cavity. Structure *12*, 1471–1480.

Suhre, K., Navaza, J., and Sanejouand, Y.H. (2006). NORMA: a tool for flexible fitting of high-resolution protein structures into low-resolution electron-microscopy-derived density maps. Acta Crystallogr. D Biol. Crystallogr. *62*, 1098–1100.

Tama, F., Wriggers, W., and Brooks, C.L., III (2002). Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. J. Mol. Biol. *321*, 297–305.

Tama, F., Miyashita, O., and Brooks, C.L., III (2004). Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. J. Mol. Biol. *337*, 985–999.

Topf, M., and Sali, A. (2005). Combining electron microscopy and comparative protein structure modeling. Curr. Opin. Struct. Biol. *15*, 578–585.

Topf, M., Baker, M.L., John, B., Chiu, W., and Sali, A. (2005). Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy. J. Struct. Biol. *149*, 191–203.

Topf, M., Baker, M.L., Marti-Renom, M.A., Chiu, W., and Sali, A. (2006). Refinement of protein structures by iterative comparative modeling and cryoEM density fitting. J. Mol. Biol. *357*, 1655–1668.

Valle, M., Zavialov, A., Li, W., Stagg, S.M., Sengupta, J., Nielsen, R.C., Nissen, P., Harvey, S.C., Ehrenberg, M., and Frank, J. (2003). Incorporation of aminoacyl-tRNA into the ribosome as seen by cryo-electron microscopy. Nat. Struct. Biol. *10*, 899–906.

Velazquez-Muriel, J.A., Valle, M., Santamaria-Pang, A., Kakadiaris, I.A., and Carazo, J.M. (2006). Flexible fitting in 3D-EM guided by the structural variability of protein superfamilies. Structure *14*, 1115–1126.

Wriggers, W., Milligan, R.A., and McCammon, J.A. (1999). Situs: a package for docking crystal structures into low-resolution maps from electron microscopy. J. Struct. Biol. *125*, 185–195.