Journal of Structural Biology 175 (2011) 264–276

Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

BCL::EM-Fit: Rigid body fitting of atomic structures into density maps using geometric hashing and real space refinement

Nils Woetzel [a,c], Steffen Lindert [a,c], Phoebe L. Stewart [b,c], Jens Meiler [a,c,*]

[a] Department of Chemistry, Vanderbilt University, TN 37212, USA
[b] Department of Molecular Physiology and Biophysics, Vanderbilt University, TN 37212, USA
[c] Center for Structural Biology, Vanderbilt University, TN 37212, USA

ARTICLE INFO

Article history:
Received 4 February 2009
Received in revised form 28 April 2011
Accepted 28 April 2011
Available online 4 May 2011

Dedicated to Dr. Brigitte Heink on occasion of her retirement.

Keywords:
Cryo-electron microscopy
CryoEM
Geometric hashing
Real space
Monte Carlo Metropolis
Fitting
Docking

ABSTRACT

Cryo-electron microscopy (cryoEM) can visualize large macromolecular assemblies at resolutions often below 10 Å and recently as good as 3.8–4.5 Å. These density maps provide important insights into the biological functioning of molecular machineries such as viruses or the ribosome, in particular if atomic-resolution crystal structures or models of individual components of the assembly can be placed into the density map. The present work introduces a novel algorithm termed BCL::EM-Fit that accurately fits atomic-detail structural models into medium resolution density maps. In an initial step, a "geometric hashing" algorithm provides a short list of likely placements. In a follow up Monte Carlo/Metropolis refinement step, the initial placements are optimized by their cross correlation coefficient. The resolution of density maps for a reliable fit was determined to be 10 Å or better using tests with simulated density maps. The algorithm was applied to fitting of capsid proteins into an experimental cryoEM density map of human adenovirus at a resolution of 6.8 and 9.0 Å, and fitting of the GroEL protein at 5.4 Å. In the process, the handedness of the cryoEM density map was unambiguously identified. The BCL::EM-Fit algorithm offers an alternative to the established Fourier/Real space fitting programs. BCL::EM-Fit is free for academic use and available from a web server or as downloadable binary file at http://www.meilerlab.org.

1. Introduction

Cryo-electron microscopy (cryoEM) (Lepault et al., 1983) has evolved in the past decade as an important tool to obtain medium resolution structures of biological macromolecular assemblies in the form of density maps. One challenge is to dock high resolution experimental structures, obtained by X-ray crystallography (Kendrew et al., 1958) and nuclear magnetic resonance (NMR) (Wüthrich, 1990), or models of individual proteins into these density maps to arrive at quasi atomic-detail representations of the macromolecular assembly. This procedure identifies regions of conformational change and regions that can be assigned to proteins of uncharacterized structure or which are characterized only in isolation.

Several protocols have been developed to fit atomic structures, usually obtained by X-ray crystallography or NMR, into low and medium resolution density maps (Fabiola and Chapman, 2005; Wriggers and Chacón, 2001). The computational problem amounts to determining six degrees of freedom, three rotational and three translational. Exhaustive searches systematically seek within this six-dimensional parameter space to optimize the cross correlation coefficient (*CCC*), which consumes significant amounts of computational time (Korostelev et al., 2002; Roseman, 2000). Computational time can be reduced by the use of a fast Fourier transformation accelerated translational search as implemented in the "Colores" program within the SITUS package (Wriggers et al., 1999). In this approach only the three rotational degrees of freedom are searched in an exhaustive fashion in real space, while the translational degrees of freedom are searched in Fourier space. For both algorithms the step size impacts the speed of the calculation, but also the reliability and quality of the solution. An optimal local fit can be found with Chimera. It provides the benefit of a graphical user interface and an implementation of gradient refinement (Goddard et al., 2007). This refinement is only local and re-

Abbreviations: EM, electron microscopy; MCM, Monte Carlo/Metropolis; $RMSD_{C\alpha}$, root mean square distance or deviation of the Cα-atom coordinates; *CCC*, cross correlation coefficient; Voxel, volume pixel.

* Corresponding author at: Vanderbilt University, Departments of Chemistry, Pharmacology, and Biomedical Informatics, Center for Structural Biology, 465 21st Ave. South, BIOSCI/MRBIII, Room 5144B, Nashville, TN 37232-8725, USA. Fax: +1 615 936 2211.

E-mail addresses: nils.woetzel@vanderbilt.edu (N. Woetzel), jens.meiler@vanderbilt.edu (J. Meiler).

URL: http://www.meilerlab.org (J. Meiler).

1047-8477/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2011.04.016

quires that the initial placement is closer to the correct solution than the protein diameter. Gradient based local minimization has been implemented on general purpose graphical processing units (GPGPU) showing speed ups of at least 30 with the same accuracy as a CPU version (Woetzel et al., 2011).

To further increase the speed of fitting, vector quantization was introduced (Wriggers and Birmanns, 2001). Single molecule data is represented by k so-called codebook vectors for high resolution protein structure data and low resolution density maps. In a search within the k! permutations the best fit is identified by the lowest residual $RMSD_{C\alpha}$ after superimposition. This "QDOCK" method in the SITUS program is fast and reliable for rigid body docking and can be used for flexible docking as well. Difficulties arise however, if the density map contains different and multiple protein structures.

Protein structures obtained by X-ray crystallography often differ from the form of the protein observed in the cryoEM experiment. This can be the case if the protein was modified to facilitate crystallization or if a comparative model was built from a crystal structure of a homologous protein. In these cases the atomic model might not reflect all of the structural and dynamical properties observed in the cryoEM map. Therefore, flexible docking protocols were developed to overcome the limitations of rigid body fitting. For example, structural alignments of one protein to proteins in the same super family can be used to sample different conformations and improve the *CCC* (Velazquez-Muriel and Carazo, 2007). Alternatively, normal mode based fitting varies the coordinates of the structure within reasonable limits while docking (Tama et al., 2004). Molecular dynamics approaches have also been tested to optimize the fit of an atomic structure into electron density maps (Schröder et al., 2007; Trabuco et al., 2009). Flexible docking can also be achieved by defining hinges between domains and varying the orientation between them using QDOCK in the SITUS package. Methods such as molecular dynamics, conjugate-gradient minimization, and Monte Carlo optimization can be integrated with different scoring functions in an iterative protocol that combines the strengths of each individual approach (Topf et al., 2008).

The present work implements for the first time a "geometric hashing" algorithm (Wolfson and Rigoutsos, 1997) termed BCL::EM-Fit for the task of fitting atomic-detail protein models into cryoEM densities. Geometric hashing was developed in the robotics field, where feature-recognition and pattern-matching give computers the ability to connect real life objects to abstract computational representations. This technique is already used in structural biology to identify similar binding sites in proteins (Shulman-Peleg et al., 2004). A second step in the BCL::EM-Fit approach involves a Monte Carlo (Metropolis and Ulam, 1949)/Metropolis (Metropolis et al., 1953) (MCM) small perturbation protocol to refine the initial fits by maximizing the *CCC*. The time and robustness of BCL::EM-Fit compares favorably with the widely used Fourier/real space fitting program "COLORES" in the SITUS package (Wriggers and Birmanns, 2001). Benchmark results are presented with simulated density, as well as examples that demonstrate fitting with experimental GroEL density (Stagg et al., 2008) and of adenovirus capsid protein crystal structures into experimental cryoEM density maps (Saban et al., 2006).

2. Methods

2.1. Geometric hashing re-casted for searching density maps with protein structures

The following paragraph gives a general overview of the steps required before a more detailed description of the present implementation is given. The basic idea of geometric hashing was developed for image recognition in robotic applications. Critical points of a complex image (features) are extracted into a feature cloud. A large number of possible rotations and translations of this feature cloud are encoded *a priori* in a hash map (Wolfson and Rigoutsos, 1997) which later allows a rapid search for objects within this image. For BCL::EM-Fit the 3D image will be the cryoEM density map. The objects to be recognized will be protein structures which will also be represented as feature clouds. Each combination of a rotation (three degrees of freedom) and translation (three degrees of freedom) of the feature cloud is a transformation with six degrees of freedom.

The general scheme for generating the geometric hash is to define many possible transformations for the density map feature cloud and store these in a memory-efficient, rapidly searchable hash map. In this process the features are "quantized", i.e. not the actual position of a feature but only the specific space bin that contains the feature is stored. This procedure not only saves memory and accelerates the search, it also limits the search to a finite (but large) set of all possible transformations. Further it compensates for experimental noise in the density map and protein structure. In the recognition step this hash map is searched with a feature cloud representation of the protein to be docked. It is expected that one of the original transformations puts the feature cloud of the density map in good overlap with the feature cloud of the protein. This can be recognized by the number of shared features, i.e. features that end up in the same space bin.

This procedure speeds up the search as not the complete image but only the features deemed important are considered. Further, not every possible transformation is considered but only a finite subset. In contrast to robotics the problem of scaling the image is absent for feature-recognition in a distance invariant cryoEM density because the units of length in the density map and atomic models are the same. Further, 3D images have an increased complexity over 2D pictures that a robot usually sees using a single camera, which changes the protocol slightly compared to plain 2D picture recognition.

2.2. Extraction of feature cloud from density map intensities (Fig. 2a)

The user inputs a density map that will be completely encoded as a point cloud for rapid fitting. If the user wants to fit into a specific segment of the density map, it is necessary to extract that from the original map in a pre-processing step. In order to generate a representation of the features in the density map two pieces of information are used (Fig. 2a): the absolute intensity of a voxel and the intensity difference to its neighboring voxel, a gradient. The higher the intensity the more likely it is that a structurally compact region such as a secondary structure element can be found in the respective position of the density maps. The higher the intensity gradient the more likely the edge of a secondary structure element can be found here. Often there is an intensity drop at the edge of secondary structure elements due to less rigid amino acid side chain atoms. The edge regions are usually close to backbone atoms of secondary structure elements and encode most of the information within the density map. In order to define the total number of features extracted from a density map Eq. (1) was derived empirically:

$$N_{\text{points}} = N_{\text{Voxel Atoms}} \times \frac{V_{\text{Voxel}}}{Max(\frac{\pi}{6} d_{fd}^3, V_{Voxel}, \rho_{\text{Atoms Protien}}^{-1})} \quad (1)$$

where $N_{\text{Voxel Atoms}}$, Number of voxels the atoms would occupy when mapped to grid of the density map; V_{Voxel}, Volume of voxel; $\frac{\pi}{6} d_{fd}^3$, Volume that one point occupies according to feature distance; V_{Vox}, Volume that one point occupies according to a Voxel's volume;

$\rho^{-1}_{\text{Atoms Protien}}$, Volume that one point occupies according to the density of selected atoms for fitting in the protein.

The number of features that represent the density map should be proportional to the number of voxels that are occupied when the selected atoms in the protein structure that is to be fitted is mapped to the grid defined by the voxel size of the density map ($N_{\text{Voxel Atoms}}$). This number is reduced by the maximal volume that one feature can occupy. The maximum is given by one feature occupying one voxel (V_{Voxel}) which reduces Eq. (1) to $N_{\text{points}} = N_{\text{Voxel Atoms}}$. If the density of atoms that are to be fitted is low, the expected Volume one feature is occupying is high which reduces Eq. (1) to $N_{\text{points}} = N_{\text{Voxel Atoms}} \times \frac{V_{Voxel}}{\rho_{\text{Atoms Protein}}}^{-1}$. If the feature distance is chosen high, the volume one feature occupies is high which reduces Eq. (1) to $N_{\text{points}} = N_{\text{Voxel Atoms}} \times \frac{V_{VoxelAtoms}}{\frac{\pi}{6}d_d^3}$. A good estimate for the number of features reduces the size of the hash map since less triangular bases are constructed and fewer features have to be transformed, quantized, and stored (read below). In addition a sufficient number of features guarantee enough triangular bases, to achieve a high precision for the fits. Custom optimization of Eq. (1) or its parameters might be required for optimal results. However, the algorithm proved robust in the presented work with respect to deviations in N_{points} of up to 25%. Hence, Eq. (1) should be applicable for most scenarios. A default choice for the feature distance is $0.15 \times r_{\text{gyr}}$ (radius of gyration of protein to be fitted), which has proven robust for the presented experiments, but can be modified. A smaller feature distance will lead to more overall features and longer fitting times. The actual scaling for the time cannot be determined since the feature distance also influences the number of triangular bases. To a first approximation, the overall time should scale quadratically with the reduction of the feature distance. Setting the feature distance to a value larger than the default value may lead to an insufficient number of encoded features.

The actual features are extracted by iterating over all voxels. For each voxel the intensity is added to the gradient intensity of the neighboring voxels. The gradient is the sum of all absolute differences to the neighboring voxels, i.e. 6 voxels adjacent on the faces, 12 on the edges and 8 on the vertices. The absolute differences are normalized by the distance between the voxels, e.g. voxels adjacent on the yz-faces are normalized by voxel length in x-direction (vl_x) or voxels on the vertices by $\sqrt{vl_x^2 + vl_y^2 + vl_z^2}$. The voxel is converted into a feature by adding the map's indices to the voxel's indices and by multiplying with the voxel width and adding the maps origin afterwards. Half of the voxel width is also added to center the feature in the voxel. The feature is inserted in a list with its intensity and gradient sum and is sorted by the sum. Finally, starting with the highest intensity-gradient-sum, the list is searched for all features that are within the feature radius of that feature, which have to be removed. Then the list is searched for all overlapping features with the second highest by the intensity-gradient-sum. This happens until no overlapping features remain. The list is then cut down to the requested number of features removing the lowest intensity-gradient-sum features.

2.3. Selection of triangular bases for coordinate transformations (Fig. 2b)

Triplets of the features f_1, f_2 and f_3 within the density map are treated as an origin of a coordinate system–a so called triangular base. Transforming all remaining features within a specified feature radius of the triangular base, this coordinate system encodes the relative position of the features with respect to this base. The internal coordinate system represented by the triangular base is invariant to the absolute position of the structure in space but encodes only relative positions of features.

It was critical to *not* consider all possible triplets of features as base. Rules were imposed that ensured that the distances $d_1 = ||f_2 - f_3||$, $d_2 = ||f_1 - f_3||$ and $d_3 = ||f_1 - f_2||$ between the features f_1, f_2 and f_3 are chosen to be between 0 and the radius of gyration of the structure to be fitted. The rationale for this approach is that within this range the relative arrangement of secondary structure elements is defined. This is ultimately the structural entity to be recognized in the search procedure. Further, it is advantageous to ensure that d_1, d_2 and d_3 are significantly different from each other and can be sorted (read below). For that purpose three thresholds are defined: r_{gyr}, the radius of gyration of the protein to be fitted, a high and a low threshold t_h and t_l. These are determined by binning all pairwise distance into 100 equal sized distance bins in the range [0, r_{gyr}]. The resulting distance histogram is used to find the two bins, at which 1/3 of all distance (t_l) and 2/3 of all distances (t_h) were observed, which typically turns out to be close to 0.5 and 0.75 times the radius of gyration of the protein to be fitted. The distances d_1, d_2 and d_3 have to fulfill the conditions:

$$r_{\text{gyr}} > d_1 > t_h > d_2 > t_l > d_3 > 0 \tag{2}$$

The arithmetic center of the triangle f_1, f_2 and f_3 is used as the origin of the coordinate system, letting f_1 be on the positive x-axis, f_2 in the positive xy-plane. This generates an ordered triplet of features and a unique transformation T_D for those three features. Without an ordering $d_1 > d_2 > d_3$, it would be necessary to store all six possible transformations for a triangular base (starting from f_1, f_2 or f_3, clockwise or counter clockwise) increasing the computational time by a factor of 6 respectively. Additionally, the geometric hashing fit step would also need to consider 6 different transformations for the chosen triangular base totaling to a factor of 36.

2.4. The maximal distance of features from the coordinate base is limited by a feature radius (Fig. 2b)

Only coordinates that are within the feature radius (outer most circle of the spherical coordinate system, Fig. 2b) are transformed and quantized. The rationale for the feature radius is that only features within the size of a typical protein domain need to be encoded. Features outside this radius arise from noise in the density map or neighboring domains and fitting results would not be improved even when considering these features. This radius restriction is particularly important if a large density map of multiple proteins is searched for individual proteins or domains. In this case the feature radius helps to reduce the memory required for storing the hash map and to reduce the computational time.

The feature radius can be seen as a maximum size of objects that can be reliably detected within the encoded density map. Hence, the feature radius should be chosen based on the size of structures that will be fitted and should have a value between the radius of gyration and the longest extent of that object. By default it is chosen to be $1.25 \times r_{\text{gyr}}$. All features f_i considered for transformation have to be within the distance r of the middle point $M = \frac{1}{3}(f_1 + f_2 + f_3)$ of the three features f_1, f_2 and f_3 used as the origin

$$r > ||f_i - \frac{1}{3}(f_1 + f_2 + f_3)|| \tag{3}$$

2.5. Quantization of features accounts for finite number of transformations, low resolution of the density map, and experimental noise (Fig. 2b–c)

To generate the keys from the transformed features f_i a quantization procedure is applied. Quantization assigns the feature to some bin in space based on its position. The advantage of such binning is that only a finite number of bins exist which will be the keys of the hash map. The precision of the quantization adjusts also the tolerance in the feature matching step (read below), i.e.

features in the density map that would map to atoms in the protein can deviate significantly if they are distant from the triangular base but should still count as a match. The density maps extracted features represent edges and high intensity density features. The feature cloud of the protein represents certain atoms (read below). However, it is not expected that these points superimpose precisely as features mark general regions not precise points. Both density map and protein structure are experimental data affiliated with errors and uncertainties. Hence, a certain tolerance between features of the density map and features of the atomic structure should be allowed for matches.

The precision of the quantization needs to be tuned to the resolution of the density map. A lower precision will tolerate a larger distance between an atomic feature and a density feature in the fit. The number of distinct keys will be small and the fitting will be faster, but accuracy might suffer. A higher precision on the other hand will give closer and more reliable fits. It will produce more distinct keys, require more time for the fitting, and should be used with higher resolution density maps.

In the present implementation a Spherical coordinate system was used to define the bins rather than a Cartesian coordinate system. The radius of the bins was chosen to increase logarithmically. The choice of the coordinate system has certain advantages and disadvantages: The use of a spherical coordinate system requires the conversion of the point cloud coordinates from Cartesian to Spherical coordinates. In contrast to the Cartesian coordinate system in the Spherical coordinate system the bin sizes increase with distance from the origin, i.e. a spherical coordinate system has a lower resolution for points that are farther away from the origin. This is beneficial as small changes in the transformation will disproportionately affect the position of features distant from the origin. In a Spherical quantization these points may remain in the same bin and can be recognized as overlapping features (read below) while in a Cartesian quantization they would wander into the next space bin. Spherical quantization gave slightly better results than Cartesian quantization in benchmark experiments (data not shown). The following equations were used to convert Cartesian coordinates $\vec{pos}_{Cartesian} = (x, y, z)$ into Spherical coordinates $\vec{pos}_{Spherical} = (\gamma, \vartheta, \varphi)$:

$$\vec{pos}_{Spherical} = \begin{pmatrix} \gamma \\ \vartheta \\ \varphi \end{pmatrix} = \begin{pmatrix} \sqrt{x^2 + y^2 + z^2} \\ \arccos\left(\frac{z}{\gamma}\right) \\ \arctan2(y, x) \end{pmatrix} \tag{4}$$

For quantization the following equations were applied to the Spherical coordinates:

$$\begin{pmatrix} \gamma_q \\ \vartheta_q \\ \varphi_q \end{pmatrix} = \begin{pmatrix} \frac{\log(\gamma)}{\log(\frac{2\pi}{res}+1)} \\ \vartheta \times \frac{res}{\pi} \\ \frac{\varphi}{\pi} \times \left\lfloor \sin\left(\frac{\pi \times \vartheta_q}{res}\right) \times res + 1 \right\rfloor \end{pmatrix} \tag{5}$$

where res is the resolution of the key and influences the quantization. The smaller res is, the more points will fall in the same bin and the more the initial fit deviates from the correct fit. Hence the hash key resolution behaves in the opposite manner to the density map resolution. A typical value is twelve, which creates twelve angular bins for φ_q on the equator of the spherical coordinate system each spanning an angle of 24°. Since $\frac{\vartheta}{\pi}$ is in the range [0, 1] and for the equator $\vartheta_q = \left\lfloor \frac{\pi}{2} \times \frac{15}{\pi} \right\rfloor$ the term $\frac{\varphi}{\pi} \times \lfloor \sin(\frac{\pi \times \vartheta_q}{res}) \times res + 1 \rfloor = \frac{\varphi}{\pi} \times \lfloor \sin(\frac{\pi \times 7}{15}) \times 15 + 1 \rfloor = \frac{\varphi}{\pi} \times \lfloor 0.99 \times 15 + 1 \rfloor = \frac{\varphi}{pi} \times 15$ creating a range [0, 15) of integer values, where 15 is at the open end of the interval because of the quantization of the floor function. The function $\lfloor x \rfloor$ = floor(x) returns the largest integer not greater than x (e.g. $\lfloor 1.1 \rfloor = 1; \lfloor 7.7 \rfloor = 7$). The key was assembled as one number using:

$$key = \gamma_q \times 10,000 + \vartheta_q \times 100 + \varphi_q \tag{6}$$

The factors 10,000 and 100 have to be increased, if the hash resolution increases to guarantee that there is no overlap between the individual quantized terms.

2.6. Hash map architecture (Fig. 2c)

For a specific transformation T_D every feature f_i within the feature radius f_r is converted into a key and stored in the hash map together with its respective transformation T_D. The resulting keys can be rapidly looked up in the hash map and all transformations T_D affiliated with a single key will be returned. It is very likely that there are multiple bases for one key, and it is also likely that certain keys will never be observed. Preprocessing of the density map and storing the hash map is the most memory and time consuming part of the algorithm. The actual implementation uses a SQL databank for larger hash maps, but can be stored in the RAM of a computer for smaller density maps accelerating the search.

2.7. Atoms within secondary structure elements are used as features to represent the protein (Fig. 2d)

A feature cloud for the protein to be fitted needs to be created. Since the atomic structure of the target protein is given it is possible to use the coordinates of atoms as features, preferably atoms that are close to regions which have high intensities in density maps. For the present purpose these are the backbone atoms within secondary structure elements. The relative rigidity of these regions coupled with the density in conjugated peptide bonds gives rise to high-intensity regions, i.e. the frequently discussed "density rods" seen for α-helices (Jiang et al., 2001; Saban et al., 2006). It is sufficient to include a fraction of all backbone atoms, i.e. C_α atoms, to reduce the number of features to be matched minimizing the time for fitting (Fig. 2d). Usage of any other backbone atom instead of C_α did not affect the accuracy of the protocol significantly (data not shown). It is recommended that the C_α atoms of all secondary structure elements be used as the feature cloud of the protein. For this purpose the program uses the secondary structure definition as given in HELIX and SHEET section of the PDB entry to automatically select the respective atoms. Atom names are taken from the ATOM lines in the PDB file. The user can alter which secondary structure regions to consider by changing the minimal length of the three secondary structure types (helix, sheet, loop) from the default values (0, 0, 999). Additionally, the user can pass a list of backbone atoms to be used although it is recommended to only use the C_α atoms as the use of additional atoms will increase the runtime and may not improve the results.

2.8. Initial fits are determined that superimpose the maximum number of features (Fig. 2e–g)

Once the feature set of the target is extracted, a possible triangular base is identified. In this procedure the same criteria are applied with respect to f_1, f_2 and f_3 that were used to encode the density map (Fig. 2e). Applying the resulting transformation T_P to the remaining features within the feature radius r and quantizing them yields a set of keys. This set of keys is now looked up in the hash map and transformations T_D are identified that are common among a maximum number of keys (Fig. 2f). Such transformations superimpose target protein and density with a maximum number of agreeing features and create a ranked list of initial fits. The transformation T_{fit} needed to fit the protein into the density is defined as $T_{fit} = T_p \times T_D^{-1}$ (Fig. 2g).

Since it cannot be expected that any three features of the target protein are necessarily represented in the feature cloud of the

density map, the fitting is repeated multiple times (Fig. 2e) and all transformations are ranked by the number of agreeing features (identical keys, Fig. 2f). The number of agreeing features is a quality measure for the initial fit. Since a large number of triangular bases within the target can be used, the following method is used to assure that the target is sampled equally, i.e. different bases with centers at sufficiently different locations within the target are picked. All bases are binned with their base centers on a Cartesian grid, with a grid width chosen, so that there are more grid elements occupied than fitting trials requested. Now, a grid element is picked randomly, and marked to not be picked again. A random triangular base within that grid element is chosen for the geometric hash fit procedure.

The accuracy of the initial fit depends on the number of features extracted from the density map and the number of features extracted from the protein model. More features increase the resolution and possibly the accuracy of the fit as more features in space are represented and more triangular bases can be identified. Since each base represents a set of translations and rotations, the space of transformations is sampled more densely. A higher agreement resulting from more superimposed features in the initial fit also results in a higher *CCC* with the density map. However, a large number of features results in longer computation times. Hence, the minimal number of features required to accurately represent the experimental information within the cryoEM density map should be used. The estimate for the number of features in the density map given in Eq. (1) represents a compromise between accuracy and computation time.

2.9. Filtering fits by translational and rotational distance

For the fitting of the penton base, hexon and GroEL, independent fits were defined by specified minimal rotational and translational differences before the geometric hash step. This is necessary, because the geometric hashing algorithm has an intrinsic property that leads to nearly identical fits being found in multiple searches with different triangular bases. In order to find a comprehensive list of independent and highly scoring fits, it is necessary to remove non-independent fits so that a few solutions do not dominate the output list.

2.10. The initial fits have to be optimized (Fig. 3)

For the purpose of optimizing initial fits, a simulated density map is computed from the atomic structure of the target with a resolution comparable to that of the experimental cryoEM density map. Starting from the position of the initial fit, small random translations and rotations are applied to the protein in order to maximize the *CCC* (Eq. (7)) in a Monte Carlo/Metropolis (MCM) simulated annealing protocol (Fig. 3).

$$CCC = \frac{k\sum_{y<k}^{y=0}\rho_s(y)\rho_E(y) - \sum_{y<k}^{y=0}\rho_s(y)\sum_{y<k}^{y=0}\rho_E(y)}{\sqrt[2]{k\sum_{y<k}^{y=0}\rho_E(y)^2 - \left(\sum_{y<k}^{y=0}\rho_E(y)\right)^2} \times \sqrt[2]{k\sum_{y<k}^{y=0}\rho_s(y)^2 - \left(\sum_{y<k}^{y=0}\rho_s(y)\right)^2}} \quad (7)$$

ρ_s and ρ_E are simulated and experimental overlapping densities. k is the number of overlapping voxels for which $\rho_s > 0$. This condition represents an "envelope" around the experimental density which will ignore noise in the region where no density was simulated from the fitted atomic structure. y is the iteration index over all voxel pairs that fulfill the $\rho_s < 0$ condition. The value of *CCC* will be 1 for best correlation, 0 for no correlation and −1 for anti-correlation.

Compared to gradient based methods Monte Carlo/Metropolis optimization is capable of sampling multiple local minima on a rugged objective function but is nevertheless accurate and fast. The scoring function is rugged due to experimental noise in the density map and due to the fact that voxel spacing quantizes the function. The input parameters for the protocol include maximum amplitude for rotations and translations, a maximum number of total iterations, and a maximum number of subsequent steps with no improvement in *CCC*. Typical translational step sizes are 0–1.0 Å; rotations are limited to 0.035 radians (∼2°). An average optimization explores between 100 and 200 steps, stops at a maximum of 250 steps, but terminates after 50 steps without an improvement in the *CCC*. The temperature parameter for the Metropolis criterion is adjusted automatically to match a certain ratio between accepted and rejected steps. This "simulated annealing" protocol starts with an estimated 50% ratio of accepted vs. rejected steps and ends with an approximate 20% ratio over the maximum of 250 steps, i.e. the final ratio of accepted steps is typically close to 0%.

2.11. Addition of noise to the synthesized density maps

Density maps were synthesized from coordinates following an implementation of pdb2vol in the SITUS package, using trilinear interpolation and Gaussian flattening kernel. This method produces density maps with zero intensity outside an envelope surrounding the protein. Different experimental deviations between the electron density map and the atomic structure can occur. First, there may be deviations in the structure or dynamics of the protein between the cryoEM conditions and the conditions used to determine the atomic-detail model. For example packing artifacts in crystals used for X-ray crystallography can result in different protein conformations than observed by cryoEM where the samples are preserved in near native conditions. Both can differ from structure and dynamics of an isolated dissolved protein observed in an NMR experiment. Further, differences in the actual proteins can occur such as length of the constructs or mutations. These deviations are not accounted for in the present algorithm but could in part be addressed through a flexible docking protocol.

However, a careful analysis was performed to test the robustness of the algorithm in the presence of noise. The noise added was Gaussian noise to mimic some of the error that is inherent in experimental density maps. While iterating over all voxels a normally distributed number was added to each voxel's intensity. After iterating over all voxels, the *CCC* between the noise-free and noise-added map was calculated. This process of adding noise was repeated, until the desired *CCC* to the noise-free density map was reached.

2.12. Specific parameters used for benchmark of 50 diverse proteins with simulated density maps

The proteins selected for the test have between 150 and 300 residues. Fifteen density maps in the resolution range of 5 to 19 Å in 1 Å steps were simulated from each of the crystal structures with Gaussian flattening (Wriggers and Birmanns 2001). The voxel size was chosen to be 1/3 of the resolution. For each protein/resolution combination four additional density maps were calculated with different levels of Gaussian noise added. The noise levels were adjusted so that the *CCC* values of the noise-added maps to the noise-free maps would be approximately 0.9, 0.8, 0.7 and 0.6. The *CCC* values were calculated according to Eq. (7). Fig. 4d shows one of the α/β benchmark proteins (1prz) with its noise-free simulated density map and its noise-added maps at a resolution of 10 Å. Visual inspection reveals that maps with noise at *CCC* value of 0.8 look comparable to the experimental map of adenovirus. The simulated maps and the corresponding atomic coordinates served as input for the BCL::EM-Fit geometric hashing and MCM optimization routines.

For the geometric hashing step the density maps were converted into feature clouds with between 22 and 232 points. These point number totals are intended to represent the structural features in a particular density map, which depends on the voxel size, the size of the protein, and the minimum distance between two resolvable features (Eq. (1)). Ten top scoring placements from the initial geometric hashing step were selected for each atomic model fit into each of its simulated density maps (the noise-free map and the four noise-added maps) at each of the 15 resolution test points. These initial hits were subjected to MCM refinement in real space.

2.13. Specific parameters used for penton base, hexon and GroEL

For the penton base fit, 709 and 631 features were extracted from the density segments at 6.8 and 9.0 Å resolution, respectively. The hexon capsid protein density segments were represented by 2890 and 3699 features for the 6.8 and 9.0 Å density maps, respectively. 2884 features were used to represent the entire 5.4 Å resolution density map of GroEL. The weight for intensity vs. gradient was the standard 1:1 ratio for all experiments (Fig. 2a). The t_l and t_h values as described in Eq. (2), the feature distances and the feature radii were derived from the radius of gyration. For all fitting procedures, a spherical coordinate system was used. The precision for the hash key quantization was set to 12 (Fig. 2b).

For the fitting of the proteins in the benchmark set, C_α atoms in helices or strands were extracted as features depending on whether it was more predominantly an α-helical, α/β or β-strand protein. For the penton base, C_α atoms in α-helices and β-strands were selected for fitting, for the hexon C_α atoms in β-strands, for GroEL C_α atoms in α-helices were selected for fitting (Fig. 2d). In all procedures 500 randomly chosen bases (Fig. 2e) were selected to generate a list of transformations T_D ordered by the number of agreeing features representing the best possible initial fits (Fig. 2f and g). For all MCM optimizations the specific parameters were derived as described in the Methods section "The initial fits have to be optimized".

In an effort to remove similar transformations $T_{fit} = T_p \times T_D^{-1}$ the list of initial fits for the penton base was filtered by removing solutions if their centers were within 5 Å and had a relative effective rotation angle smaller than 1 radian (~60°) using a previously described protocol (Urzhumtseva and Urzhumtsev, 2002). The list of initial fits for the hexon was filtered by removing solutions that were closer than 60 Å and had a relative effective rotation angle of less than 2 radians (~120°). Two fits for the GroEL experiment were considered identical within a translational difference of 5 Å and rotational difference of 3 radians (~170°). This filtering was necessary to find symmetrically related copies (since the hexon and penton base proteins are multimers) and to find translationally independent copies (the hexon map density segment had density for at least 4 full hexon proteins, the GroEL density map contained density for all 14 subunits).

2.14. Fold recognition and construction of comparative models using bioinfo.pl and Modeller

To identify template folds and construct comparative models for the benchmark proteins their primary sequences were submitted to the bioinfo.pl meta server. The output with the best aligned sequence, and with sequence similarity <99% to the original sequence, was chosen as a homologous structure. This helps to ensure that the template protein and homologous structure will have some differences. It is appreciated that in real-word applications the template and target structures may be considerably more distinct. However, a more detailed analysis of usage of comparative models for fitting is beyond the scope of the present work. The homologous proteins were downloaded from the PDB (Dutta and Berman, 2005) and used for cross-fitting experiments. Comparative models were acquired by submitting the bioinfo.pl alignment to the Modeller server using the "model" link provided on the bioinfo.pl website. This approach was chosen to keep the protocol as straight-forward and unbiased as possible. A more elaborate construction yielding possibly more accurate comparative models for fitting into cryoEM density maps remains to be pursued in future studies.

3. Results

3.1. An efficient two-stage low and high resolution fitting protocol

The BCL::EM-Fit protocol consists of several steps including geometric hashing to find initial fits, and Monte Carlo/Metropolis (MCM) optimization for refinement (Fig. 1). Features are extracted from the density map and stored in a hash map (either in computer memory or a databank, see also Fig. 2a–c). The fitting procedure involves feature extraction from the atomic protein structure and comparison with saved features from the density map. The best initial fits are determined by counting matching quantized features between the atomic structure and density map (see also Fig. 2d–g). Finally, a MCM optimization step is used to refine the initial fits based on real space *CCC*. The following paragraphs give a brief summary of the major steps. Implementation details are discussed in the Section 2.

In the first step the density map is converted into a feature cloud using several user inputs, such as the number of structural features expected in the density map and minimal distance between structural features (Fig. 2a). Regions of high intensity and with large

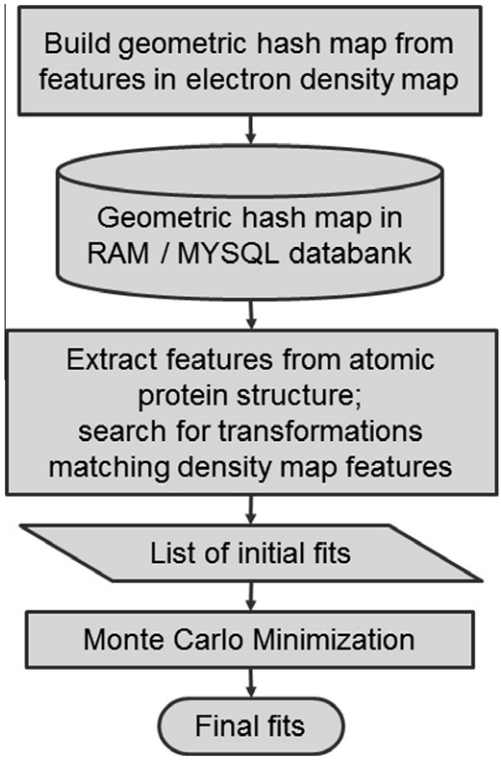

Fig.1. Schematic flowchart of BCL::EM-Fit. The general scheme of BCL::EM-Fit starts with the extraction of geometric features from the density map. These features are transformed into different orientations and saved together with their respective transformation in a hash map that is stored in computer RAM or in a MySQL databank. This process must be completed once for an experimental density map. In order to dock an atomic structure representative features are extracted from the coordinate set and compared to the hash map. The geometric hashing algorithm identifies a list of transformations that maximize the number of shared features between density map and atomic structure. Each of these initial fits is optimized in a MCM refinement step.

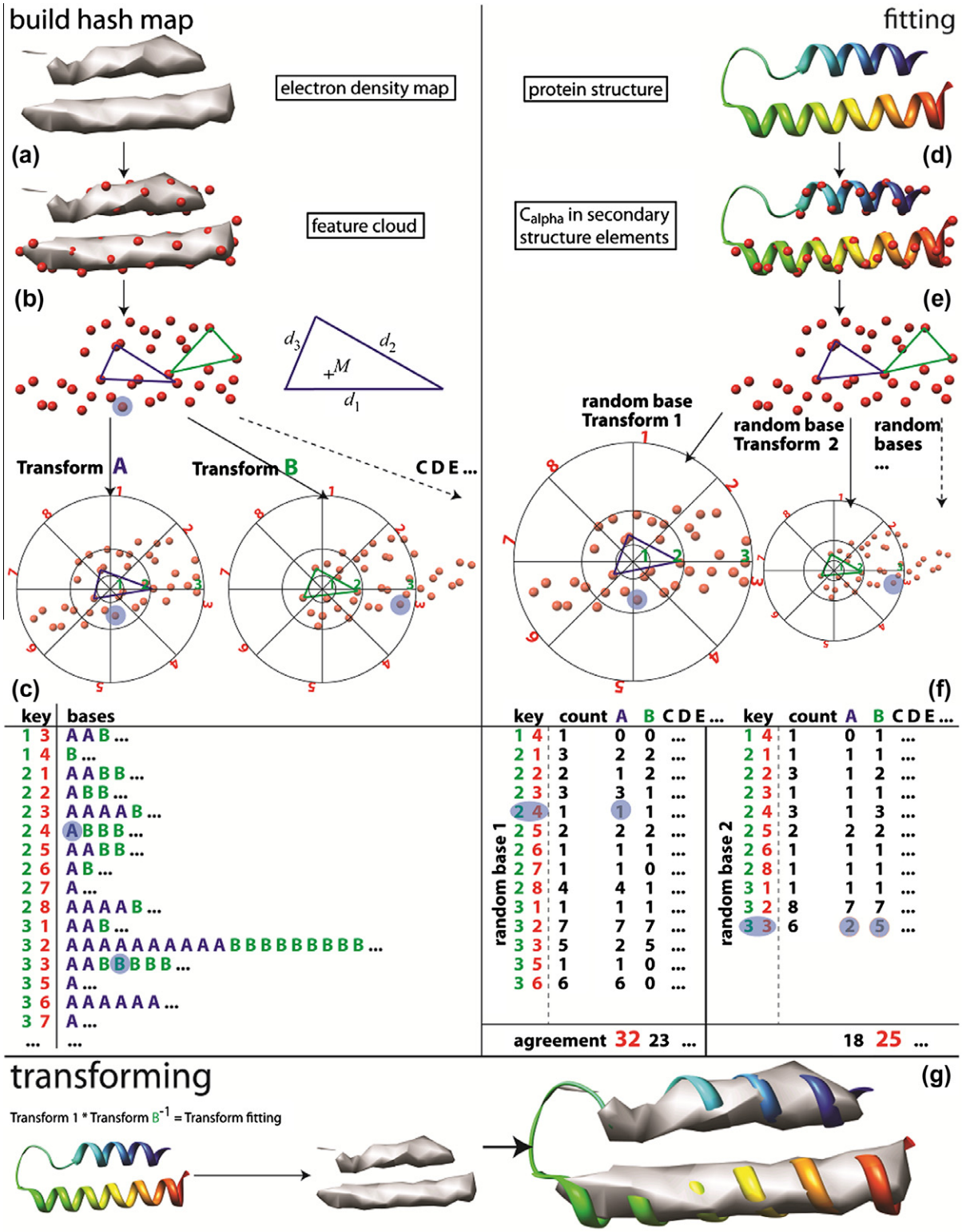

Fig.2. Detailed flowchart of geometric hashing protocol. The geometric hashing protocol is illustrated with an example protein structure and its density map in two dimensions. Building the hash map starts with (a) extracting a feature cloud from the density map. (b) Each possible combination of three features represents a triangular base with the sides d_1, d_2, and d_3. Triangles that satisfy Eq. (2) represent a base that is transformed to be the origin of a new coordinate system. (c) All remaining points that satisfy Eq. (3) in terms of their distance to the base (outermost circle) are transformed and quantized using a spherical coordinate system (Eqs. (4)–(6)). Quantized coordinates are stored in the hash map with the respective triangular base. The blue highlighted point will occur in the hash map multiple times affiliated with different keys and different bases. Steps (a–c) are performed once for every density map. (d) The fitting starts with extracting features from the protein structure i.e. C_{α}-atoms in α-helices. (e) Subsequently random bases are picked in this feature cloud and all features of the protein structure are transformed with respect to these random bases. (f) Now, all keys affiliated with a random base are looked up in the hash map. From this procedure original triangular bases are identified that share a maximum number of keys. Each shared key represents one agreeing feature between protein and density map and increments the hash score by one. The blue highlighted point adds to the agreement, if it corresponds to the matching base in the hash map. (g) The transformations with the highest hash scores will be chosen as the best initial fit.

intensity gradients are automatically selected from the density map. High intensity regions describe the centers of structural features, such as observed density rods for α-helices, which typically have high intensity values. Large gradients are observed along iso-surfaces of structural features and can be thought of as points along structural edges. This information is stored in a feature cloud corresponding to the selected voxel (volume pixel) centers. Within this feature cloud triangular bases are selected according to minimal and maximal distances between the three points (Eq. (2) and Fig. 2b). These triangular bases serve as a coordinate framework in which all other features of the cloud are expressed. Each triangular base is described by a unique transformation consisting of three rotational and three translational parameters. After transforming the feature cloud for each triangular base, the features within a given feature radius (Eq. (3)) are quantized (Eqs. (4)–(6)) and stored in a geometric hash map together with the respective transformation (Fig. 2c). The feature radius is chosen depending on the dimensions of the atomic structures to be fitted. This procedure effectively stores the feature cloud as seen from many different perspectives in space. This preprocessing procedure is only performed once for a given density map.

In order to fit a given atomic model into the previously encoded density map, a user-defined subset of backbone atoms (C_α, N, O, or C) within secondary structure elements must be extracted from the full coordinate file (Fig. 2d) (see details in Section 2). The rationale for using only backbone atoms is that these atoms are usually close to the edges of high-density regions in the density map and typically define edges of regular secondary structure elements such as α-helices. From within this set of atoms three features are chosen as a triangular base and all other features are transformed so that the triangular base ends up at the origin (Fig. 2e). The transformed features within the feature radius are quantized and then searched for within the hash map representation of the density map (Fig. 2f). The geometric hashing algorithm results in the identification of transformations that superimpose a maximum number of features between the atomic resolution model and the density map (Fig. 2g). Henceforth the maximum number of superimposable features will be termed the "hash score".

In the second stage of the BCL::EM-Fit protocol, a small number of top scoring initial placements are refined with MCM optimization applying rotational and translational perturbations (Fig. 3). The real space *CCC* (Eq. (7)) is maximized between a simulated density map based on the atomic model and the experimental density map. The refined placements are ranked by *CCC*.

3.2. Protein fitting procedure is highly reliable for resolutions of10 Å or better

In order to evaluate the reliability of the BCL::EM-Fit algorithm a benchmark was performed with 21 α-helical, 7 β-strand and 22 α/β proteins (Table S1). Specific parameters can be found in the Section 2. Fig. 4 presents the BCL::EM-Fit results for all of the benchmark proteins fit within their simulated density maps with various noise levels as a function of resolution (5–19 Å). The results were analyzed for each atomic model/simulated map combination to see if at least one of the initial 10 best fits by hash score was refined by MCM to have a final placement with an $RMSD_{C\alpha}$ value of <5 Å with respect to the correct position. Note that for the set of α-helical benchmark proteins fit within the noise-free maps, essentially all of the BCL::EM-Fit runs resulted in at least one MCM refined fit with an $RMSD_{C\alpha} < 5$Å. This is shown in Fig. 4a as black bars with heights of 20%, or close to 20%, at all resolutions in the range of 5–19 Å. Since the noise-free maps represent 20% of the total maps tested, this level represents the fact that a correctly fit solution was found for almost all atomic model/simulated density combinations in the α-helical benchmark proteins category using noise-free maps. As the plot indicates, the BCL::EM-Fit results are not quite as good with the noise-added maps. Nevertheless, an overall success rate of 90% is achieved for the α-helical benchmark proteins with simulated density maps up to ~14 Å resolutions. The BCL::EM-Fit results for the set of α/β benchmark proteins (with more than 2 helices and 2 strands in the structure) indicate an overall success rate of 90% with simulated density maps up to ~11 Å resolution (Fig. 4b). The β-only benchmark proteins were the most challenging, with a 70% success rate up to ~10 Å resolution (Fig. 4c).

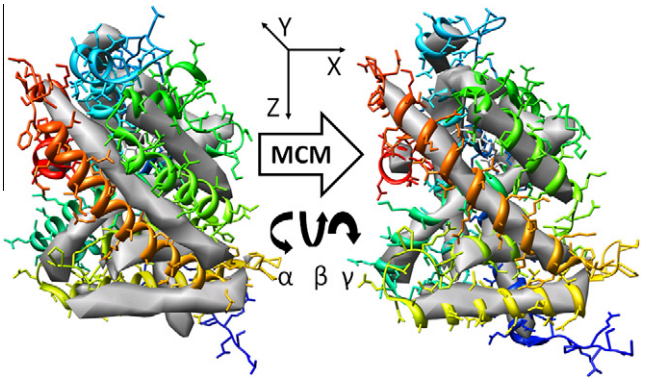

Fig.3. MCM refinement through a real-space rigid body six-dimensional search. Schematic representation of the Monte Carlo Metropolis (MCM) refinement step in which rigid body movements (translations in *X*, *Y*, and *Z* and rotational changes around α, β, and γ) are applied to the atomic protein structure relative to the density map in order to maximize *CCC*. After each movement the *CCC* between the experimental density and the simulated density map (derived from the atomic protein structure, Eq. (7)) is calculated.

As the results presented in Fig. 4 show, there are combinations of atomic models and simulated density maps for which refinement of the initial 10 best fits by hash score did not result in any correct final positions (i.e. within $RMSD_{C\alpha} < 5$Å). However, the trends reflected in Fig. 4 indicate that fitting failures occur with greater frequency when simulated maps with higher noise levels or of lower resolution are used. This implies that at a certain point the simulated density maps lack a sufficient number of unique features for this method to find the correct fit of the atomic model within the best 10 placements.

In general, these benchmark tests show that α-helical proteins are fitted with higher success rates than α/β-proteins, followed by β-strand proteins. It should be noted that these benchmark tests were designed to reveal the theoretical limits of the hashing algorithm and the MCM protocol. Admittedly, the benchmark tests were performed with single protein molecules in isolation and do not reflect the results one might expect when there are neighboring molecules or symmetry related subunits present in the density map. Also other than Gaussian noise, no attempts were made to mimic additional sources of error that might be present in an experimental cryoEM density map. These include errors due to conformational flexibility and heterogeneity. However, these benchmarks do show that the BCL::EM-Fit protocol performs well for isolated α-helical proteins, mixed α/β and β-strand proteins, albeit with different resolution limitations. In addition, they can serve as a useful guide for the experimentalist regarding the resolutions that may be required for robust fitting of atomic coordinates for α-helical proteins, mixed α/β and β-strand proteins.

3.3. BCL::EM-Fit identifies the correct density for a given atomic resolution structure, homolog, or comparative model

Often atomic resolution structures of proteins are placed into cryoEM density maps of macromolecular systems in order to

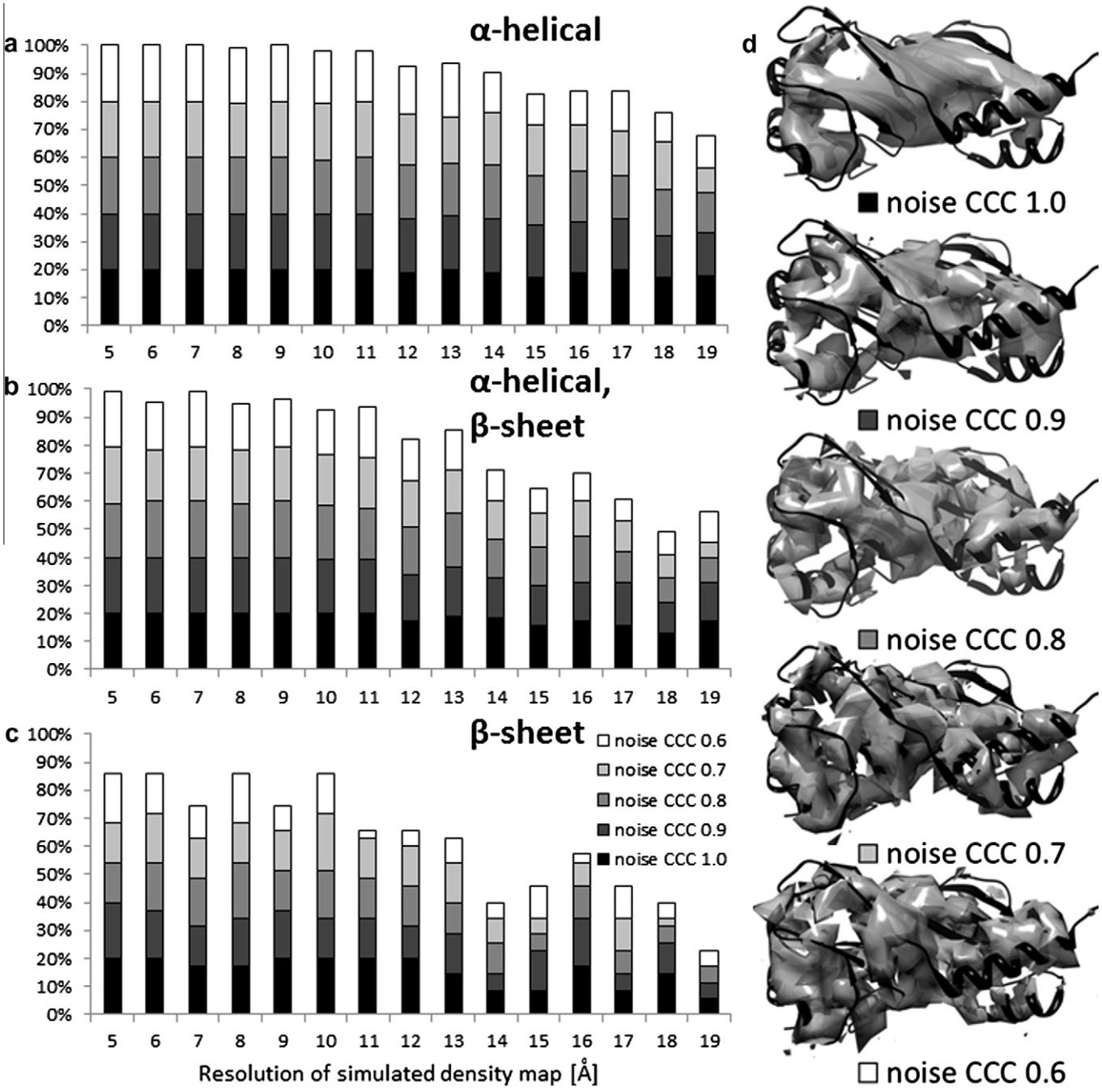

Fig.4. Fitting of benchmark proteins at different resolutions. (a) Results of fitting 21 α-helical proteins into simulated density maps calculated in the resolution range of 5–19 Å both with and without added noise. The *CCCs* of the noise-added maps to the noise-free maps are 0.9, 0.8, 0.7 and 0.6. The *x*-axis represents the resolution of the simulated density map in Å. The *y*-axis represents the percentage of atomic model/simulated map combinations that had at least one fit within the initial 10 best fits by hash score that refined to the correct position (within RMSDCα < 5 Å). The results with noise-free maps (noise *CCC* 1.0) are plotted with black bars, and those with noise-added maps are plotted in shades of gray to white. The maximum height of any bar (noise-free, or with noise) is 20%, corresponding to the percentage for that category of maps. (b) Results of fitting 22 α/β proteins. (c) Results of fitting 7 β-sheet proteins. (d) Simulated density maps for one of the α/β benchmark proteins (1prz) at 10 Å resolution with and without added noise shown together with the input atomic structure.

assign density regions to specific proteins. This proves even more challenging if no experimental atomic resolution structure is available and the structure of a homolog or comparative model is used. To test the robustness of the algorithm in this respect a *cross-fitting* experiment was performed where 9 of the α-helical benchmark proteins were fitted into all 12 Å resolution noise-free density maps (Table S2). The experiment was repeated for 6 β-strand proteins with 11 Å resolution noise-free density maps (Table S3). In all cases the correct match was identified with *CCCs* of 1.00. The best fit into a wrong density map never had a CCC higher than 0.95.

This experiment was repeated using homologous structures, identified by bioinfo.pl meta server (Ginalski et al., 2003), and comparative models generated by MODELLER (Sánchez and Sali, 2000), for three of the α-helical and two β-strand benchmark proteins (Table 1). Density maps were generated with a resolution of 11 Å and with noise levels designed to yield *CCCs* of 0.8 with respect to the noise-free maps. All but one homologous structure showed the highest *CCC* when fitted into the density of the respective homologous protein (Table 1 left). The exception is 1LN1, which is a homolog of β-strand protein 2E3S. In tests with the 1LN1 atom-

ic coordinates, roughly equivalent *CCC* values (0.73–0.75) were found after docking into four different simulated maps. One of these four maps was the intended simulated map for the homolog 2E3S, but there was not a clear peak in the *CCC* value with the correct simulated density map (Table columns). Similarly, the simulated density map for 2E3S had high correlations (0.71–0.75) with coordinates of 3 non-homologues structures (Table rows). The lesson implied by these results, which is not unexpected, is that some protein folds will be more difficult to fit than other folds at certain resolution cutoffs.

Comparative models were built with MODELLER using these same homologous structures as templates and the bioinfo.pl alignment. Details are given in the Methods section. For all comparative models the highest *CCC* value was found for the correct density map, as indicated by the diagonal (Table 1 right). Correct placement of the model into the density was validated by visual inspection. Although the comparative models did not have a significantly higher *CCC* for the fitted structures compared to the values found for the homologous structures (compare Table 1 left and right), the comparative models were fit unambiguously to the correct density maps.

3.4. Adenovirus capsid proteins are docked with high confidence into cryoEM density

The crystal structures for two adenovirus capsid proteins were docked into two experimental cryoEM density maps of adenovirus at 6.8 Å (Saban et al., 2006; Lindert et al., 2009b) and 9.0 Å resolution (Lindert et al., 2009a) (FSC 0.5 criterion). Note that the 6.8 Å resolution map is of the Ad35F (Ad5.F35) vector, which contains human adenovirus type 5 (HAdV5) hexon and penton base capsid proteins. The 9.0 Å resolution map is HAdV12 in complex with integrin and is based on a subset of the full dataset in order to limit the resolution to 9 Å. The penton base structure (pdb: 1X9T) (Zubieta et al., 2005) is a homopentamer (2615 residues) formed by an N-terminally truncated form of HAdV2 penton base (residues 49–571) together with a 21 residue N-terminal tail of the HAdV2 fiber. The hexonstructure (pdb: 1P30) (Rux et al., 2003) is a homotrimer (2853 residues) with 951 residues per monomer of the HAdV5 hexon. The hexon and penton base proteins from HAdV2, 5, and 12 all highly homologous, with percent identities in the range of 77–99%.

The penton base fitting experiments were performed using comparable segments from the same location in the two different resolution cryo-EM density maps (details can be found in the Supplementary Material). The segments contained one tightly cut copy of the penton base oligomer. Due to the fivefold symmetry of the penton base five distinct fitting positions are possible. Three different fits within seven correct solutions were identified by BCL::EM-Fit among the best 10 scoring fits for the 6.8 Å segment (Fig. 5a), two different fits were identified among the 10 best scoring fits within the 9.0 Å density segment. *CCC* values between 0.06 and 0.31 were found for the 6.8 Å segment and *CCCs* between 0.02 and 0.54 for the 9.0 Å segment before the refinement (Table 2).

The MCM refinement procedure was performed on the 10 top-scoring initial placements to optimize the *CCC* further. Details are given in the Methods section. For seven of the initial placements the *CCC* was optimized to 0.53 or better for the 6.8 Å segment; two placements were refined to *CCC* 0.66 for the 9.0 Å segment (Table 2). The accurate placement of the penton base was confirmed visually (6.8 Å segment is shown in Fig. 5b). Comparison of the initial and refined positions for the atomic coordinates yields $RMSD_{C\alpha}$ values in the range of 6.2–11.6 Å, indicating movements on this order during refinement.

The hexon capsid protein was docked into different segments of the reconstructed adenovirus density maps at 6.8 and 9.0 Å resolution, which contained all four independent positions of this protein within the asymmetric unit (details can be found in the Supplementary Material). Seven correct placements were identified in the 6.8 Å resolution density segment, of which four represent sym-

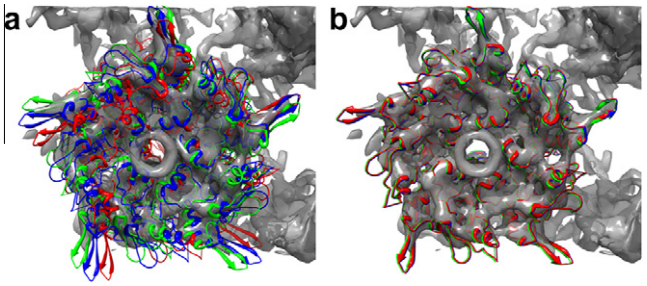

Fig.5. BCL::EM-Fit docking of penton base into adenovirus cryoEM density map segment at 6.8 Å resolution. (a) The best three unique fits out of ten initial fits by *CCC* are shown docked into the cryoEM density segment (gray) displayed with an iso surface level chosen to reveal the strongest density features. (b) The same three fits after 250 steps of MCM refinement. The optimal placement of all three fits is confirmed visually by the good superimposition of α-helices with density rods.

Table 1
Cross-docking *CCC* matrix for benchmark proteins with homologous structures and comparative models.

Density map[a]	Homologous structures[b]					Comparative models[c]				
	1RJK	1PVL	1L2 J	1T5 J	1LN1	1RJK	1PVL	1L2 J	1T5 J	1LN1
%seqsim.	91	71	98	26	17					
CATH	α	β	α	α	αβ	α	β	α	α	αβ
#residues	292	301	271	313	214	259	299	255	303	255
Helix/strand	13/3	3/22	12/2	20/2	6/17	10/3	1/19	8/2	14/0	6/10
$RMSD_{C\alpha}$[d]	2.75	1.51	2.58	3.13	3.92	3.16	1.09	1.68	3.48	5.35
SSE-$RMSD_{C\alpha}$[e]						0.42	0.65	0.91	1.52	3.03
1IE9	**0.82**	0.68	0.74	0.70	0.73	**0.81**	0.67	0.74	0.70	0.70
1LKF	0.68	**0.83**	0.66	0.62	0.63	0.68	**0.82**	0.67	0.62	0.60
1QKM	0.68	0.63	**0.82**	0.72	0.73	0.77	0.63	**0.81**	0.73	0.71
2CWC	0.72	0.58	0.73	**0.79**	0.75	0.72	0.58	0.73	**0.81**	0.74
2E3S	0.73	0.60	0.71	**0.75**	0.73	0.71	0.61	0.74	0.74	**0.78**

[a] Simulated density maps for five proteins:three α-helical (1IE9, 1QKM, 2CWC) and two β-strand (1LKF, 2E3S) at 11 Å resolution and with added noise (*CCC* 0.8 with respect to noise-free map).
[b] Homologous structures were identified with bioinfo.pl.
[c] Comparative models were built for 1IE9, 1LKF, 1QKM, 2CWC, and 2E3S from the homologous structures (1RJK, 1PVL, 1L2J, 1T5J, and 1LN1, respectively) using MODELLER.
[d] $RMSD_{C\alpha}$ of the original PDB vs. the homologous structure (using mammoth structure alignment) and vs. the comparative model.
[e] SSE-$RMSD_{C\alpha}$ only using secondary structure elements that are common to both PDBs.

Table 2
Docking of penton base into adenovirus cryoEM density mapsat 6.8 and 9.0 Å resolution with BCL::EM-Fit.

Map resolution [Å]	Rank by hash score	Hash score	Initial *CCC*	Optimized[b] *CCC*	$RMSD_{C\alpha}$[c] of optimized to initial fit [Å]
6.8	**5**[a]	**181**	**0.18**	**0.54**	**11.59**
6.8	**1**[a]	**192**	**0.31**	**0.53**	**6.19**
6.8	**4**[a]	**182**	**0.29**	**0.53**	**6.32**
6.8	**6**[d]	**181**	**0.18**	**0.53**	**10.03**
6.8	**10**[d]	**179**	**0.19**	**0.53**	**12.23**
6.8	**3**[d]	**186**	**0.30**	**0.53**	**6.65**
6.8	**2**[d]	**191**	**0.27**	**0.53**	**8.29**
6.8	8	181	0.14	0.16	6.07
6.8	9	180	0.10	0.12	6.91
6.8	7	181	0.06	0.10	7.49
9.0	**1**[a]	**128**	**0.54**	**0.66**	**9.28**
9.0	**2**[a]	**128**	**0.48**	**0.66**	**11.29**
9.0	4	126	0.15	0.32	16.59
9.0	3	127	0.29	0.31	2.73
9.0	6	125	0.19	0.31	17.58
9.0	8	125	0.12	0.18	11.83
9.0	7	125	0.10	0.13	8.80
9.0	9	125	0.02	0.12	12.31
9.0	10	125	0.04	0.12	12.53
9.0	5	126	0.05	0.12	13.18

[a,d] All of the fits that are correct have a high *CCC* value after optimization (bold).
[a] Best independent fits after MCM optimization by *CCC*. The three best fits that yield different placements with respect to the 6.8 Å resolution map are shown in Fig. 5a.
[b] MCM refinement (see Section 2). The refined positions of the three best independent fits with respect to the 6.8 Å resolution map are shown in Fig. 5b.
[c] The$RMSD_{C\alpha}$ of initial to refined fit is shown to indicate the amount of movement of the atomic model during the refinement step.
[d] Fits which duplicate positions of the three best fits marked '[a]'.

metrically independent, non-overlapping positions in the asymmetric unit (Table S4). These four initial fits have *CCCs* above 0.13, with the best being 0.25. Fig. S1 shows superimpositions of the transformed hexon with the 6.8 Å resolution density segment confirming correct placements for this protein. A MCM refinement was performed on the 50 best initial placements. After optimization the *CCCs* for the symmetrically unrelated copies were in the range of 0.47–0.48 (Table S4 and Fig. S2). Ten correct placements were identified in the 9.0 Å resolution density segment, of which four are symmetrically independent and non-overlapping positions. These four positions have *CCCs* above 0.53. After MCM refinement *CCCs* are between 0.68 and 0.73 (Table S4).

The adenovirus capsid protein fitting experiments indicate that the BCL::EM-Fit algorithm can identify initial fits of the atomic structures in question. The subsequent MCM refinement procedure delivers results in visually improved fits with higher CCCs.

3.5. Four copies of 1OELG are docked into the chaperonin GroEL density map at 5.4 Å resolution

A single chain (id: G) of the crystal structure of the chaperonin GroEL (pdb: 1OEL) (Braig et al., 1995) was docked into the complete 5.4 Å resolution density map of GroEL (EMDB: 1457) (Stagg et al., 2008; Lawson et al., 2011). GroEL is a dual heptameric particle with a main 7-fold axis and a perpendicular 2-fold axis (dihedral 7-fold symmetry). Detailed density derived parameters can be found in the Supplementary Material.

The BCL::EM-Fit algorithm identified six correct fits (Table S5) which could be confirmed visually. Four of them are in different positions (Fig. S3). Initial fits had *CCCs* between 0.39 and 0.62; refined fits had *CCCs* between 0.62 and 0.75. The entire procedure took 51 min on a single core of an Intel(R) Xeon(R) CPU W3570 @ 3.20 GHz.

Table 3
Comparison of the initial fitting and refinement step by BCL::EM-Fit for penton base into the correct and the symmetry-inverted density maps at 6.8 Å resolution.

Correct				Flipped[a]			
Rank by hash score	Hash score	Initial *CCC*	Optimized *CCC*	Rank by hash score	Hash score	Initial *CCC*	Optimized *CCC*
10	179	0.19	0.54	4	177	0.16	0.27
2	191	0.27	0.53	6	175	0.18	0.27
3	186	0.30	0.53	2	179	0.17	0.18
6	181	0.19	0.53	8	173	0.06	0.15
5	181	0.19	0.53	7	175	0.10	0.13
1	192	0.31	0.53	3	179	0.11	0.12
4	182	0.29	0.53	1	179	0.10	0.11
8	181	0.14	0.17	9	173	0.08	0.10
9	180	0.10	0.16	0	179	0.07	0.08
7	181	0.06	0.07	5	175	0.05	0.08
Mean	183	0.20	0.41		176	0.11	0.15
SD	5	0.09	0.19		3	0.05	0.07

[a] The flipped density map was created to have the opposite handedness compared to the correct density map.

3.6. Correct handedness of a density maps can be verified by the CCC of the initial fit

Imaging a macromolecular assembly by transmission electron microscopy results in the loss of the absolute hand of the structure because the three-dimensional information is projected into a two-dimensional plane. Several methods for determining the absolute hand of a cryoEM single particle reconstruction have been developed, which involve collecting tilted images (Belnap et al., 1997; Rosenthal and Henderson, 2003). Often however the absolute hand of a cryoEM structure is not experimentally determined, and thus both possible hands of the density should be tested when docking atomic resolution structures. To test the BCL::EM-Fit algorithm's ability to distinguish correct from incorrect handedness, two versions of the experimental density map segment around the adenovirus penton base were created (correct and flipped). The refined fits for the correct map have *CCCs* of as high as 0.54. In contrast, the refined fits for the flipped map have a *CCC* only as high as 0.27 (Table 3). This indicates that given a density map with a sufficiently high resolution (6.8 Å resolution in this example), the BCL::EM-Fit algorithm can differentiate between the two possible hands of the density map and select the map with the correct hand.

5. Discussion

5.1. Docking works best when secondary structural elements are resolved within the density map

A new algorithm, BCL::EM-Fit, is presented for rapid and accurate docking of atomic resolution structures within moderate resolution (5–12 Å) density maps. The protocol consists of feature extraction from the density map and encoding of this information into a geometric hash map, followed by searching of the hash map with features extracted from the coordinate file of an atomic resolution structure or model. The resulting initial fits are then refined in an MCM refinement step. Docking experiments with benchmark proteins demonstrate reliable fitting of atomic structures if the density map has a resolution of ~10 Å or better. The docking experiments also indicate that the *CCC* between simulated and experimental density maps is a satisfactory way to identify optimal positions, since the highest CCC is observed for positions that have an RMSD <5 Å to the correct placement.

Benchmark tests were performed with α-helical proteins, mixed α/β-proteins, and predominantly β-strand proteins. The algorithm works reliably for α-helical proteins with nearly no incorrect fits at resolutions up to 12 Å. The algorithm also works well for α/β and β-strand proteins for resolutions up to ~11 or 10 Å, respectively. The better performance of BCL::EM-Fit with mostly α-helical proteins is attributed to the fact that α-helices can be resolved at more moderate resolution than β-strands (Zhou, 2008). For resolutions in the range of 12–19 Å the secondary structure elements that help to accurately position atomic models are not well enough resolved for the BCL::EM-Fit algorithm to find the correct fit in all cases.

5.2. BCL::EM-Fit correctly identifies and places homologous structures and comparative models

A cross fitting experiment with five simulated density maps and homologous structures or comparative models was performed (Table 1). The ambiguous docking results with one simulated density map (that of 2E3S, a mostly β-strand benchmark protein) might have been alleviated if higher resolution density maps were used. The results indicate that BCL::EM-Fit works reasonably well with both homologous structures and comparative models, however better docking results were obtained with comparative models.

5.3. BCL::EM-Fit is applicable to fitting of large adenovirus capsid proteins

For human adenovirus penton base and hexon capsid protein were fitted within 6.8 and 9 Å resolution sections of experimental cryoEM density maps of the entire virus. The generated fits of the atomic resolution protein structures cover all symmetrically unrelated placements which can be used to rebuild the 3D structure of the entire virus capsid. BCL::EM-Fit was further capable of identifying the correct handedness of the reconstructed cryoEM density map by superior hash score and *CCCs* at the initial and refinement stage of fitting.

5.4. BCL::EM-Fit can fit subunits within a larger assembly

In addition to the tests with the multimeric adenovirus capsid proteins, BCL::EM-Fit was also used to successfully fit a single chain of 1OEL into the GroEL density map at 5.4 Å resolution. Although only four of the 14 copies were found, the knowledge of the7-fold dihedral symmetry of GroEL would enable the construction of the complete assembly from only one correctly docked subunit. Alternatively, one could refine more of the initial fits and expect to find more independent positions at the cost of a longer fitting time.

5.5. BCL::EM-Fit and flexible docking

All benchmarks and examples shown here are rigid body fitting experiments that provide an initial fit. This experimental design allows testing the geometric hashing approach which is tailored for the rigid body fitting problem. One possible way to explore protein flexibility on the domain level is to separate the coordinates of the protein of interest into independent domains and fit them into the density map separately. Internal flexibility could be simulated with Molecular Dynamics programs and a selected set of representative conformations could be saved and subsequently fit into a density map. Additional tools have been developed that perform flexible docking once an initial fit is identified, e.g. using BCL::EM-Fit. These include Qplasty in the SITUS package (Wriggers and Birmanns, 2001), Rosetta (Tyka et al., 2009), molecular dynamics flexible docking (MDFF) (Trabuco et al., 2009) and DireX (Schröder et al., 2007).

5.6. Advantages and disadvantages of Geometric Hashing compared to Fourier/Real Space fitting

The geometric hashing approach is presented as an alternative method for fitting atomic resolution structures into multiple positions within large density maps. The BCL::EM-Fit results demonstrate good performance for fitting proteins into density maps of a resolution up to 12 Å. All orientations and positions of interest for the hexon and penton base proteins in adenovirus could be determined within sections of the virus density map at 6.8 and 9 Å resolution. A time comparison to the exhaustive Fourier/Real Space search method as implemented in Colores revealed a 3-fold advantage for BCL::EM-Fit using a single CPU (Supplementary Material). Colores may still be advantageous in several scenarios. It samples all regions of the density map evenly and therefore it can identify matches that might be missed by the geometric hashing approach. This is especially true for lower resolution density maps (>12 Å) that often lack distinctive features. A second advantage relates to the fact that closely packed protein domains in obligate oligomers might appear as one continuous domain to the feature matching algorithm of BCL::EM-Fit. In cases like this a Fourier/Real Space search has an increased chance of identifying all monomeric copies of the protein. These disadvantages of BCL::EM-Fit will be addressed in future versions of the program. Nevertheless, given the growing importance of docking atomic models into cryoEM density maps it should prove useful to have multiple algorithms to accomplish this task.

6. Conclusions

The intensities in a cryoEM density map represent structural features of rigid and dense parts of the structure, in particular secondary structure elements at resolutions better than ~10 Å. The position of these features can be pre-encoded in a geometric hash map. Using the C_α atom positions in α-helices and β-strands, atomic models can be fit into density maps by enumerating features in common between the density map and the atomic model. In BCL::EM-Fit tests presented here with both simulated and experimental density, initial fits that led to correct positions during refinement were distinguishable by their CCC. The accuracy of the final fit is dependent on the resolution of the density map, the voxel size within the density map, and the resolution that is used to quantize the features within the hash map. MCM optimization with rigid body perturbation quickly and reliably refines the initial fit to a fit with the maximum CCC between the experimental and the simulated density map created from the atomic model. The BCL::EM-Fit algorithm provides an alternative method for docking of atomic models within cryoEM density maps.

Acknowledgments

The authors acknowledge the help of Susan Saban, who was engaged in the initial discussions for developing an alternative fitting algorithm. N.W. is supported by the Warren research fellowship of the chemistry department at Vanderbilt University, Nashville, TN. J.M. and N.W. are supported by grant 0742762 from the National Science Foundation and grant R01-GM080403 National Institutes of Health. P.L.S. is supported by grants R01-AI42929 and R01-CA140538 and R01-CA141439 from the National Institutes of Health.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jsb.2011.04.016.

References

Belnap, D.M., Olson, N.H., Baker, T.S., 1997. A method for establishing the handedness of biological macromolecules. Journal of structural biology 120, 44–51.

Braig, K., Adams, P.D., Brünger, A.T., 1995. Conformational variability in the refined structure of the chaperonin GroEL at 2.8 A resolution. Nature structural biology 2, 1083–1094.

Dutta, S., Berman, H.M., 2005. Large macromolecular complexes in the Protein Data Bank: a status report. Structure (London, England: 1993) 13, 381–388.

Fabiola, F., Chapman, M.S., 2005. Fitting of high-resolution structures into electron microscopy reconstruction images. Structure (London, England: 1993) 13, 389–400.

Ginalski, K., Elofsson, A., Fischer, D., Rychlewski, L., 2003. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics (Oxford, England) 19, 1015–1018.

Goddard, T.D., Huang, C.C., Ferrin, T.E., 2007. Visualizing density maps with UCSF Chimera. Journal of structural biology 157, 281–287.

Jiang, W., Baker, M.L., Ludtke, S.J., Chiu, W., 2001. Bridging the information gap: computational tools for intermediate resolution structure interpretation. Journal of molecular biology 308, 1033–1044.

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., et al., 1958. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. Nature 181, 662–666.

Korostelev, A., Bertram, R., Chapman, M.S., 2002. Simulated-annealing real-space refinement as a tool in model building. Acta Crystallographica Section D Biological Crystallography 58, 761–767.

Lawson, C.L., Baker, M.L., Best, C., Bi, C., Dougherty, M., et al., 2011. EMDataBank.org: unified data resource for CryoEM. Nucleic acids research 39, D456–D464.

Lepault, J., Booy, F.P., Dubochet, J., 1983. Electron microscopy of frozen biological suspensions. Journal of microscopy 129, 89–102.

Lindert, S., Silvestry, M., Mullen, T.-M., Nemerow, G.R., Stewart, P.L., 2009a. Cryo-electron microscopy structure of an adenovirus-integrin complex indicates conformational changes in both penton base and integrin. Journal of Virology 83, 11491–11501.

Lindert, S., Staritzbichler, R., Wötzel, N., Karakaş, M., Stewart, P.L., et al., 2009b. EM-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps. Structure (London, England: 1993) 17, 990–1003.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. The Journal of Chemical Physics 21, 1087.

Metropolis, N., Ulam, S., 1949. The Monte Carlo method. Journal of the American Statistical Association 44, 335–341.

Roseman, A.M., 2000. Docking structures of domains into maps from cryo-electron microscopy using local correlation. Acta Crystallographica. Section D, Biological Crystallography 56, 1332–1340.

Rosenthal, P.B., Henderson, R., 2003. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. Journal of Molecular Biology 333, 721–745.

Rux, J.J., Kuser, P.R., Burnett, R.M., 2003. Structural and phylogenetic analysis of adenovirus hexons by use of high-resolution X-ray crystallographic, molecular modeling, and sequence-based methods. Journal of Virology 77, 9553–9566.

Saban, S.D., Silvestry, M., Nemerow, G.R., Stewart, P.L., 2006. Visualization of alpha-helices in a 6-angstrom resolution cryoelectron microscopy structure of adenovirus allows refinement of capsid protein assignments. Journal of Virology 80, 12049–12059.

Schröder, G.F., Brunger, A.T., Levitt, M., 2007. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. Structure (London, England: 1993) 15, 1630–1641.

Shulman-peleg, A., Nussinov, R., Wolfson, H.J., 2004. Recognition of functional sites in protein structures. Journal of Molecular Biology 339, 607–633.

Stagg, S.M., Lander, G.C., Quispe, J., Voss, N.R., Cheng, A., et al., 2008. A test-bed for optimizing high-resolution single particle reconstructions. Journal of Structural Biology 163, 29–39.

Sánchez, R., Sali, A., 2000. Comparative protein structure modeling. Introduction and practical examples with modeller. Methods in molecular biology (Clifton, NJ) 143, 97–129.

Tama, F., Miyashita, O., Brooks, C.L., 2004. Normal mode based flexible fitting of high-resolution structure into low-resolution experimental data from cryo-EM. Journal of Structural Biology 147, 315–326.

Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W., et al., 2008. Protein structure fitting and refinement guided by cryo-EM density. Structure (London, England: 1993) 16, 295–307.

Trabuco, L.G., Villa, E., Schreiner, E., Harrison, C.B., Schulten, K., 2009. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. Methods (San Diego, Calif.) 49, 174–180.

Tyka, M.D., Dimaio, F., Baker, M.L., Chiu, W., Baker, D., 2009. Refinement of protein structures into low-resolution density maps using rosetta. Journal of Molecular Biology 392, 181–190.

Urzhumtseva, L., Urzhumtsev, A., 2002. COMPANG: automated comparison of orientations. Journal of Applied Crystallography 35, 644–647.

Velazquez-muriel, J.A., Carazo, J.-M.A., 2007. Flexible fitting in 3D-EM with incomplete data on superfamily variability. Journal of Structural Biology 158, 165–181.

Woetzel, N., Lowe, E.W., Meiler, J., 2011. *Poster: GPU-accelerated rigid body fitting of atomic structures into electron density maps.* 2011 IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences (ICCABS).

Wolfson, H.J., Rigoutsos, I., 1997. Geometric hashing: an overview. IEEE Computational Science and Engineering 4, 10–21.

Wriggers, W., Birmanns, S., 2001. Using situs for flexible and rigid-body fitting of multi resolution single-molecule data. Journal of structural biology 133, 193–202.

Wriggers, W., Chacón, P., 2001. Modeling tricks and fitting techniques for multi resolution structures. Structure (London, England: 1993) 9, 779–788.

Wriggers, W., Milligan, R.A., Mccammon, J.A., 1999. Situs: a package for docking crystal structures into low-resolution maps from electron microscopy. Journal of Structural Biology 125, 185–195.

Wüthrich, K., 1990. Protein structure determination in solution by NMR spectroscopy. The Journal of Biological Chemistry 265, 22059–22062.

Zhou, Z.H., 2008. Towards atomic resolution structural determination by single-particle cryo-electron microscopy. Current opinion in structural biology 18, 218–228.

Zubieta, C., Schoehn, G., Chroboczek, J., Cusack, S., 2005. The structure of the human adenovirus 2 penton. Molecular Cell 17, 121–135.