Journal of Structural Biology xxx (2011) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Journal of Structural Biology

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

Finding rigid bodies in protein structures: Application to flexible fitting into cryoEM maps

Arun Prasad Pandurangan, Maya Topf*

Institute of Structural and Molecular Biology, Crystallography/Department of Biological Sciences, Birkbeck College, University of London, London WC1E 7HX, United Kingdom

ARTICLE INFO

Article history:
Received 20 August 2011
Received in revised form 22 October 2011
Accepted 27 October 2011
Available online xxxx

Keywords:
Flexible fitting
Rigid bodies
Clustering
Optimisation
Low-resolution maps
Cryo electron microscopy density fitting

ABSTRACT

We present RIBFIND, a method for detecting flexibility in protein structures via the clustering of secondary structural elements (SSEs) into rigid bodies. To test the usefulness of the method in refining atomic structures within cryoEM density we incorporated it into our flexible fitting protocol (Flex-EM). Our benchmark includes 13 pairs of protein structures in two conformations each, one of which is represented by a corresponding cryoEM map. Refining the structures in simulated and experimental maps at the 5–15 Å resolution range using rigid bodies identified by RIBFIND shows a significant improvement over using individual SSEs as rigid bodies. For the 15 Å resolution simulated maps, using RIBFIND-based rigid bodies improves the initial fits by 40.64% on average, as compared to 26.52% when using individual SSEs. Furthermore, for some test cases we show that at the sub-nanometer resolution range the fits can be further improved by applying a two-stage refinement protocol (using RIBFIND-based refinement followed by an SSE-based refinement). The method is stand-alone and could serve as a general interactive tool for guiding flexible fitting into EM maps.

1. Introduction

One of the keys to understanding cellular processes at the molecular level is the determination of the structures of macromolecular assemblies (Alber et al., 2008). In recent years, cryo electron microscopy (cryoEM) has become one of the most prominent techniques in the field (Frank, 2006; Lawson et al., 2011). Combined with image processing, single-particle cryoEM has been applied to characterise the purified forms of assemblies at different conformational states, mostly at intermediate (~5–15 Å) and low (>15 Å) resolutions (Frank, 2009). It is almost standard now to obtain pseudo-atomic models of those assemblies by fitting into them atomic structures of components of proteins and nucleic acids if available (from X-ray crystallography, NMR spectroscopy or comparative and *ab initio* modelling) (Beck et al., 2011; Chiu et al., 2005; Fabiola and Chapman, 2005; Rossmann et al., 2005). Manual fitting using visualisation programs such as Chimera (Pettersen et al., 2004) can be affected heavily by user bias and can become obstinately difficult, especially with maps of large assemblies containing many different components. Automated fitting can alleviate these problems and therefore has become increasingly popular.

Indeed, automated rigid fitting has been very successful in providing many pseudo-atomic models of macromolecular assemblies (Fabiola and Chapman, 2005; Rossmann et al., 2005; Topf and Sali, 2005; Wriggers and Chacon, 2001). In most of these methods, the goodness-of-fit measure for the placement of an atomic structure in a density map is the cross-correlation between the cryoEM density map and a simulated density map of the structure, computed by convolving its atomic coordinates with a point-spread function (Fabiola and Chapman, 2005). However, the isolated component structure may exhibit a different conformation than that reflected in the assembly density map due to the experimental conditions under which it was determined or errors in modelling (Alber et al., 2008; Baker and Sali, 2001; Topf et al., 2008). In addition, heterogeneity in the imaged sample often results in a number of maps describing different conformational states of the intact assembly (Spahn and Penczek, 2009). Thus there is often a need to modify the position and orientation not only of the entire component but also of its parts, a process referred to as "flexible fitting".

One way to tackle this problem is to divide the atomic structure of the component into rigid bodies, such as domains, and fit each of them independently into the map (Volkmann et al., 2000; Wendt et al., 2001). This approach often results in the distortion of the mechanical properties of the structure. A more objective approach is to generate multiple "valid" conformations for the component and select the top ranking conformation based on its fit into the density (Topf et al., 2005). The component structure is usually first placed into the density map by rigid fitting to reduce the sampling of degrees of freedom and candidate conformations are then

* Corresponding author.
E-mail address: m.topf@cryst.bbk.ac.uk (M. Topf).
URL: http://www.cryst.bbk.ac.uk/~ubcg67a (M. Topf).

1047-8477/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2011.10.011

generated using normal mode analysis (NMA) (Ma, 2005; Suhre et al., 2006; Tama et al., 2002), comparative modelling (Chandramouli et al., 2008; Rawi et al., 2010; Taylor et al., 2009; Topf et al., 2006), *ab intio* modelling (Baker et al., 2006), geometric hashing (Woetzel et al., 2011) or by exploring the structural variability of protein domains within a given superfamily (Velazquez-Muriel et al., 2006).

An alternative approach is to simultaneously refine the position, orientation, and conformation of the component structure in the cryoEM map (Fabiola and Chapman, 2005) while maintaining its mechanical properties (Beck et al., 2011; Chen et al., 2003; DiMaio et al., 2009; Grubisic et al., 2010; Schroder et al., 2007; Topf et al., 2008; Trabuco et al., 2008; Zheng, 2011). Many refinement methods optimise the conformation using coarse-grained approaches, including grouping atoms together into rigid bodies connected by flexible regions (Beck et al., 2011). This can be done manually or by automated methods, such as those based on graph theory (Jacobs et al., 2001; Jolley et al., 2008); hinge identification based on energetic interaction (Flores and Gerstein, 2007) and a comparison of pairs of proteins (Abyzov et al., 2010; Hayward and Berendsen, 1998; Wriggers and Schulten, 1997). Unfortunately, the use of rigid bodies can often limit the conformational degrees of freedom of the atomic structure in ways detrimental to the fitting process. If the number of rigid bodies is too small, the optimisation may not reach the global minimum because a more detailed modification of the conformation is needed. On the other hand, if an all-atom representation is chosen the computational efficiency is largely reduced and the system is likely to get trapped in local minima. An optimal partitioning of the structure into flexible and rigid bodies would help guide the trajectory of the optimisation, and result in a better fit.

Here we introduce RIBFIND, a new method for finding rigid bodies in protein structures based on the clustering of SSEs (Section 2). By incorporating the method into our flexible fitting protocol Flex-EM (Topf et al., 2008), we show how flexible fitting of atomic structures into cryoEM maps can be significantly improved by a superior partition of rigid bodies and flexible regions. We tested RIBFIND on a benchmark of 10 protein structures. Each of these was refined into a simulated density map representing a differing (known) conformation at 5–15 Å resolution (Section 3). We also tested the method on three structures using experimentally determined cryoEM maps at the same resolution range (Section 3). Finally, we discuss our approach and its implications for refining structures and models using cryoEM density maps (Section 4).

2. Methods

2.1. Neighbourhood-based clustering

The RIBFIND method identifies rigid bodies in protein structures using a clustering approach based on the spatial proximity between secondary structural elements (SSEs). Fig. 1a shows the steps involved. Starting from the input atomic coordinates, the SSEs are assigned using the program DSSP (Kabsch and Sander, 1983). These SSEs form an initial pool of members for clustering. Next, a temporary "neighbourhood" list is created by adding an SSE member selected randomly from the pool. All the neighbours from the pool spatially proximal to this SSE are then identified and added to the list. This process is repeated iteratively for each member in the list until no new members can be added from the pool. After completion of the iterative process, if the list contains more than one SSE member then all the members are removed from the pool, and a "cluster" is formed by these members; otherwise the individual SSE member is added to a non-clustered list. The whole process of clustering is repeated as long as the pool contains more than one member. Finally, the method outputs each cluster, and the loops connecting its SSE members, as a rigid body. Each non-clustered SSE is defined as an individual rigid body, and each of the atoms connecting them is treated as an individual rigid body. The list of rigid bodies can then be used in flexible fitting.

The method defines two parameters to measure the spatial proximity between any two SSEs (see Fig. 1b). The first parameter is the *residue contact distance*, which represents the contact between any two residues. This parameter is measured by the distance between the side-chain centroids of the two residues (Miyazawa and Jernigan, 1996). The second parameter is the *cluster cutoff*, which helps to cluster any two SSEs based on the percentage of residues in contact between them. The values of this parameter run from 0% to 100%. Also, we do not treat any two helices in the protein to form one cluster if the number of residues in either one of the helix is less than the other by 40%. This condition will avoid building up unrealistic clusters caused by long helices. Both the contact distance and cluster cutoff can be user defined and can be varied to achieve different levels of clusters. Their choice is discussed in Section 3.

2.2. Application to flexible fitting

To test the improvement in flexible fitting using RIBFIND we use our Flex-EM method (Topf et al., 2008). The method can optimise the position and conformation of an atomic structure in a cryoEM map in three stages: (i) a Monte Carlo search; (ii) a conjugate-gradients minimisation; and (iii) a simulated annealing molecular dynamics (MD) refinement. The scoring function of Flex-EM includes stereochemical and non-bonded interaction terms in addition to the cross-correlation function (CCF) term. We typically apply the method at the 'domain level' first (*i.e.*, the rigid bodies correspond to domains and the individual atoms that connect the domains), followed by the 'SSE level' (*i.e.*, the rigid bodies correspond to the SSEs and the individual atoms that connect them) (Topf et al., 2008). However, the method is flexible, allowing any of the optimisation procedures to be applied to any groups of rigid bodies, including user-defined rigid bodies (for example, based on prior knowledge of the structure or visual inspection in the context of the density). For our purposes we applied the Flex-EM simulated annealing MD refinement to two sets of rigid bodies – one based on RIBFIND (*i.e.*, the 'clustered set') and the other based on the SSE level (*i.e.*, the 'non-clustered set').

2.3. Benchmark

The refinement protocol was tested on two benchmarks representing proteins with a maximal number of five domains mostly from α/β fold classes. The first benchmark (*simulated benchmark*) contained 10 proteins (see Table 1) with two conformations each. From each of these pairs of conformations, one conformation was defined as the *initial conformation*. The other conformation (*target conformation*) was used for simulating density maps (*target maps*) at 5, 10, and 15 Å resolution. To minimise bias, the simulated maps were not produced with the same blurring technique used for fitting (as described in Topf et al., 2005), but with the *molmap* command in Chimera (Pettersen et al., 2004), using a different sigma factor for the Gaussian blurring (sigma factor = 0.356 × resolution, grid spacing = resolution/3). Five of the 10 pairs of benchmark proteins (PDB IDs of the initial conformation: 1wdnA, 1f6mE, 1k89A, 1mrpA, and 1dpeA) were selected from the molecular motions database (Flores et al., 2006) that stores experimentally determined structures of macromolecules in two distinct conformations. For actin, the initial (act2A) and target (act1A) conformations are the models of actin taken from a Situs package tutorial (Wriggers, 2010). The remaining pairs of proteins (PDB IDs of the

Please cite this article in press as: Pandurangan, A.P., Topf, M. Finding rigid bodies in protein structures: Application to flexible fitting into cryoEM maps. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.10.011

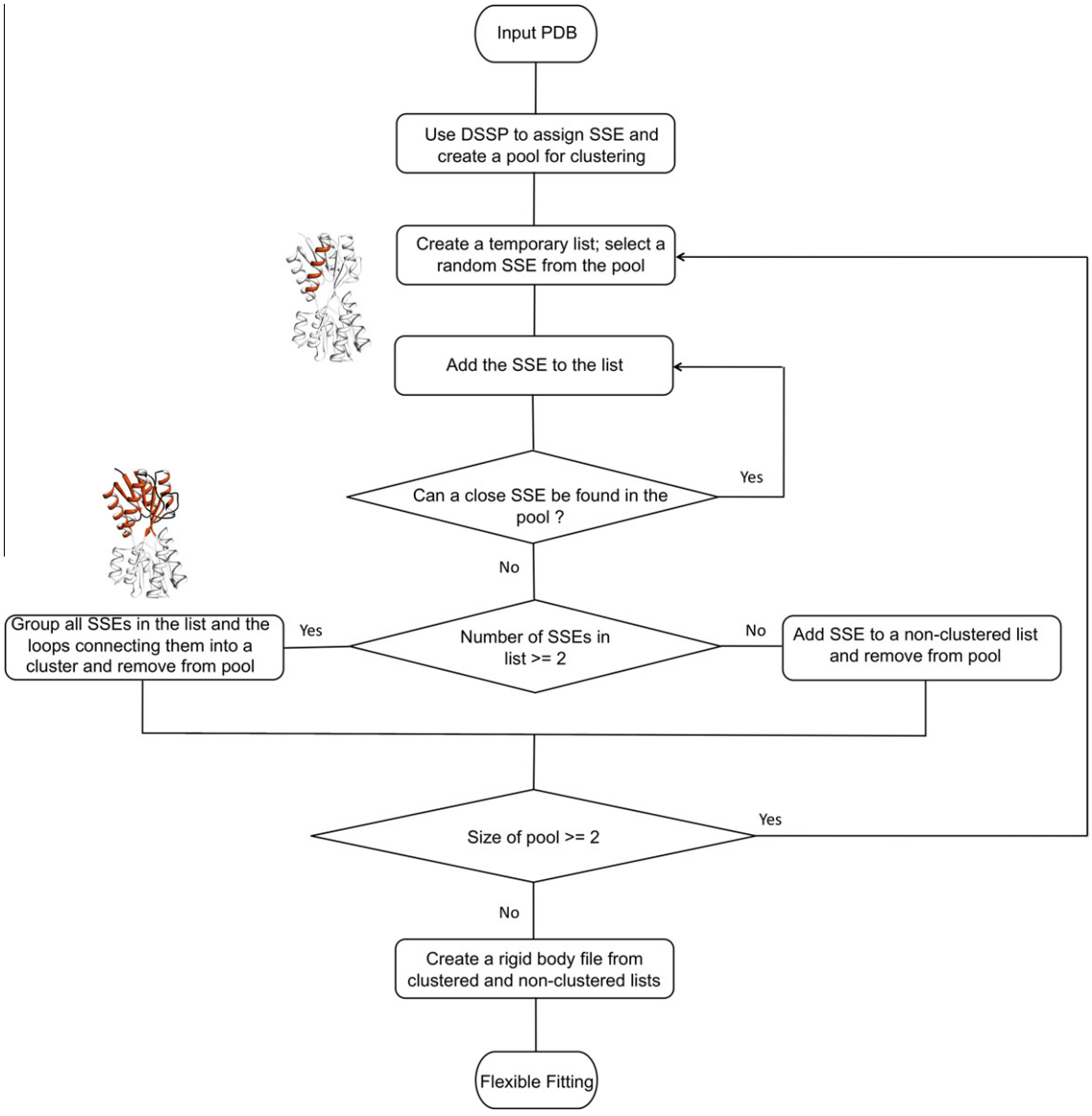


Fig.1a. A flowchart of the RIBFIND method for finding sets of rigid bodies based on the spatial proximity between secondary structure elements (SSEs). The protein under investigation is shown in a cartoon representation. The input to the program is a structure of a protein (PDB format) and the output is a set of rigid bodies, each of which is composed of a group of SSEs selected within a cluster and the loops connecting them.

initial conformation: 1oaoD, 1lfgA, 2driA, and 1ompA) were taken from other studies (Grubisic et al., 2010).

To demonstrate the ability of the method to refine atomic structures in more realistic cases, we tested it on experimentally determined cryoEM maps. These are included in a second benchmark (*experimental benchmark*) containing three pairs of proteins, for which experimental density maps were used (see Table 1). These proteins are a monomer of actin, the elongation factor EF4 (lepA), and a subunit of the heptameric ring of the GroEL chaperonin. All target fits in the experimental benchmark are based on fits deposited in the PDB. These are treated as the "gold standard" fits for our quality assessment of fits. For the actin monomer, the initial conformation was taken from the Situs package tutorial (Wriggers, 2010) (act2A), the experimental map of rabbit actin at 6.6 Å resolution from the EM database (EMDB) (EMD-5168) (Lawson et al., 2011), and the corresponding fit (target conformation) from the PDB (PDB ID: 3mfpA) (Fujii et al., 2010). For EF4, the initial conformation was the crystal structure of EF4 from *Escherichia coli* (PDB ID: 3cb4F) (Evans et al., 2008). The experimental map at 11 Å and its corresponding target conformation were taken from EMDB (EMD-1524) and PDB (PDB ID: 3degC), respectively (Connell et al., 2008). For GroEL, the initial conformation was a subunit of the crystal structure from *E. coli* (PDB ID: 1oelA) (Braig et al., 1995). The experimental map at 15 Å and its corresponding target conformation were taken from EMDB (EMD-1047) and PDB (PDB ID: 2c7eA), respectively (Ranson et al., 2001).

For each of the 13 pairs of benchmark proteins, the initial conformation was first superposed onto its target and then fitted rigidly in the corresponding density map, using the *match* and *fit_in_map* commands in Chimera, respectively. Flex-EM MD refinement was then performed for 15 cycles, except for two proteins (1oelA and 1oaoD), which required 35 cycles due to their large size. The resulting conformation was defined as the *final conformation*.

2.4. Measures of accuracy

We have used the following four scores to assess the accuracy of a structure in a given conformation.

Please cite this article in press as: Pandurangan, A.P., Topf, M. Finding rigid bodies in protein structures: Application to flexible fitting into cryoEM maps. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.10.011

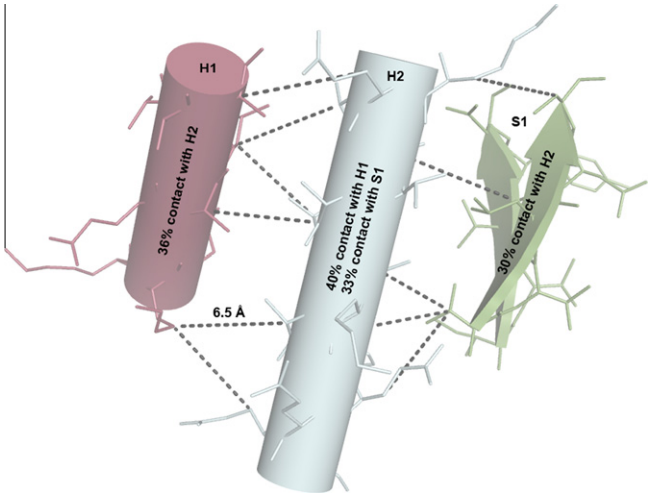


Fig.1b. A schematic diagram representing the RIBFIND parameters – contact distance and cluster cutoff. The two helices (H1 and H2) and the sheet (S1) are shown in a cartoon representation and their corresponding side-chains in a stick representation. In this illustration, the number of residues in H1, H2, and S1 are 11, 15, and 10, respectively. Only the contacts between pairs of residues for which the distance between their side-chains centroids is within 6.5 Å are shown as dotted lines. H1 has 36% of its residues in contact with H2 and H2 has 40% of its residues in contact with H1. H2 has 33% of its residues in contact with S1 and S1 has 30% of its residues in contact with H2. Thus, each of the three SSEs shows 30% or more residues in contact with the nearby SSEs. A *cluster cutoff* of 30% or more will cluster all the three SSEs into one cluster.

2.4.1. Cα RMSD

The root-mean square deviation (RMSD) is calculated between the Cα atoms of a given conformation (*i.e.*, at the initial conformation, a conformation during refinement, or at the final conformation) and the corresponding atoms in the target conformation. The score is calculated upon superposition of the two structures using a rigid-body least-squares minimisation, as implemented in the *superpose* method of MODELLER-9.0 (Sali and Blundell, 1993).

2.4.2. minRMSD

We calculated a "minimal" Cα RMSD (*minRMSD*) for each conformation, corresponding to the best-possible conformation that our refinement could theoretically reach given that the SSEs (in the non-clustered set) or the clusters of atoms (in the clustered set) are treated as rigid bodies. For the non-clustered set, the best-possible conformation has all loop atoms overlapping perfectly with the equivalent target positions and each SSE in the initial conformation superposed independently onto the corresponding SSE in the target conformation. For the clustered set, the best-possible conformation has each cluster in the initial conformation superposed onto its corresponding cluster in the target conformation, and all the non-clustered SSEs in the initial conformation superposed independently onto the corresponding SSE in the target conformation; the rest of the atoms (non-clustered loops) overlap perfectly with the equivalent target positions.

2.4.3. Area-based component placement score (ACPS)

An assessment score previously used by us, the component placement score (CPS) (Lasker et al., 2009; Topf et al., 2008; Zhang et al., 2010), was implemented here in a modified fashion. The CPS calculates the difference between the orientation and position of equivalent clusters (identified by RIBFIND) in the initial or final conformation and the target conformation. This gives two values for each rigid body, *i.e.*, the shift and rotation angle needed to superpose the rigid body in the given initial or final conformation onto the corresponding rigid body in the target conformation. In order to better visualise the deviation, we combine the shift and rotation angle into one score. A sector is defined based on these two parameters, and the new score (ACPS) is the area of the sector: $ACPS = (\pi/360°) \times (shift)^2 \times angle$. If the method identifies more than one cluster, then the ACPS scores for each cluster are summed and reported. The units of the ACPS are in $Å^2$.

2.4.4. Cross-correlation coefficient (CCC)

A local CCC between the rigid bodies identified by RIBFIND in a given conformation (*i.e.*, the initial or the final conformation) and the target density map was calculated with Chimera (Pettersen et al., 2004). First a density map is simulated for each identified rigid body at the appropriate resolution, using Chimera's *molmap* command (with an *edgepadding* parameter of 8 Å). The local CCC is then calculated between the simulated map and the target map using the *measure correlation* command in Chimera. Additionally, CCCs for the fit of the entire protein was calculated.

Table 1
Summary of the simulated and experimental benchmarks.

Initial PDB ID	Target PDB ID	Range	Cluster cutoff (%)	No. of clusters	No. of SSEs in clusters	Total No. of SSEs	Percentage of SSEs in clusters (%)
Benchmark 1 (simulated)							
1wdnA	1gggA	5–224	25	2	11	13	85
1f6mE	1tdeA	1–316	30	2	9	12	75
1oaoD	1oaoC	2–729	33	6	22	28	79
2driA	1urpA	1–271	37	3	9	11	82
1k89A	1hrdC	1–449	40	3	13	21	62
1mrpA	1d9vA	1–309	35	3	13	14	93
1dpeA	1dppG	1–507	15	3	22	24	92
1ompA	1anfA	1–370	30	4	16	21	76
act2A[a]	act1A[a]	1–375	20	2	18	20	90
1lfgA	1lfhA	1–691	44	6	19	29	66
Benchmark 2 (experimental)							
act2A	3mfpA	1–375	20	2	18	20	90
3cb4F	3degC	1–13, 18–30, 57–545	35	4	15	18	83
1oelA	2c7eA	2–525	30	3	22	25	88

Descriptions for the items are: initial and target PDB ID, the PDB and chain ID for the initial and target conformations, respectively; range, the start and end residues; cluster cutoff, one of the RIBFIND parameters; No. of clusters, the number of clusters identified by RIBFIND; No. of SSEs in clusters, the total number of SSE included in clusters; total No. of SSEs, the total number of SSEs in the whole protein.

[a] act2A and act1A represents the models of initial and target conformations of actin taken from the Situs tutorial (Wriggers, 2010).

Please cite this article in press as: Pandurangan, A.P., Topf, M. Finding rigid bodies in protein structures: Application to flexible fitting into cryoEM maps. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.10.011

3. Results and discussion

3.1. Selection of parameters

The value of the contact distance parameter was fixed to 6.5 Å for all the test cases (Miyazawa and Jernigan, 1996). As the cluster cutoff increase from 0% to 100%, RIBFIND tends to find clusters of rigid bodies that were more compact. To determine what values to use for the cluster cutoff, we ran RIBFIND with all possible values from 0% to 100% (in increments of 1, see Fig. 2). For each such value, the method calculates a specific number of clusters and its members in the protein structure. Because two different cluster cutoffs can result in the same number of clusters but different numbers of members in them, we calculated a "unique" number of clusters by summing the number of clusters and the fraction of total SSE members included in each. We then used the cluster cutoff value that corresponds to the maximal unique number of clusters (see Table 1 and Fig. 2). The rationale behind this was to have as many rigid bodies as possible in the protein to allow more flexibility during fitting. Although the clustering method may result in some SSEs not being part of any cluster (see Table 1), in 11 out of 13 test cases 75% of the total SSEs were captured by the clustering (the exceptions are 1k89 and 1lfg). For these cluster cutoff values, the composition of the clusters identified by the method had no direct correlation with the domains identified by the CATH database (Orengo et al., 1997).

3.2. Accuracy of refinement

The results of the flexible fitting of the 10 proteins in the simulated benchmark are summarised in Table 2. The average percentage improvement in CCC (calculated from the initial and final averages of CCC) for the clustered set is 9.761%, 5.126%, and 4.720% for 5, 10, and 15 Å resolution maps, respectively, and for the non-clustered set is 9.569%, 5.345%, and 4.615%, respectively. The improvement in CCC is comparable between the clustered and non-clustered set, and in both cases it drops when using lower resolution maps. The average improvement in Cα RMSD for the clustered set is 76.70%, 51.30%, and 40.64% for 5, 10, and 15 Å resolution, respectively, and for the non-clustered set is 65.97%, 43.77%, and 26.52%, respectively. Irrespective of the use of clustered and non-clustered sets, the average improvement in Cα RMSD decreases with the resolution. Although at 15 Å resolution the average improvement in Cα RMSD using the clustered set is double that of the non-clustered set, the best Cα RMSD obtained at this resolution is 2.95 Å. The Cα RMSDs of the best models obtained at 5 and 10 Å resolution are 0.82 and 1.71 Å, respectively (Table 2). Thus, for resolutions worse than 10 Å, obtaining an accurate refined model remains a challenge. In most cases, the Cα RMSD between the final conformation and the target conformation is lower for the clustered set than the non-clustered set. Only 6 out of the 30 test cases (act2A and 1dpeA at 5 Å; 1f6mE and 1dpeA at 10 Å; and 1f6mE and 1ompA at 15 Å) have higher Cα RMSD when using the clustered set. In three of these test cases (act2A and 1dpeA at 5 Å, and 1ompA at 15 Å), the difference in Cα RMSD for the clustered and non-clustered set is almost comparable (see Table 2a and c). In many cases, it is clear that the ACPS score serves as a better indicator than the Cα RMSD for quantifying the subtle differences between final conformations based on the clustered and non-clustered. For example, the difference between the final Cα RMSD of 1mrpA using the clustered and non-clustered sets at 15 Å resolution (4.24 vs. 5.08 Å, respectively) (see Table 2c), may appear less significant than the corresponding difference in the ACPS scores (8.128 vs. 16.566 $Å^2$, respectively, see Table S2 in the Supporting material for the translational and rotational values for the respective rigid bodies).

3.3. Specific examples

Three specific examples are selected from the simulated benchmark for further analysis. These are 1oaoD, 2driA, and 1dpeA, each refined at 5, 10, and 15 Å resolution, respectively. Below, the results of the flexible fitting using the clustered and non-clustered sets of rigid bodies in each case are compared and discussed.

3.3.1. 1oaoD

1oaoD is the largest protein in the benchmark, with 729 residues. It has a total of 28 SSEs composing 22 helices and 6 beta sheets. Initial RIBFIND clustering with a cluster cutoff of 33% resulted in a maximal number of six clusters, containing a total of 22 SSEs (see Table 1). The CCC for the initial conformation is 0.788. For the two final conformations corresponding to the clustered and non-clustered sets refined at 5 Å resolution the CCCs are 0.896 and 0.843, respectively (see Table 2a). The Cα RMSDs for the initial conformation and final conformations (clustered and non-clustered sets) are 14.22, 5.37, and 9.44 Å, respectively (see Fig. 3a). Clearly, flexible fitting using rigid bodies created from the clustered set is far better than the non-clustered set in this case (see Table 2a and Fig. 3a). The improvement in Cα RMSD for the

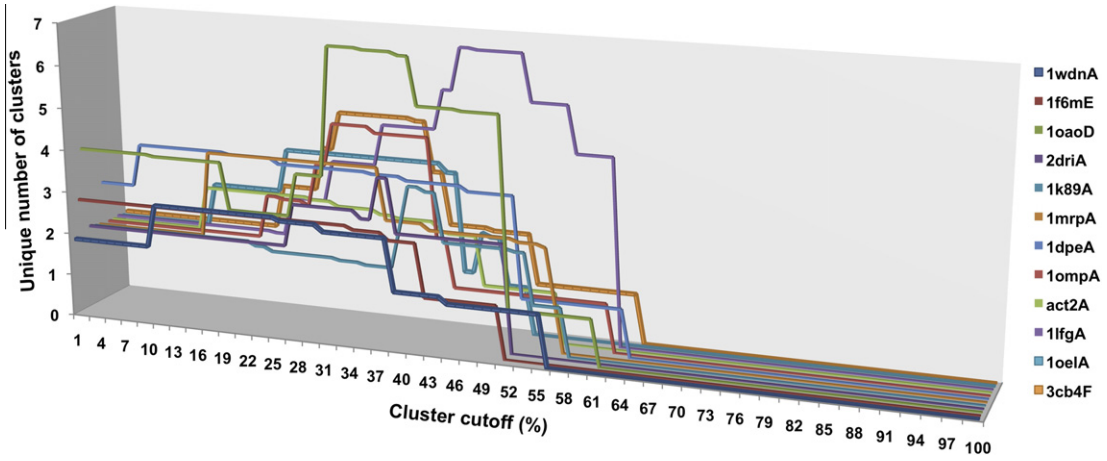


Fig.2. Unique number of clusters calculated by RIBFIND with all possible cluster cutoff values for all benchmark proteins. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: Pandurangan, A.P., Topf, M. Finding rigid bodies in protein structures: Application to flexible fitting into cryoEM maps. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.10.011

Table 2
Flexible fitting into simulated density maps at 5 Å (a), 10 Å (b), and 15 Å (c) resolution using clustered (RIBFIND) and non-clustered (SSE elements) sets of rigid bodies.

	CCC			Cα RMSD (Å)					ACPS (Å^2)		
	Initial	Final		Initial	Min		Final		Initial	Final	
		C	NC		C	NC	C	NC		C	NC
(a) Initial and target PDB IDs											
1wdnA–1gggA	0.797	0.932	0.937	10.60	0.30	0.33	1.25	1.79	76.91	0.06	0.13
1f6mE–1tdeA	0.850	0.927	0.927	9.89	0.21	0.22	1.45	2.04	90.87	0.05	0.44
1oaoD–loaoC	0.788	0.896	0.843	14.22	0.56	0.22	5.37	9.44	168.52	0.61	74.76
2driA–1urpA	0.832	0.942	0.945	7.55	0.21	0.18	0.82	0.97	26.62	0.04	0.08
1k89A–1hrdC	0.872	0.939	0.933	3.78	0.33	0.25	1.01	1.46	3.64	0.03	0.17
1mrpA–1d9vA	0.877	0.932	0.934	3.73	0.10	0.15	1.76	2.07	6.56	0.05	0.14
1dpeA–1dppG	0.814	0.867	0.891	12.28	0.61	0.20	3.92	3.69	126.73	4.41	1.97
1ompA–1anfA	0.842	0.938	0.939	7.24	0.21	0.26	1.05	1.33	50.32	0.04	0.16
act2A–act1A	0.828	0.898	0.925	5.87	0.31	0.43	2.54	2.12	2.45	1.09	0.52
1lfgA–1lfhA	0.861	0.905	0.886	8.13	0.32	0.23	2.66	5.07	153.37	0.76	41.89
Average	0.836	0.918	0.916	8.33	0.32	0.25	2.18	3.00			
SD	0.030	0.025	0.032	3.46	0.16	0.08	1.48	2.57			
Average % improvement[a]		9.761	9.569				76.70	65.97			
(b) Initial and target PDB IDs											
1wdnA–1gggA	0.868	0.919	0.924	10.60	0.30	0.33	5.32	6.39	76.91	4.57	10.10
1f6mE–1tdeA	0.920	0.948	0.962	9.89	0.21	0.22	5.88	4.66	90.87	22.16	7.67
1oaoD–loaoC	0.864	0.947	0.933	14.22	0.56	0.22	8.64	10.66	168.52	64.88	143.51
2driA–1urpA	0.914	0.977	0.974	7.55	0.21	0.18	1.71	3.04	26.62	0.37	2.77
1k89A–1hrdC	0.947	0.975	0.973	3.78	0.33	0.25	2.17	2.74	3.64	0.40	1.09
1mrpA–1d9vA	0.960	0.973	0.972	3.73	0.10	0.15	3.14	4.09	6.56	2.52	7.01
1dpeA–1dppG	0.878	0.941	0.956	12.28	0.61	0.20	5.91	4.89	126.73	15.56	7.30
1ompA–1anfA	0.918	0.976	0.980	7.24	0.21	0.26	2.32	2.65	50.32	0.58	1.07
act2A–act1A	0.919	0.964	0.963	5.87	0.31	0.43	2.76	4.41	2.45	1.61	13.47
1lfgA–1lfhA	0.923	0.957	0.961	8.13	0.32	0.23	4.34	4.40	153.37	6.66	8.03
Average	0.911	0.958	0.960	8.33	0.32	0.25	4.22	4.79			
SD	0.032	0.019	0.018	3.46	0.16	0.08	2.21	2.35			
Average % improvement[a]		5.126	5.345				51.30	43.77			
(c) Initial and target PDB IDs											
1wdnA–1gggA	0.886	0.975	0.975	10.60	0.30	0.33	5.07	5.27	76.91	11.69	8.06
1f6mE–1tdeA	0.934	0.970	0.969	9.89	0.21	0.22	7.05	5.58	90.87	34.02	17.31
1oaoD–loaoC	0.883	0.960	0.948	14.22	0.56	0.22	10.14	13.00	168.52	103.90	250.36
2driA–1urpA	0.934	0.977	0.974	7.55	0.21	0.18	4.36	5.37	26.62	7.09	14.32
1k89A–1hrdC	0.965	0.976	0.966	3.78	0.33	0.25	3.69	5.40	3.64	2.70	10.39
1mrpA–1d9vA	0.977	0.981	0.975	3.73	0.10	0.15	4.24	5.08	6.56	8.13	16.57
1dpeA–1dppG	0.887	0.964	0.969	12.28	0.61	0.20	4.17	7.00	126.73	5.32	25.76
1ompA–1anfA	0.933	0.980	0.984	7.24	0.21	0.26	3.93	3.79	50.32	5.85	4.78
act2A–act1A	0.938	0.974	0.973	5.87	0.31	0.43	2.95	5.74	2.45	1.60	20.26
1lfgA–1lfhA	0.938	0.961	0.970	8.13	0.32	0.23	5.13	5.64	153.37	13.17	13.84
Average	0.928	0.972	0.970	8.33	0.32	0.25	5.07	6.19			
SD	0.033	0.008	0.009	3.46	0.16	0.08	2.09	2.52			
Average% improvement[a]		4.720	4.615				40.64	26.52			

Descriptions for the items are: initial and final PDB IDs, the PDB and chain ID for the initial and final conformations; initial and final, initial and final conformation; C and NC, final conformations based on the clustered and non-clustered sets of rigid bodies, respectively; CCC, the value of the cross-correlation coefficient of a given conformation with the density map calculated with Chimera; Cα RMSD, Cα root-mean square deviation between the a given conformation and the target conformation; ACPS, the area based component placement score between a given conformation and the target conformation; min, the minimal RMSD (see minRMSD in Section 2); SD, the standard deviation.

[a] Average % improvement in CCC for clustered and non-clustered sets is calculated by (average_initial − average_final)/average_initial × 100; Average % improvement in RMSD for clustered and non-clustered set is calculated by (average_initial − average_final)/(average_initial − average_minRMSD)/100.

clustered set (65%) is twice higher than the non-clustered set (34%).

Although the fit based on the clustered set is better than the one based on the non-clustered set for all three resolutions, it is still far from the target conformation at 10 and 15 Å resolutions. The final Cα RMSDs for the clustered set at 10 and 15 Å resolution maps are 8.64 and 10.14 Å, respectively, compared with 10.66 and 13.00 Å for the non-clustered set (see Table 2b and c). Interestingly, flexible fitting with a new clustered set obtained using a cluster cutoff of 10% resulted in final conformations whose Cα RMSDs with the target conformation are 1.74, 2.93, and 6.99 Å at 5, 10, and 15 Å resolution, respectively. At this value of cluster cutoff only three clusters were identified by the method (vs. six with the cluster cutoff of 33%, Table 1), containing a total of 25 SSEs. Out of these three clusters, one cluster contains two helices (residues 627–634 and 671–684) and a sheet (residues 599–604, 609–614, 643–646, 666–668), which were not clustered with the cluster cutoff of 33% (therefore represented as individual rigid bodies). The deviation of these three rigid bodies from the target conformation is quite large (with ACPS = 208.21 Å^2: shift = 21.3 Å, angle = 52.6°), and therefore, treating them as one cluster resulted in a better fit.

3.3.2. 2driA

2driA has a total of 11 SSEs composing 9 helices and 2 beta sheets. With the cluster cutoff of 37% the method identified the maximal number of three clusters containing a total of 9 SSEs (see Table 1). The CCCs for initial conformation and the final conformations corresponding to the clustered and non-clustered sets refined at 10 Å resolution are 0.914, 0.977, and 0.974, respectively (see Table 2b). The Cα RMSDs for the initial conformation

Please cite this article in press as: Pandurangan, A.P., Topf, M. Finding rigid bodies in protein structures: Application to flexible fitting into cryoEM maps. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.10.011

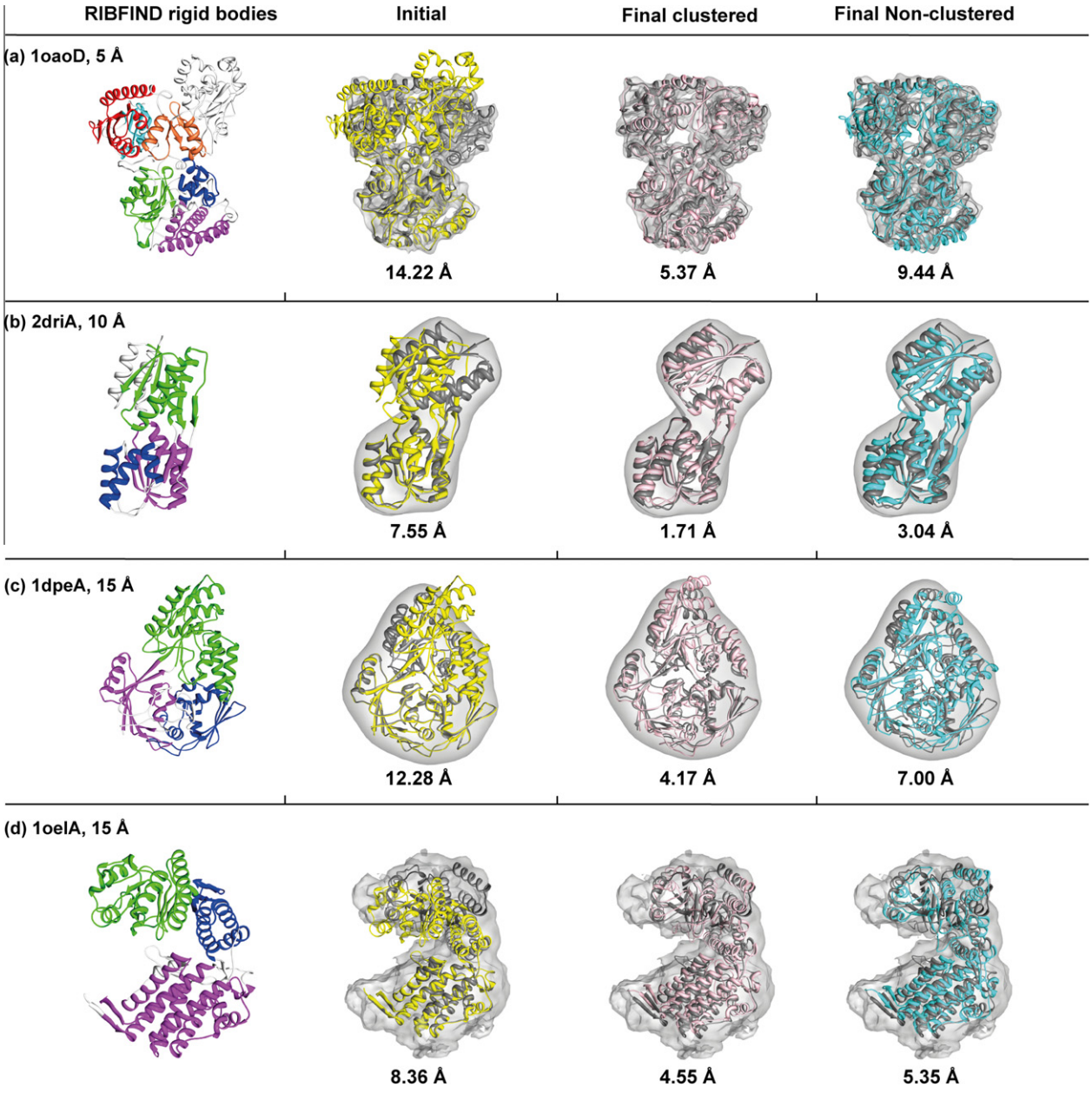


Fig.3. Flexible fitting with and without RIBFIND of 1oaoD (a), 2driA (b), 1dpeA (c) into simulated cryoEM maps at 5, 10, and 15 Å, resolution, respectively, and of 1oelA (d) into the segmented experimental cryoEM map (EMD-1047) at 15 Å resolution. On the left we show the rigid bodies identified by RIBFIND (uniquely coloured) on the initial conformation (cartoon representation). For 1oaoD (a), the six rigid bodies are coloured magenta, blue, green, cyan, red, and coral in order. For 2driA, 1dpeA, and 1oelA, the first, second, and third rigid bodies are coloured magenta, blue, and green, respectively. On the right we show the initial conformation (yellow) and the final conformations based on the clustered (pink) and non-clustered (cyan) sets of rigid bodies superposed onto the target conformation (grey) and fitted in the density map (grey). The Cα RMSDs of the initial and the two final conformations (corresponding to the clustered and non-clustered sets) from the target conformation are shown below the respective examples. The figures were generated with Chimera (Pettersen et al., 2004).

and final conformations (clustered and non-clustered set) are 7.55, 1.71, and 3.04 Å, respectively. At 15 Å resolution, the Cα RMSDs for the two final conformations are 4.36 and 5.37 Å, respectively. The higher Cα RMSDs for the final conformation based on the non-clustered set are caused by the rigid bodies being drawn towards the centre of the density map during refinement (see Fig. 3b).

This result is more profound for the rigid bodies that correspond to the third cluster in the clustered set (residues 3–38, 58–63, 85–103, 237–263), for which the final fit of part of it was located at a low-density region of the map. This is revealed by the cluster-based CCC and ACPS scores (see Tables S5 and S2): the cluster-based CCC scores for the first, second, and third rigid bodies are 0.917, 0.860, and 0.934 at 10 Å resolution, and 0.891, 0.819, and 0.900 at 15 Å resolution for the final fits based on the clustered set, and 0.920, 0.866, and 0.923 at 10 Å and 0.882, 0.833, and 0.896 at 15 Å for the final fits based on the non-clustered set; The corresponding ACPS scores are 0.113, 0.170, and 0.086 $Å^2$ at 10 Å resolution, and 0.878, 1.951, and 4.259 $Å^2$ at 15 Å resolution for the clustered set, and 0.162, 0.354, and 2.255 $Å^2$ at 10 Å resolution, and 0.617, 1.093, and 12.605 $Å^2$ at 15 Å resolution for the non-clustered set. The improvement in Cα RMSD for the clustered and non-clustered set is 80% and 60%, respectively.

Please cite this article in press as: Pandurangan, A.P., Topf, M. Finding rigid bodies in protein structures: Application to flexible fitting into cryoEM maps. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.10.011

3.3.3. 1dpeA

1dpeA represents a challenging case of a large protein with numerous loops. It is the third largest protein in the simulated benchmark with 507 residues. It has a total of 24 SSEs composed of 19 helices and 5 beta sheets, and 29 loops with varying lengths ranging from 1 to 22 residues. Using a cluster cutoff of 15% RIBFIND identified three clusters containing 22 SSEs (see Table 1). At 5 and 10 Å resolution, the Cα RMSDs of the final conformations based on the non-clustered set are 3.69 and 4.89 Å and based on the clustered set are 3.92 and 5.91 Å, respectively (see Table 2a and b). At 15 Å resolution, however, using the clustered set resulted in a significantly better fit than using the non-clustered set. At this resolution, the CCCs for the initial conformation and the two final conformations (clustered and non-clustered sets) are 0.887, 0.964, and 0.969, respectively (see Table 2c). The Cα RMSDs for the initial conformation and the final conformations (clustered and non-clustered sets) are 12.28, 4.17, and 7.00 Å, respectively (see Fig. 3c). The ACPS scores between the initial and target conformations calculated for the rigid bodies corresponding to the first and second clusters (0.004 and 0.004 Å^2) show that the clusters are already close to the target conformation (see Table S1). In the final conformation using the non-clustered set rigid bodies, the corresponding ACPS scores (first and second clusters) are 4.582 and 6.391 Å^2, respectively (see Table S2), which is worse than the initial conformation ACPSs and worse than the corresponding scores for the final conformation using the clustered set rigid bodies (0.353 and 0.425 Å^2). Hence, in low-resolutions, refinement with the clustered set restricts over fitting.

The over fitting of the second cluster in particular is well pronounced in the cluster-based CCC, which is higher for the non-clustered set (0.899 vs. 0.880 for the clustered set, see Table S5). For the rigid bodies corresponding to the third cluster, there is an improvement in fit using both the clustered and non-clustered sets, but the improvement is much more profound for the clustered set (ACPS scores are 126.721, 4.539, and 14.788, for the initial and the final conformations based on the clustered and non-clustered sets, respectively, see Tables S1 and S2).

3.4. Experimental results

The results of the second benchmark, containing proteins which were refined using cryoEM maps of an actin monomer, the elongation factor EF4, and one subunit of GroEL, at 6.6, 11, and 15 Å, respectively, are discussed below.

3.4.1. Actin (act2A, 6.6 Å)

The actin monomer has a total of 20 SSEs containing 16 helices and 4 beta sheets. The initial and target conformations represent the open and closed forms of actin, respectively. With a cluster cutoff of 20%, RIBFIND obtained two clusters containing a total of 18 SSEs (see Table 1). The CCCs for the initial conformation and the final conformations corresponding to the clustered and non-clustered sets are 0.482, 0.619, and 0.685, respectively (see Table 3). The corresponding Cα RMSDs are 6.09, 3.37, and 3.15 Å. Thus, in this case the non-clustered based refinement produced a slightly better fit. Interestingly, although a similar result was observed at a comparable 5 Å simulated map in the first benchmark (see Table 2a), at resolutions of 10 and 15 Å, the same clustered set of actin proved to be better than the non-clustered set. For the simulated actin map at 10 Å, the Cα RMSDs for the initial conformation and two final conformations (clustered and non-clustered sets) are 5.87, 2.76, and 4.41 Å respectively (see Table 2b), and for the corresponding 15 Å map are 5.87, 2.95, and 5.74 Å, respectively (see Table 2c).

It is worth noting that in the case of actin, reversing the test by refining the structure from the target conformation (closed conformation) into the initial conformation (open conformation) was not successful using RIBFIND cluster-based rigid bodies (data not shown). In this case, the SSEs in the closed conformation are tightly packed and therefore the method was not able to identify rigid bodies that allowed flexibility between the closed and open conformation.

3.4.2. EF4 (3cb4F, 11 Å)

EF4 has 18 SSEs composed of 11 helices and 7 beta sheets. Using a cluster cutoff of 35% RIBFIND identified 4 clusters containing a total of 15 SSEs (see Table 1). The Cα RMSDs of the initial conformation and the two final conformations (clustered and non-clustered sets) are 6.16, 3.95, and 5.43 Å, respectively (see Table 3). The CCCs for the clustered and non-clustered sets are comparable (0.763 vs. 0.782, respectively) while the percentage improvement in Cα RMSD (relative to the initial conformation) of the former is twice the latter (36.35% vs. 12.27%, respectively). Interestingly, the ACPS scores of the 4th cluster in the initial conformation and final conformations resulting from the clustered and non-clustered sets are 2.262, 1.374, and 31.000 Å^2, respectively (see Tables S3 and S4), indicating a poor fit using non-clustered rigid bodies during the refinement.

3.4.3. GroEL (1oelA,15 Å)

The subunit of GroEL has 25 SSEs composing 18 helices and 7 beta sheets. With a cluster cutoff of 30% the subunit is divided into three clusters (corresponding to the equatorial, intermediate and apical domains) containing 22 SSEs (see Table 1 and Fig. 3d). The Cα RMSDs of the initial conformation and the two final conformations (clustered and non-clustered sets) are 8.36, 4.55, and 5.35 Å, respectively (see Table 3). The CCC scores for the clustered and non-clustered sets are comparable (0.926 and 0.928, respectively).

For the clustered set, the ACPS scores of the first, second, and third cluster are 0.378, 1.618, and 2.748 Å^2 and for the corresponding non-clustered set are 0.651, 3.326, and 6.977 Å^2 (see Table S4), respectively. For the initial conformation the ACPS

Table 3
Flexible fitting of into experimental cryoEM maps using clustered (RIBFIND) and non-clustered (SSEs) sets of rigid bodies.

Initial and target PDB IDs	EMDB ID	CCC			Cα RMSD (Å)					ACPS (Å^2)		
		Initial	Final		Initial	Min		Final		Initial	Final	
			C	NC		C	NC	C	NC		C	NC
act2A–3mfpA	EMD-5168 (6.6 Å)	0.482	0.619	0.685	6.09	1.67	1.00	3.37	3.15	116.09	15.90	15.38
3cb4F–3degC	EMD-1524 (11 Å)	0.672	0.763	0.782	6.16	0.08	0.21	3.95	5.43	23.85	4.01	38.75
1oelA–2c7eA	EMD-1047 (15 Å)	0.897	0.926	0.928	8.36	0.38	0.48	4.55	5.35	51.75	4.74	10.95

Descriptions for the items are: initial and final PDB IDs, the PDB and chain ID for the initial and final conformations; EMDB ID, the EMDB ID of the target map (the resolution of the target map in parentheses); initial and final, initial and final conformation; C and NC, final conformations based on the clustered and non-clustered sets of rigid bodies, respectively; CCC, the value of the cross-correlation coefficient of a given conformation with the density map calculated with Chimera; Cα RMSD, Cα root-mean square deviation between a given conformation and the target conformation; ACPS, the area based component placement score between the a given conformation and the target conformation; min, the minimal RMSD (see minRMSD in Section 2).

Please cite this article in press as: Pandurangan, A.P., Topf, M. Finding rigid bodies in protein structures: Application to flexible fitting into cryoEM maps. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.10.011

scores of the first, second, and third cluster are 0.009, 6.507, and 45.232 Å^2, respectively (see Table S3). Thus, the ACPS score indicates a better fit for the second and third clusters (corresponding to the intermediate and apical domains) in the final conformation based on the clustered set than for the corresponding fit based on the non-clustered set. Over fitting is observed in the first cluster in both cases, but it is more pronounced in the non-clustered set.

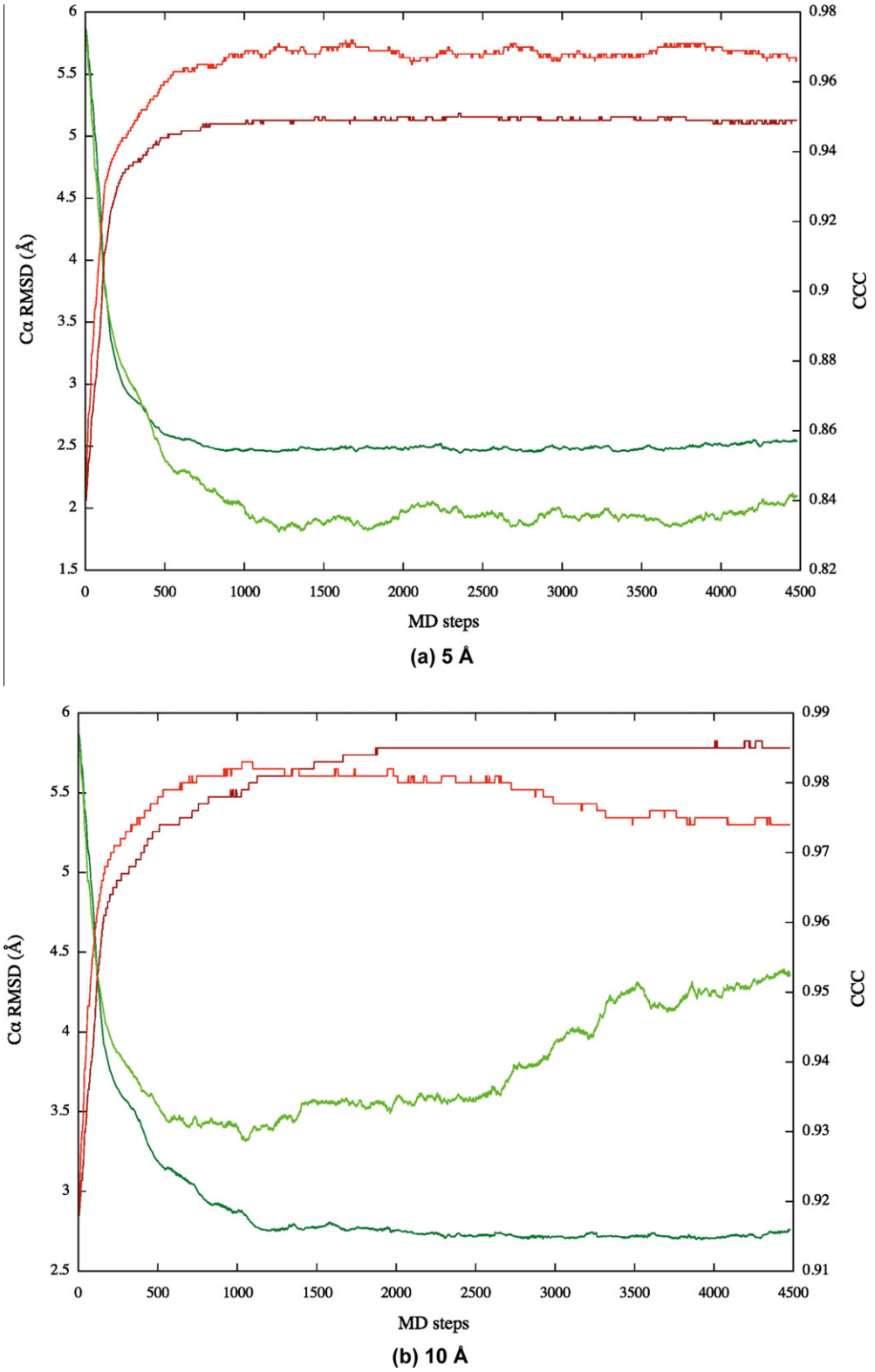
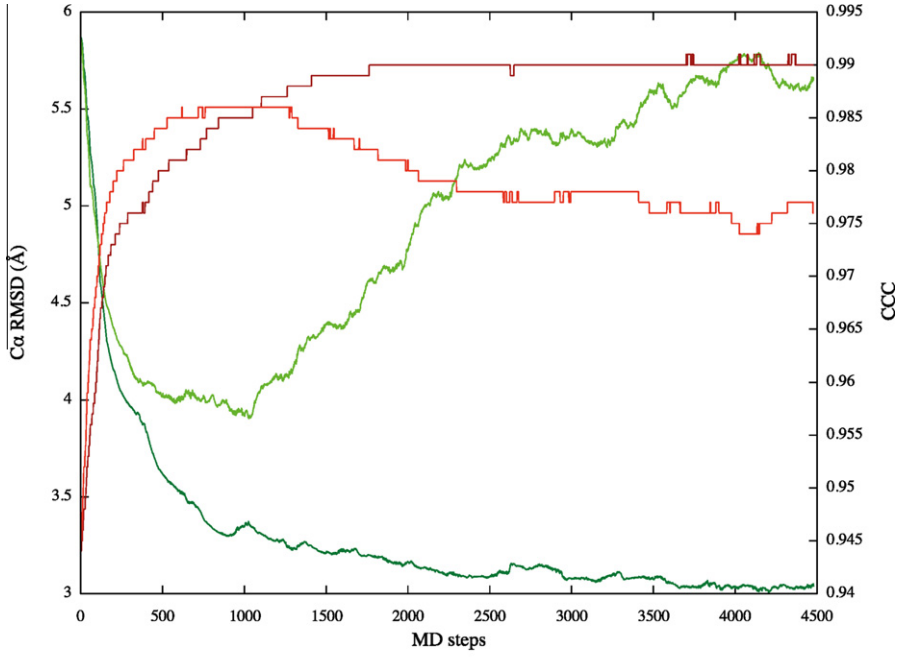


Fig.4. The Cα RMSD and the Flex-EM CCC profiles for the clustered and non-clustered sets calculated along the Flex-EM refinement of actin (act2A) in the (a) 5 Å, (b) 10 Å, and (c) 15 Å simulated cryoEM maps. The Cα RMSD profiles for the clustered and non-clustered are shown in dark red and red. The corresponding CCC profiles are shown in dark green and green, respectively. (Note that the Cα RMSD of the final conformation, which is usually obtained after a short conjugate gradient minimisation at the end of the MD simulation, is not included in the figure.)

Please cite this article in press as: Pandurangan, A.P., Topf, M. Finding rigid bodies in protein structures: Application to flexible fitting into cryoEM maps. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.10.011



(c) 15 Å

Fig. 4 (*continued*)

3.5. Optimisation

The course of flexible fitting during the Flex-EM refinement is very much dependent on the choice of rigid bodies. Fig. 4 shows the Cα RMSD and CCC profiles of the actin (act2A) refinement in the simulated maps at all three resolutions (5, 10, and 15 Å). Starting from a Cα RMSD of ~6.0 Å between initial conformation and the target conformation, the simulations using rigid bodies corresponding to the clustered and non-clustered sets show similar convergence until around 400, 130, and 100 simulation steps in the 5, 10, and 15 Å resolutions maps, respectively, where the Cα RMSD approaches 2.70, 4.30, and 4.90 Å, respectively. Unlike the case corresponding to the clustered set, at 10 and 15 Å resolution the Cα RMSD corresponding to the non-clustered set diverges early and increases with time, indicating a poor fit that moves away from target conformation. Therefore, the lower the resolution, the more essential is the identification of proper rigid bodies in order to drive the rest of the simulation towards the target conformation. Even at 5 Å resolution, although the non-clustered set shows slightly better convergence, the Cα RMSD and the CCC profiles of the clustered set converge steadily (without fluctuations), *i.e.*, less likely to allow over fitting. The Cα RMSD and the CCC profiles of non-clustered set show even greater fluctuations with the 10 and 15 Å resolution simulated maps. The steady convergence of the CCC for the clustered set is an added advantage if the decision on the length of the simulation is unclear.

3.6. Two-stage refinement protocol

Flexible fitting of the atomic structure into the density map can be further improved if the refinement is done in independent stages starting by defining larger rigid bodies (*i.e.*, clustered set) and moving toward a selection of smaller rigid bodies (*i.e.*, non-clustered set) in a hierarchical fashion. To demonstrate this, we applied a second round of refinement to all test cases (in the simulated benchmark) for which the non-clustered results performed better than the clustered ones at sub-nanometer resolution (1dpe and act2 at 5 Å and 1dpe and 1f6m at 10 Å, see Table 2a and b). In each of these cases, starting from the final conformation obtained using the clustered set of rigid bodies, a second stage of refinement with five simulated annealing MD cycles was performed by treating individual SSEs as rigid bodies (non-clustered set). This procedure not only improved the results obtained in the first round of refinement (using the clustered set) but also resulted in conformations better than those of initial non-clustered based refinement. After applying the second stage of refinement, the Cα RMSDs of 1dpe and act2 at 5 Å resolution improved from 3.92 to 2.12 Å (see Fig. S1 in the Supporting material) and from 2.54 to 1.93 Å, respectively. The corresponding CCC scores improved from 0.867 to 0.919 and from 0.898 to 0.921, respectively. The Cα RMSDs of 1dpe and 1f6m at 10 Å resolution improved from 5.91 to 3.09 Å and from 5.88 to 4.07 Å, respectively. The corresponding CCC scores improved from 0.941 to 0.972 and from 0.948 to 0.963, respectively. The final Cα RMSDs of the original non-clustered based refinement for 1dpe and act2 at 5 Å are 3.69 and 2.12 Å, respectively, and for 1dpe and 1f6m at 10 Å are 4.89 and 4.66 Å, respectively (Table 2a and b).

4. Conclusion

Our broad objective is to be able to characterise the structure of macromolecular assemblies as accurately as possible at different functional states. Flexible fitting of atomic structures into cryoEM maps has become an important tool in achieving this goal. However, optimising the fit of an atomic structure in a low-resolution density map is a multiple minima problem, and adding flexibility to the atomic structure (in comparison to rigid fitting) further increases the complexity of the problem. Dividing the atomic structure into a series of smaller rigid bodies may help the optimisation by limiting the conformational degrees of freedom relative to full flexibility (where each atom is essentially a rigid body), and preventing it from getting trapped in many of the local minima. Inappropriate identification of rigid bodies, however, can result in too limited flexibility, preventing the optimisation from exploring the

Please cite this article in press as: Pandurangan, A.P., Topf, M. Finding rigid bodies in protein structures: Application to flexible fitting into cryoEM maps. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.10.011

directions leading to the global minimum. We presented here a method for finding appropriate rigid bodies solely based on structural information in protein structures (RIBFIND). We compared the flexible fits obtained using rigid bodies identified by the method (clustered set) with fits obtained using each SSE as an individual rigid body (non-clustered set). Our results show that the quality of the fit using the clustered set was better for most of the proteins in the benchmark, and in those cases for which the fits were not better (at the sub-nanometer resolution) they were further improved by applying a second round of refinement using the non-clustered set (two-stage refinement protocol). Thus, at resolutions between 5 and 15 Å we recommend to apply cluster-based refinement first.

The method (which is applicable to any size of an atomic structure) outputs a set of rigid bodies that can be used when refining the structure in a cryoEM map (typically at intermediate resolutions). The method is stand-alone and fully automated while also allowing user intervention to select different sets of rigid bodies based on two parameters – the contact distance and cluster cutoff. Despite the possibility of changing the values of both parameters we kept a constant value for the former throughout the benchmark while the value of the latter was selected based on the maximum unique number of clusters. However, depending on the complexity of the problem, using different values (especially for the cluster cutoff) will allow the identification of different sets of rigid bodies. Using the corresponding different sets of rigid bodies, one, in principle, could design a hierarchical optimisation for the refinement protocol that is beyond the scope of this paper. Other methods for identifying rigid bodies in proteins could also be used for that purpose. However, our two-stage refinement protocol is a first example showing the potential of such approach.

RIBFIND was tested on three different resolutions at the intermediate range (5, 10, and 15 Å). It generally performed best on the low-resolution maps (15 Å). At these resolutions, there is a tendency for over fitting where the rigid bodies are "pulled" into the centre of the map. This is because at this resolution many fits have similar scoring and therefore reaching the optimal fit is much more challenging. Based on our benchmark, clustering the SSEs into larger rigid bodies clearly minimises this effect. In addition, it allowes the optimisation to converge better and stay steadier, without over fitting. This effect of a more stable optimisation was evident also at the higher resolutions but to a lesser extent.

As expected, at the highest resolution (5 Å), where the density information is greater, the differences between results based on clustered and non-clustered rigid bodies were smaller on average. Furthermore, one problem with our method, which is based on the proximity of the SSEs is that in principle if the protein is very compact it sometimes fails to capture appropriate rigid bodies, especially if the protein under consideration moves from a tightly packed conformation to a open conformation (e.g., actin – from close to open). In those cases using cluster-based rigid bodies may not allow any refinement to take place. Manual intervention and the use of additional experimental information may help to resolve the problem.

Finally, when fitting into cryoEM maps it is often the case that the structure of the individual components is not known. In those cases one could use protein structure prediction methods (homology or *ab initio* modelling). However, these methods carry modelling errors, which could complicate the optimisation in the cryoEM map. Clustering individual SSEs into groups of larger rigid bodies may help minimising this problem. Similarly, having larger rigid bodies that work as an additional constraint on the optimisation may help minimising the effect of noise.

In summary, although our results suggest that obtaining an accurate model for low-resolution maps (10–15 Å) remains a challenge irrespective of the rigid body definition, they also show that flexible fitting at these resolutions can often be significantly improved using a proper selection of rigid bodies. Our method, in principle, can be applied to other fitting programs or other techniques that rely on refinement of rigid bodies using experimental information, such as SAXS (Forster et al., 2008). Due to the simplicity of the method it can be easily integrated into molecular visualisation programs (such as Chimera, Pettersen et al., 2004), thereby helping the process of manual and automated fitting. Future directions will involve extending the method to identify optimal rigid bodies for refinement using lower resolutions density maps and design of optimal parameters for hierarchical clustering.

Acknowledgments

The authors are grateful to Dr. Daven Vasishtan for very helpful discussions and Drs. David Houldershaw and Richard Westlake for computer support. This research was supported by a grant from the Human Frontier Science Program (RGY0079/2009-C) and an MRC Career Development Award (G0600084).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jsb.2011.10.011.

References

Abyzov, A., Bjornson, R., Felipe, M., Gerstein, M., 2010. RigidFinder: a fast and sensitive method to detect rigid blocks in large macromolecular complexes. Proteins 78, 309–324.

Alber, F., Forster, F., Korkin, D., Topf, M., Sali, A., 2008. Integrating diverse data for structure determination of macromolecular assemblies. Annu. Rev. Biochem. 77, 443–477.

Baker, D., 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., Chiu, W., 2006. Ab initio modeling of the herpesvirus VP26 core domain assessed by CryoEM density. PLoS Comput. Biol. 2, e146.

Beck, M., Topf, M., Frazier, Z., Tjong, H., Xu, M., Zhang, S., Alber, F., 2011. Exploring the spatial and temporal organization of a cell's proteome. J. Struct. Biol. 173, 483–496.

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

Chandramouli, P., Topf, M., Menetret, J.F., Eswar, N., Cannone, J.J., Gutell, R.R., Sali, A., Akey, C.W., 2008. Structure of the mammalian 80S ribosome at 8.7 A resolution. Structure 16, 535–548.

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

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

Connell, S.R., Topf, M., Qin, Y., Wilson, D.N., Mielke, T., Fucini, P., Nierhaus, K.H., Spahn, C.M., 2008. A new tRNA intermediate revealed on the ribosome during EF4-mediated back-translocation. Nat. Struct. Mol. Biol. 15, 910–915.

DiMaio, F., Tyka, M.D., Baker, M.L., Chiu, W., Baker, D., 2009. Refinement of protein structures into low-resolution density maps using rosetta. J. Mol. Biol. 392, 181–190.

Evans, R.N., Blaha, G., Bailey, S., Steitz, T.A., 2008. The structure of LepA, the ribosomal back translocase. Proc. Natl. Acad. Sci. USA 105, 4673–4678.

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

Flores, S., Echols, N., Milburn, D., Hespenheide, B., Keating, K., Lu, J., Wells, S., Yu, E.Z., Thorpe, M., Gerstein, M., 2006. The Database of Macromolecular Motions: new features added at the decade mark. Nucleic Acids Res. 34, D296–D301.

Flores, S.C., Gerstein, M.B., 2007. FlexOracle: predicting flexible hinges by identification of stable domains. BMC Bioinf. 8, 215.

Forster, F., Webb, B., Krukenberg, K.A., Tsuruta, H., Agard, D.A., Sali, A., 2008. Integration of small-angle X-ray scattering data into structural modeling of proteins and their assemblies. J. Mol. Biol. 382, 1089–1106.

Frank, J., 2006. Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State. Oxford University Press, New York.

Frank, J., 2009. Single-particle reconstruction of biological macromolecules in electron microscopy – 30 years. Q. Rev. Biophys. 42, 139–158.

Fujii, T., Iwane, A.H., Yanagida, T., Namba, K., 2010. Direct visualization of secondary structures of F-actin by electron cryomicroscopy. Nature 467, 724–728.

Please cite this article in press as: Pandurangan, A.P., Topf, M. Finding rigid bodies in protein structures: Application to flexible fitting into cryoEM maps. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.10.011

Grubisic, I., Shokhirev, M.N., Orzechowski, M., Miyashita, O., Tama, F., 2010. Biased coarse-grained molecular dynamics simulation approach for flexible fitting of X-ray structure into cryo electron microscopy maps. J. Struct. Biol. 169, 95–105.

Hayward, S., Berendsen, H.J., 1998. Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme. Proteins 30, 144–154.

Jacobs, D.J., Rader, A.J., Kuhn, L.A., Thorpe, M.F., 2001. Protein flexibility predictions using graph theory. Proteins 44, 150–165.

Jolley, C.C., Wells, S.A., Fromme, P., Thorpe, M.F., 2008. Fitting low-resolution cryo-EM maps of proteins using constrained geometric simulations. Biophys. J. 94, 1613–1621.

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

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

Lawson, C.L., Baker, M.L., Best, C., Bi, C., Dougherty, M., Feng, P., van Ginkel, G., Devkota, B., Lagerstedt, I., Ludtke, S.J., Newman, R.H., Oldfield, T.J., Rees, I., Sahni, G., Sala, R., Velankar, S., Warren, J., Westbrook, J.D., Henrick, K., Kleywegt, G.J., Berman, H.M., Chiu, W., 2011. EMDataBank.org: unified data resource for CryoEM. Nucleic Acids Res. 39, D456–D464.

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

Miyazawa, S., Jernigan, R.L., 1996. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J. Mol. Biol. 256, 623–644.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M., 1997. CATH – a hierarchic classification of protein domain structures. Structure 5, 1093–1108.

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

Ranson, N.A., Farr, G.W., Roseman, A.M., Gowen, B., Fenton, W.A., Horwich, A.L., Saibil, H.R., 2001. ATP-bound states of GroEL captured by cryo-electron microscopy. Cell 107, 869–879.

Rawi, R., Whitmore, L., Topf, M., 2010. CHOYCE – a web server for constrained homology modelling with cryoEM maps. Bioinformatics 26, 1673–1674.

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

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

Schroder, 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 15, 1630–1641.

Spahn, C.M., Penczek, P.A., 2009. Exploring conformational modes of macromolecular assemblies by multiparticle cryo-EM. Curr. Opin. Struct. Biol. 19, 623–631.

Suhre, K., Navaza, J., 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., Brooks 3rd, C.L., 2002. Exploring global distortions of biological macromolecules and assemblies from low-resolution structural information and elastic network theory. J. Mol. Biol. 321, 297–305.

Taylor, D.J., Devkota, B., Huang, A.D., Topf, M., Narayanan, E., Sali, A., Harvey, S.C., Frank, J., 2009. Comprehensive molecular structure of the eukaryotic ribosome. Structure 17, 1591–1604.

Topf, M., 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., 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., Sali, A., 2006. Refinement of protein structures by iterative comparative modeling and CryoEM density fitting. J. Mol. Biol. 357, 1655–1668.

Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W., Sali, A., 2008. Protein structure fitting and refinement guided by cryo-EM density. Structure 16, 295–307.

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

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

Volkmann, N., Hanein, D., Ouyang, G., Trybus, K.M., DeRosier, D.J., Lowey, S., 2000. Evidence for cleft closure in actomyosin upon ADP release. Nat. Struct. Biol. 7, 1147–1155.

Wendt, T., Taylor, D., Trybus, K.M., Taylor, K., 2001. Three-dimensional image reconstruction of dephosphorylated smooth muscle heavy meromyosin reveals asymmetry in the interaction between myosin heads and placement of subfragment 2. Proc. Natl. Acad. Sci. USA 98, 4361–4366.

Woetzel, N., Lindert, S., Stewart, P.L., Meiler, J., 2011. BCL::EM-Fit: rigid body fitting of atomic structures into density maps using geometric hashing and real space refinement. J. Struct. Biol. 175, 264–276.

Wriggers, W., 2010. Using Situs for the integration of multi-resolution structures. Biophys. Rev. 2, 21–27.

Wriggers, W., Schulten, K., 1997. Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates. Proteins 29, 1–14.

Wriggers, W., Chacon, P., 2001. Modeling tricks and fitting techniques for multiresolution structures. Structure (Cambridge) 9, 779–788.

Zhang, S., Vasishtan, D., Xu, M., Topf, M., Alber, F., 2010. A fast mathematical programming procedure for simultaneous fitting of assembly components into cryoEM density maps. Bioinformatics 26, i261–i268.

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

Please cite this article in press as: Pandurangan, A.P., Topf, M. Finding rigid bodies in protein structures: Application to flexible fitting into cryoEM maps. J. Struct. Biol. (2011), doi:10.1016/j.jsb.2011.10.011