SUPPLEMENTARY INFORMATION (SI)

Fully Automated Fold Recognition in Low Resolution Electron Density Maps

Mitul Saha and Marc C. Morais

Supplementary Figure 1 (Fig. S1)	Outline of the algorithm to solve **P**
Supplementary Figure 2 (Fig. S2)	Outline of algorithm for computing local Cartesian reference frames in Step 1 of Supplementary Figure 1
Supplementary Figure 3 (Fig. S3)	Outline of algorithm for computing local region descriptors (LRDs) in Step 2 of Supplementary Figure 1
Supplementary Figure 4 (Fig. S4)	Cartoon representations of the steps to solve **P**
Supplementary Figure 5 (Fig. S5)	Flexibility in ribosome 70S
Supplementary Figure 6 (Fig. S6)	Additional flexible fitting test-cases
Supplementary Figure 7 (Fig. S7)	The four domain test-case
Supplementary Table 1 (a)-(b) (Table S1)	RMSD error in docking/fitting, using FOLD-EM
Supplementary Table 2 (a)-(d) (Table S2)	RMSD errors in docking/fitting in the presence of extraneous regions.
Supplementary Table 3 (a)-(d) (Table S3)	Flexible fitting
Supplementary Table 4 (a)-(i) (Table S4)	Automated fold recognition
Supplementary Table 5 (Table S5)	Evaluation of the FOLD-EM generated fittings
Supplementary Text 1 (Text S1)	MOTIF-EM
Supplementary Text 2 (Text S2)	Validation of P22 results
Supplementary Text 3 (Text S3)	Fold recognition in ribosome 70S

The Algorithm to solve P:

Notations:

M_i: map i, i=1,2

p^i_j: grid point p_j in map i.

Λ^i_j: LRD at grid point p_j in map i.

$m(p^i_j, p^k_l)$: A match pair of grid points p^i_j (from map i) and p^k_l (from map k), i~=k

$O(p^i_j)$ or O^i_j: O-XYZ Cartesian reference frame at grid point p^i_j

If S is a set, $S(i)$ is the i-th element of S

X': transpose of X

The Algorithm to solve **P**
Inputs: Volumetric density maps M_1, M_2

1. Compute Cartesian frame sets for M_1 and M_2:
   $\boldsymbol{O}(M_1)$:$\{O^1_1, O^1_2, \ldots\}$=**compute_frame_set**(M_1)
   $\boldsymbol{O}(M_2)$:$\{O^2_1, O^2_2, \ldots\}$=**compute_frame_set**(M_2)

2. Compute LRD sets for M_1 and M_2:
   $\boldsymbol{\Lambda}(M_1)$=$\{\Lambda^1_1, \Lambda^1_2, \ldots\}$=**compute_LRD_set**($M_1$, $\boldsymbol{O}(M_1)$)
   $\boldsymbol{\Lambda}(M_2)$=$\{\Lambda^2_1, \Lambda^2_2, \ldots\}$=**compute_LRD_set**($M_2$, $\boldsymbol{O}(M_2)$)

3. For a given LRD Λ^1_i in $\boldsymbol{\Lambda}(M_1)$, find k closest LRDs from $\boldsymbol{\Lambda}(M_2)$:
   $\boldsymbol{\Lambda}^1_i\_closest$=$\{\Lambda^2_{i1},\Lambda^2_{i2}, \ldots,\Lambda^2_{ik}\}$, $\Lambda^2_j \in \boldsymbol{\Lambda}(M_2)$;
   Let $m(p^1_i,p^2_{ij})$: $\{\Lambda^1_i, \Lambda^2_{ij}\}$ define a match pair,
   Λ^2_{ij} is the j-th element in $\boldsymbol{\Lambda}^1_i\_closest$

4. For every match pair $m(p^1_a,p^2_b)$, obtained in step 3, find the
   corresponding $6DOF(p^1_a,p^2_b)$=**find_dof**($[O^1_a\ p^1_a]$, $[O^2_b\ p^2_b]$)

5. Cluster the 6DOFs obtained in step 4.

6. For each large cluster C_i, from step 5, construct an un-weighted graph G_i. A node in G_i is a match pair from C_i. An edge exists between two nodes in G_i if inter-point distances, corresponding to the match pairs in the two nodes, are preserved. Find the largest clique $S(G_i)$ in G_i and return the match pairs in $S(G_i)$ as the rigidly conserved domain pair.

Fig. S1. Outline of the algorithm to solve P from (Saha *et al*, 2010)

(Notations: See Fig. S1)

Algorithm **compute_frame_set**

Input: map M

1. At a grid location p_i of M,

   Cartesian reference frame O_i = **compute_frame**(M, p_i)

2. Return {O_1, O_2, …}

Algorithm **compute_frame**

Inputs: map M, grid location p_o in M

S1. Sample k points {p_1, p_2, ..., p_k} uniformly in the

   neighborhood (within r_o radius) of p_o

S2. Let v_i be the density value at p_i in M

   Define matrix P_{kx3} as [w_1*v_1*(p_1-p_o); w_2*v_2*(p_2-p_o); ....]

   - i-th row of P_{kx3} is w_i*v_i*(p_i-p_o)

   - w_i is a Gaussian wt: w_{o1}*exp(-(w_{o2}*|p_o-p_i|^2))

S3. [U_{3x3} D_{3xk} V_{3xk}]=SVD(P_{kx3})

S4. Return the Cartesian reference frame at p_o,

   O-XYZ(p_o): [O_x O_y O_z]=U_{3x3}

Fig. S2. Outline of algorithm for computing local Cartesian reference frames in Step 1 of Fig. S1, from (Saha *et al*, 2010)

(Notations: See Fig. S1)

Algorithm **compute_LRD_set**

Input: Map M, Cartesian frame set: $\{O_1, O_2, \ldots\}$ (O_i is the frame at p_i in M)

1. At a grid location p_i of M, LRD Λ_i = **compute_LRD**(M, p_i, O_i)

2. Return $\{\Lambda_1, \Lambda_2, \ldots\}$

Algorithm **compute_LRD**

Inputs: Map M, grid location p_o in M, Cartesian frame $O(p_o)$:$[O_x\ O_y\ O_z]$ at p_o

S1. Let H be a gradient histogram with m bins: $\{b_1, b_2, ..b_{8*26}\}$

S1.1 divide the region around p_o into 8 equal quadrants: $\{q_1, q_2, \ldots, q_8\}$, in the local frame $O(p_o)$. Let each quadrant have 26 representative directions: **D**:$\{d_1, d_2, \ldots, d_{26}\}$={{[-1/0/1, -1/0/1, -1/0/1]}-[0, 0, 0]}. d_i is finally normalized.

S1.2 bin b_i corresponds to $\{q(\text{ceil}(i/26)), d(1+i\%26)\}$

S1.3 initialize b_i=0

S2. Sample k points $\{p_1, p_2, ...p_k\}$ uniformly in the neighborhood (withen r radius) of p_o

S2.1 let V_i=$O(p_i)_x$

S2.2 let V_{i2}=$(O(p_o).V_i)'$

S2.3 let p_{i2}=$(O(p_o).(p_i-p_o))'$

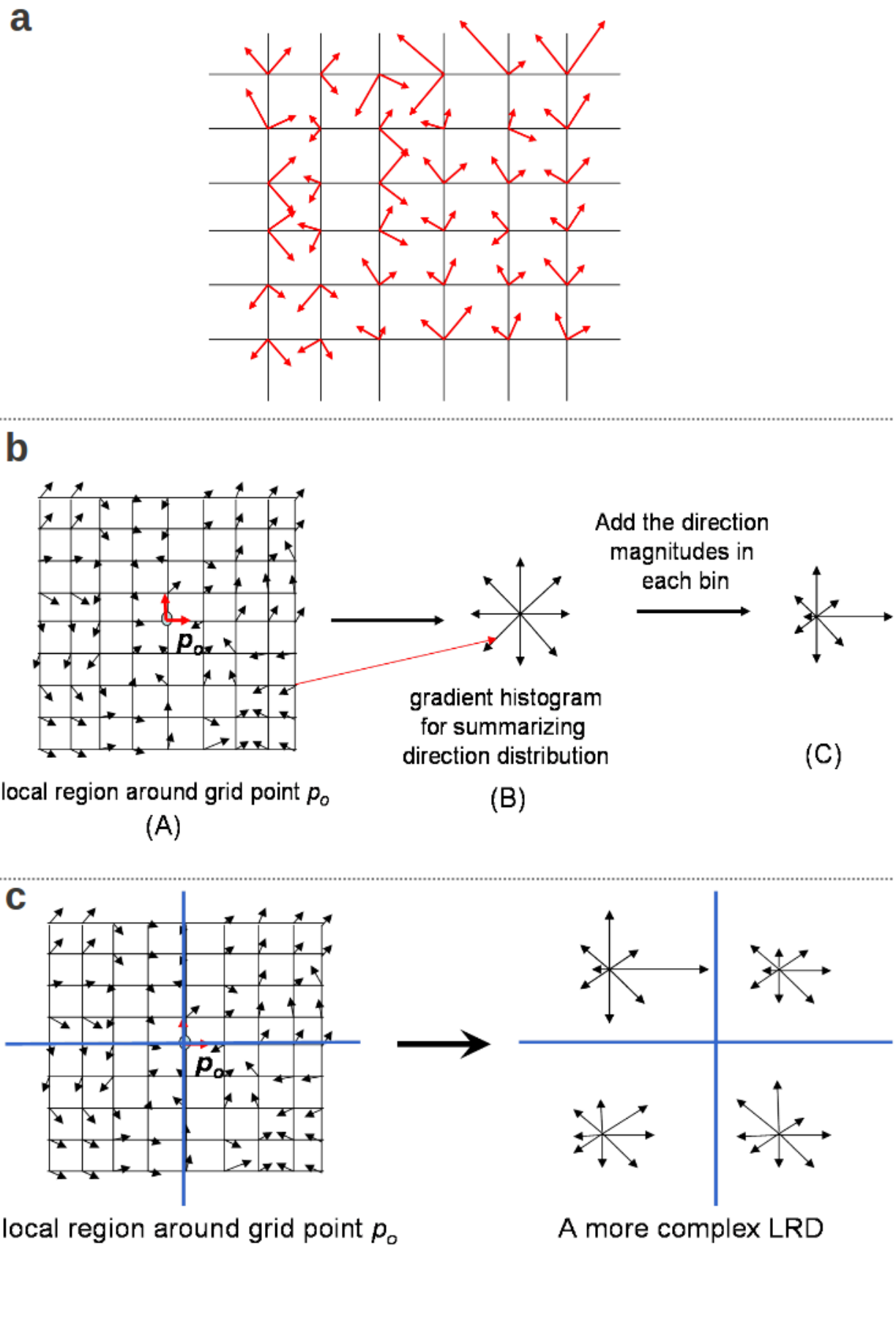
S2.4 find a bin b_i=$\{q_a, d_b\}$, such that p_{i2} is in q_a and d_b is the direction from **D** closest to V_{i2}.

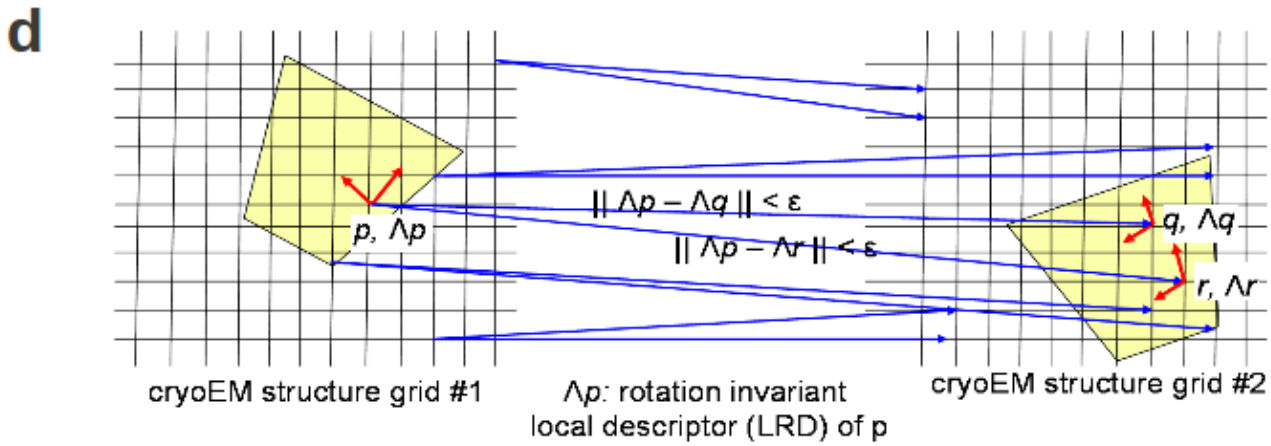
S2.5 let b_i+=$v_i.w_i$

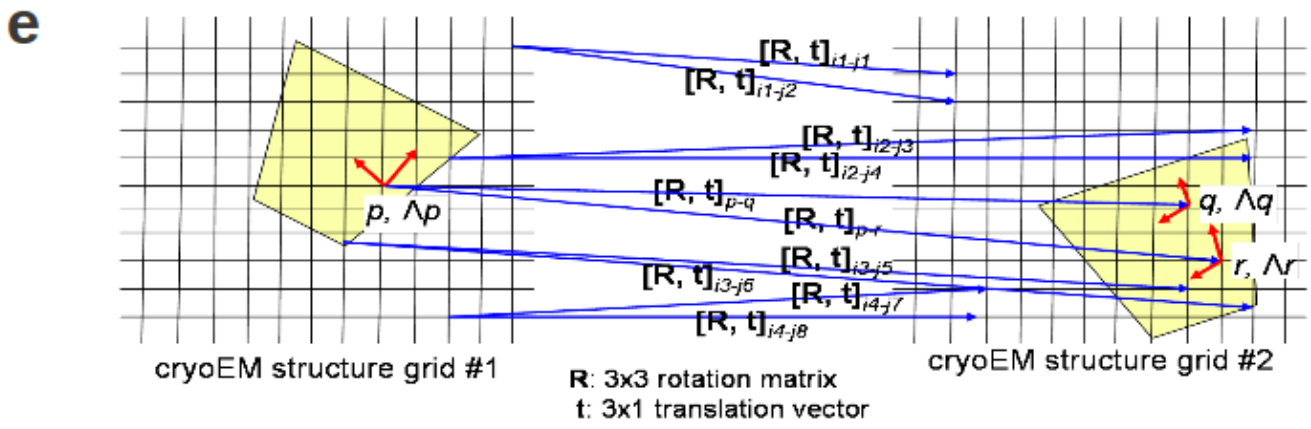
-v_i: magnitude of V_i or $D_{3xk}(1)$ obtained from step S3 in Fig. S2

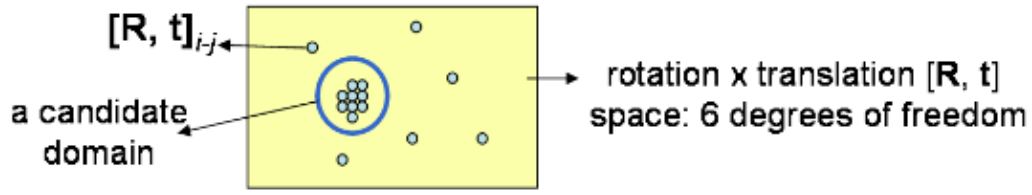
-w_i: Gaussian wt: $w_{o1}.\exp(-w_{o2}.|p_o-p_i|^2)$

Fig. S3. Outline of algorithm for computing local region descriptors (LRDs) in Step 2 of Fig. S1, from (Saha *et al*, 2010)


a
b
p_o
Add the direction magnitudes in each bin
gradient histogram for summarizing direction distribution
(C)
local region around grid point p_o
(A)
(B)
c
p_o
local region around grid point p_o
A more complex LRD


d
p, Λp
|| Λp – Λq || < ε
|| Λp – Λr || < ε
q, Λq
r, Λr
cryoEM structure grid #1
Λp: rotation invariant
local descriptor (LRD) of p
cryoEM structure grid #2


e
[R, t]i1-j1
[R, t]i1-j2
[R, t]i2-j3
[R, t]i2-j4
[R, t]p-q
[R, t]p-r
[R, t]i3-j5
[R, t]i3-j6
[R, t]i4-j7
[R, t]i4-j8
p, Λp
q, Λq
r, Λr
cryoEM structure grid #1
R: 3x3 rotation matrix
t: 3x1 translation vector
cryoEM structure grid #2


f
[R, t]i-j
a candidate
domain
rotation x translation [R, t]
space: 6 degrees of freedom

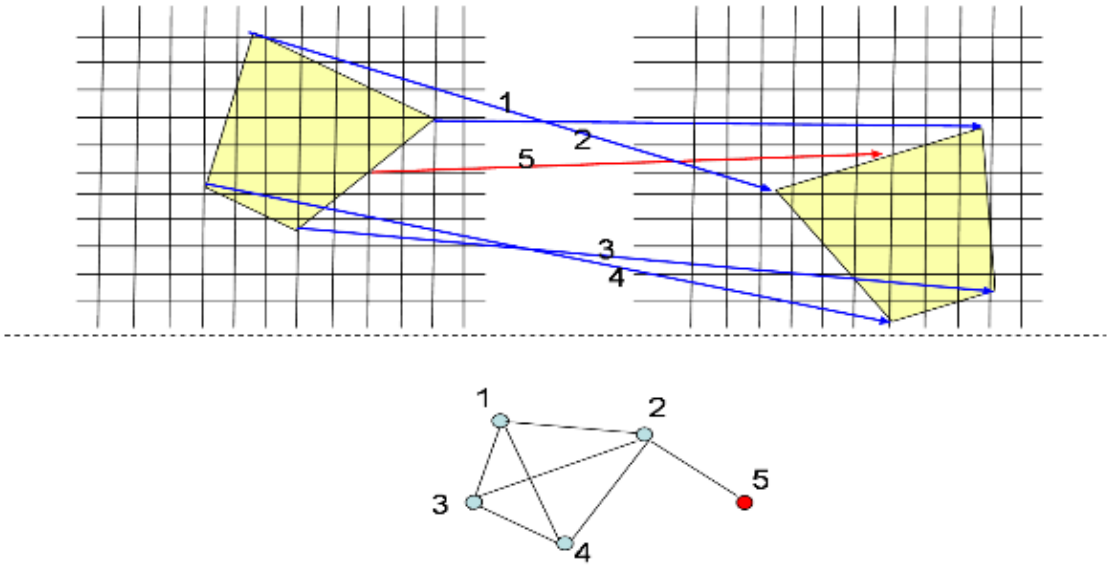

g
1
2
3
4
5
1
2
3
4
5

Fig. S4. Cartoon representation of the steps (of algorithm in Fig. S1) to solve **P**, from (Saha *et al*, 2010).

(a): Step 1 of Fig. S1, mimicked in 2D. A Cartesian reference frame is placed at each of the grid points. The length of a frame axis reflects the extent of local density variation along the axis.

(b): Step 2 of Fig. S1, LRD or gradient histogram construction, mimicked in 2D. The principal direction (X axis) of the reference frame of a grid point around p_o is first re-expressed in p_o's reference frame and then stored in the bin (of the gradient histogram) representing the direction closest to the re-expressed one. The magnitudes of the stored gradients in a bin are summed up to obtain a numerical value for each bin (reflected in the length of the directions in (C)).

(c): The local region around p_o can be divided into quadrants. LRDs, one from each quadrant, can be stacked together as a single vector to construct a more complex LRD.

(d): Step 3 of Fig. S1, mimicked in 2D. For a given grid point p in input cryoEM grid 1, locally similar grid points are found in the input cryoEM grid 2 by comparing the LRD at p with LRDs in grid 2.

(e): Step 4 of Fig. S1, mimicked in 2D. For a given match, there exists a spatial rotation **R** and a spatial translation **t**, that transforms the match pair onto each other.

(f): Step 5 of Fig. S1, mimicked in 2D. The match pairs obtained in Step 4 of Fig. S1 are clustered in the [rotation x translation] space.

(g): Step 6 of Fig. S1, mimicked in 2D. A graph is constructed such that match pairs are nodes. An edge between two nodes indicates that the distance between the corresponding two grid points is preserved between the two maps. A clique in the graph (formed by blue nodes) is a collection of grid points whose inter-point distances are preserved between the maps.

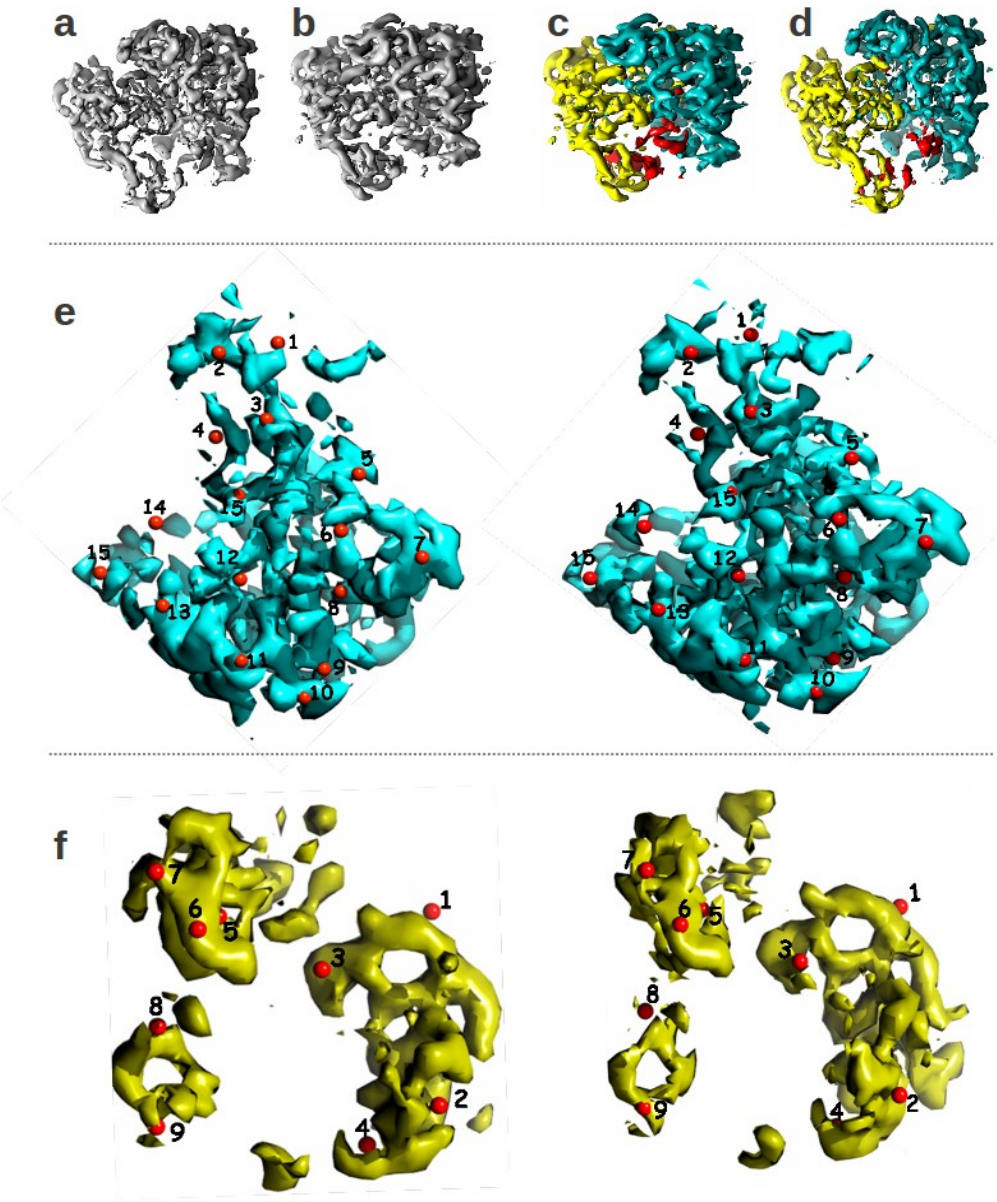


Fig. S5. A result from (Saha *et al*, 2010): Solving **P** using MOTIF-EM for a pair of ribosome 70S conformations. (a) & (b) show two low resolution (10 Å+) conformations (pre- and post-translocational states, respectively) of ribosome 70S. MOTIF-EM in (Saha *et al*, 2010) decomposes the two conformations into two rigid domains (the two regions colored as yellow and cyan) as shown in (c) & (d). The red region in (c) & (d) is the remnant non-conserved region in the input maps (a) & (b). (e) & (f) (enlarged compared to a & b) show the correspondences (numbered red balls), established by MOTIF-EM, between the first extracted domain pair (e) and the second extracted domain pair (f), respectively. The first extracted pair is predominantly the 30S subunit of the 70S ribosome, as per (Valle *et al*, 2003). The second extracted pair is predominantly the 50S subunit of the 70S ribosome, as per (Valle *et al*, 2003). Putting these

results together leads to the inference of ratchet like conformation change between the pre- and post- translocational states: http://cs.stanford.edu/~mitul/motifEM/rna_anim.gif

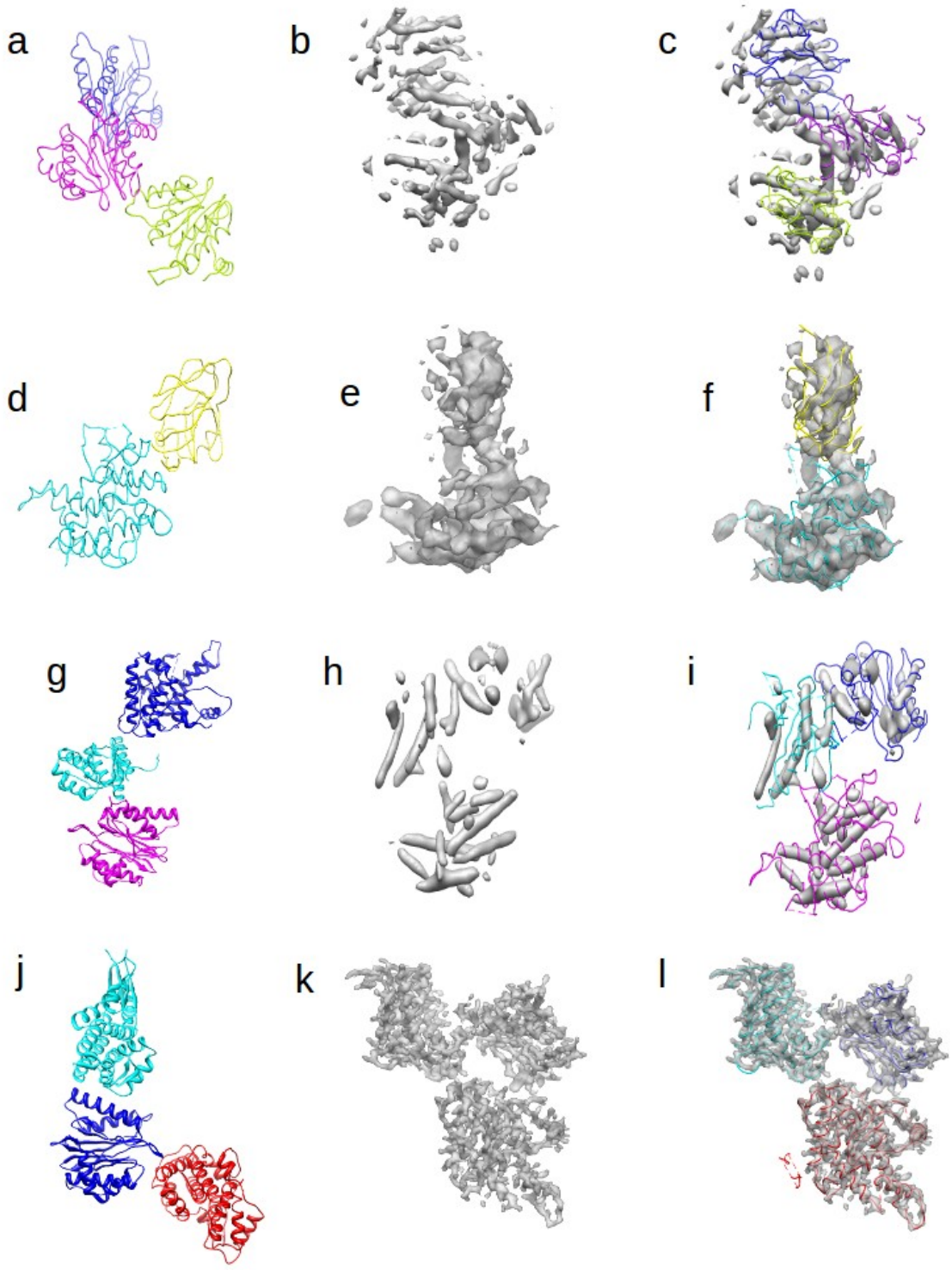


Fig. S6. Additional flexible fitting test-cases. (a): A synthetic atomic resolution conformation build using three copies of atomic resolution domain 1YAR (H:1-203). FOLD-EM flexible fitting was used to fit the synthetic conformation into a low resolution cryo-EM map of 20S proteasome shown in (b). (c) shows the fit of (a) into (b). The FOLD-EM flexible fitter had to alter the relative orientation and position of the individual domains in order to complete the fitting. Along the

same line, the rest of Fig. S6 shows three other test-cases ((d,e,f): flexible fitting into a rice dwarf cryo-EM map; (g,h,i): flexible fitting into synthetic map made from arbitrary spatial arrangement of domains 1KP8 (A:2-526), 1UF2 (C:1:147, C:301-421) & 1YAR (H:1-203); (j,k,l); flexible fitting into a synthetic map made from arbitrary spatial arrangement: 1KID (A), 1UF2 (C:1:147, C:301-421) & 1YAR (H:1-203)), where FOLD-EM flexible fitting was applied with similar outcome.

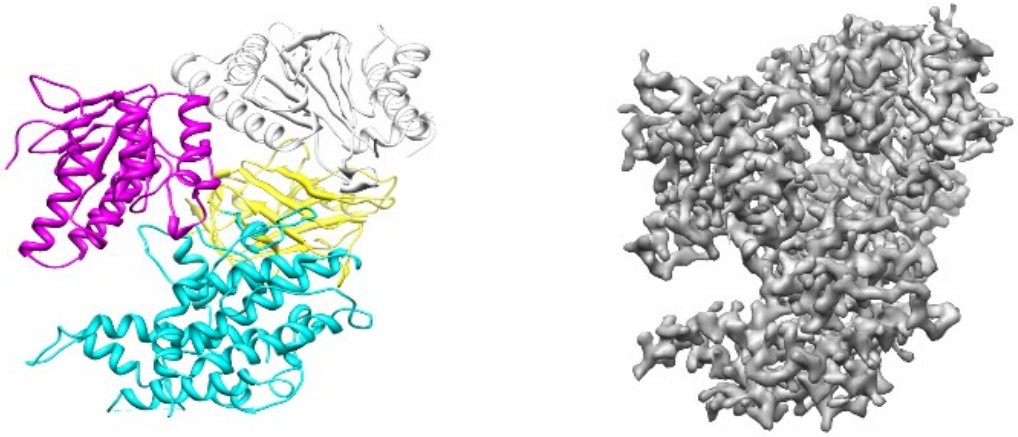


Fig. S7. Four atomic resolution domains (1KID (A), 1UF2 (C:1-147, C:301-421), 1UF2 (C:148-300), 1YAR (H:1-203)) were arbitrarily arranged in space to create a conformation shown in the figure at left. Synthetic cryo-EM maps (like the one at left) were created using EMAN from this conformation in the resolution range 5-15 Å.

Map resolution (Å)	Fitting Error (RMSD Å)	Fitting Error (RMSD Å)	Fitting Error (RMSD Å)
5	0.29	0.23	0.23
10	0.29	0.26	0.31
15	0.47	0.17	0.33
20	0.53	0.36	0.43

Table S1 (a). RMSD error in docking/fitting, using FOLD-EM

Column 2: Fitting errors for the intermediate domain of GroEL (size: 90 residues),

Column 3: Fitting errors for the apical domain of GroEL (size: 182 residues),

Column 4: Fitting errors for the equatorial domain of GroEL (size: 249 residues).

The fittings are done into simulated GroEL cryo-EM maps with resolution ranging from 5-20 Å (column 1).

(A RMSD (all-atom) error for a fitting is computed between the fitted atomic-resolution domain and the atomic-resolution domain used to simulate the map region where the fitting is supposed to occur).

ID of target Cryo-EM map; PDB ID of domain to be fitted	Fitting error at original map resolution (RMSD Å)	Fitting error at 10 Å map resolution (RMSD Å)	Fitting error at 15 Å map resolution (RMSD Å)
5001 (GroEL 4Å); 1AON (equatorial)	0.97	0.9	1.67
5001 (GroEL 4Å); 1AON (apical)	0.86	0.9	0.76
1060 (Rice Dwarf Virus 6.8 Å); 1UF2 (C:1-147& C:301-421)	0.87	1.12	1.58
1060 (Rice Dwarf Virus 6.8 Å); 1UF2 (C:148-300)	0.86	0.92	3.34
1120 (Phi29 7.9 Å); HK97	3.31	2.31	3.39
1740 (20S proteasome 6.8 Å); 1YAR (H:1-203)	0.4	0.78	0.73
1623 (Yeast FAS 5.9 Å); 2VKZ (A:392-933)	0.63	1.04	1.58
2005 (GTPgammaS microtubules	1.04	0.9	1.7

8.6 Å); 4ABO (A)			
1552 (Stressome; 8 Å); 2VY9 (A)	2.17	1.72	1.13
1079 (Metarhodopsin 5.5 Å); 1GZM (A)	0.5	0.71	1.05
5223 (Human Ndc80 microtubule 8.6 Å); 3IZ0 (A)	2.27	1.72	0.77
5155 (Bovine Papillomavirus 4.2 Å); 3IYJ (A)	0.44	1.14	1.87

Table S1 (b). Evaluation of the fitting/docking module of FOLD-EM on additional experimentally determined maps (from Electron Microscopy Data Bank (EMDB): http://www.ebi.ac.uk/pdbe-srv/emsearch/). For a given map, a corresponding atomic resolution domain (2nd entry in column 1) was fitted using FOLD-EM. Column #2 lists corresponding RMSD fitting errors. The following two columns list fitting errors when the same maps were filtered to lower resolutions (10 Å and 15 Å, respectively). The low pass filtering of the maps were done using EMAN. In all these additional thirty (ten maps, each at three levels of resolution) test-cases, FOLD-EM was able to successfully fit domains with reasonably low RMSD errors.

Tables S2 (a)-(d). RMSD errors in docking/fitting in the presence of extraneous regions.

Map resolution (Å)	Fitting Error (RMSD Å) (10% extra noise residues)	Fitting Error (RMSD Å) (20% extra noise residues)	Fitting Error (RMSD Å) (30% extra noise residues)
5	0.11	0.20	0.26
10	0.16	0.25	0.24
15	0.36	0.37	0.47
20	0.68	0.68	0.59

Table S2 (a): RMSD error in FOLD-EM generated fitting of an atomic resolution domain (intermediate domain of GroEL; size: 90 residues), with extraneous residues, into a simulated GroEL monomer with resolution ranging from 5-20 Å. Specifically, columns #2, #3, and #4 list errors when 10%, 20%, and 30%, respectively, extra residues were added as noise to the domain to be docked.

Map resolution (Å)	Fitting Error (RMSD Å) (10% extra noise residues)	Fitting Error (RMSD Å) (20% extra noise residues)	Fitting Error (RMSD Å) (30% extra noise residues)
5	0.11	0.15	0.12
10	0.11	0.18	0.16
15	0.18	0.23	0.20
20	0.24	0.24	0.32

Table S2 (b): RMSD error in FOLD-EM generated fitting of an atomic resolution domain (apical domain of GroEL; size: 182 residues), with extraneous residues, into a simulated GroEL monomer with resolution ranging from 5-20 Å. Specifically, columns #2, #3, and #4 list errors when 10%, 20%, and 30%, respectively, extra residues were added as noise to the domain to be docked.

Map resolution (Å)	Fitting Error (RMSD Å) (10% extra noise residues)	Fitting Error (RMSD Å) (20% extra noise residues)	Fitting Error (RMSD Å) (30% extra noise residues)
5	0.25	0.20	0.22
10	0.19	0.28	0.26
15	0.35	0.36	0.28
20	0.49	0.43	0.42

Table S2(c): RMSD error in FOLD-EM generated fitting of an atomic resolution domain (equatorial domain of GroEL; size: 249 residues), with extraneous residues, into a simulated GroEL monomer with resolution ranging from 5-20 Å. Specifically, columns #2, #3, and #4 list errors when 10%, 20%, and 30%, respectively, extra residues were added as noise to the domain to be docked.

Map ID; Domain ID	Fitting Error (RMSD Å) (X=10 % extra noise residue)	Fitting Error (RMSD Å) (X=20 % extra noise residue)	Fitting Error (RMSD Å) (X=30 % extra noise residue)	Fitting Error (RMSD Å) (X=10 % extra noise residue) (map filtered to Y=10 Å)	Fitting Error (RMSD Å) (X=20 % extra noise residue) (map filtered to Y=10 Å)	Fitting Error (RMSD Å) (X=30 % extra noise residue) (map filtered to Y=10 Å)	Fitting Error (RMSD Å) (X=10 % extra noise residue) (map filtered to Y=15 Å)	Fitting Error (RMSD Å) (X=20 % extra noise residue) (map filtered to Y=15 Å)	Fitting Error (RMSD Å) (X=30 % extra noise residue) (map filtered to Y=15 Å)
5001 (GroEL 4Å); 1AON (equatorial)	0.89	0.81	1.07	0.93	0.78	1.18	1.17	1.52	1.53
5001 (GroEL 4Å); 1AON (apical)	1.26	1.2	1.49	1.08	1	0.93	0.96	1.35	1.16

1060 (Rice Dwarf Virus 6.8 Å); 1UF2 (C:1-147& C:301-421)	0.69	0.56	0.77	1.18	1.09	0.65	1.52	1.83	2.36
1060 (Rice Dwarf Virus 6.8 Å); 1UF2 (C:148-300)	1.12	1.55	1.43	0.96	1.07	1.01	4.13	3.11	0.94
1740 (20S proteasome 6.8 Å); 1YAR (H:1-203)	0.62	0.6	0.57	1.49	0.64	0.62	0.48	0.96	0.87

Table S2(d): RMSD errors in FOLD-EM generated fitting of atomic resolution domains, with added extraneous residues, into experimentally determined cryo-EM maps. Column #1 lists the map EMDB ids and PDB ids of domains that were fitted. A subsequent column lists the fitting RMSD errors, when X% of extra noise residues are added to the domain to be fitted into a target map at resolution Y Å. Y is either the original resolution of the map or the new resolution (10 Å or 15 Å) of the map, after it being low pass filtered. The low pass filtering of the maps were done using EMAN. In all these 45 test-cases FOLD-EM was able to fit domains with reasonably low RMSD errors.

Tables S3 (a)-(d). Flexible fitting.

Map resolution (Å)	Fitting Error (RMSD Å) (equatorial domain)	Fitting Error (RMSD Å) (apical domain)
6	0.12	0.07
12	0.24	0.1
18	0.30	0.15

Table S3 (a): Error in the FOLD-EM predicted fitting, of conformation #1 (Fig. 7(a), left) into it's target map (Fig. 7(a), right), for each of the two domains in conformation #1.

Map resolution (Å)	Fitting Error (RMSD Å) (equatorial domain)	Fitting Error (RMSD Å) (apical domain)	Fitting Error (RMSD Å) (intermediate domain)
6	0.25	0.07	0.07
12	0.28	0.07	0.11
18	0.35	0.17	0.18

Table S3 (b): Error in the FOLD-EM predicted fitting, of conformation #2 (Fig. 7(b), left) into it's target map (Fig. 7(b), right), for each of the three domains in conformation #2.

Map resolution (Å)	Fitting Error (RMSD Å) (equatorial domain)	Fitting Error (RMSD Å) (apical domain)	Fitting Error (RMSD Å) (intermediate domain)	Fitting Error (RMSD Å) (intermediate domain #2)
6	0.09	0.08	0.07	0.06
12	0.14	0.15	0.14	0.09
18	0.20	0.16	0.36	0.49

Table S3 (c): Error in the FOLD-EM predicted fitting, of conformation #3 (Fig. 7(c), left) into it's target map (Fig. 7(c), right), for each of the four domains in conformation #3.

Map ID; domain in the atomic resolution conformation to be flexed and fitted	Fitting Errors (RMSD Å) (original resolution)	Fitting Errors (RMSD Å) (10 Å)	Fitting Errors (RMSD Å) (15 Å)
5001 (GroEL 4Å); equatorial domain of the conformation in Fig. 8 (a)	1.03	0.98	1.4
5001 (GroEL 4Å); apical domain of the conformation in Fig. 8 (a)	2.94	1.59	1.16
5001 (GroEL 4Å); intermediate domain of the conformation in Fig. 8 (a)	1.19	2.94	-
1740 (20S proteasome 6.8 Å); domain 1 of the conformation in Fig. S6 (a)	0.78	0.75	1.12
1740 (20S proteasome 6.8 Å); domain 2 of the conformation in Fig. S6 (a)	1.6	1.07	1.98
1740 (20S proteasome 6.8 Å); domain 3 of the conformation in Fig. S6 (a)	1.06	1.92	1.56
1060 (Rice Dwarf Virus 6.8 Å); domain 1 of the conformation in Fig. S6 (d)	0.52	0.65	1.08
1060 (Rice Dwarf Virus 6.8 Å); domain 2 of the conformation in Fig. S6 (d)	0.6	0.44	2.5
Synthetic cryo-EM map (Fig. S6 (h); domain 1 of the conformation in Fig. S6 (g)	0.04	0.06	0.1
Synthetic cryo-EM map (Fig. S6 (h); domain 2 of the conformation in Fig. S6 (g)	0.09	0.11	0.08
Synthetic cryo-EM map (Fig. S6 (h); domain 3 of the conformation in Fig. S6 (g)	0.32	0.23	1.38
Synthetic cryo-EM map (Fig. S6 (k); domain 1 of the conformation in Fig. S6 (j)	0.16	0.17	0.15
Synthetic cryo-EM map (Fig. S6 (k); domain 2 of the conformation in Fig. S6 (j)	0.14	0.25	0.35
Synthetic cryo-EM map (Fig. S6 (k); domain 3 of the conformation in Fig. S6 (j)	0.14	0.13	0.16

Table S3 (d): Error in the FOLD-EM predicted flexible fitting in more simulated and experimentally determined cryo-EM maps. A given row lists RMSD error in docking a domain into it's corresponding region in the target map at it's original resolution (column #2), at a

reduced resolution 10 Å resolution (column #3), and at a further reduced resolution of 15 Å (column #4). The resolutions of the target maps were reduced using the low pas filtering module in EMAN. In all cases (except for GroEL 15 Å, intermediate domain), FOLD-EM successfully fitted respective domains with reasonably low RMSD errors. In the case of fitting the intermediate domain into the GroEL 15 Å, we believe failure occurred because the domain is quite small and the map resolution is quite low too.

Tables S4 (a)-(i). Automated fold recognition

Top candidates for domain #1	Score (S_{AV}, S_{FE})	Top candidates for domain #2	Score (S_{AV},S_{FE})	Top candidates for domain #3	Score (S_{AV}, S_{FE})
1KP8 (A:2-136,A:410-526)	0.37, 250	1KP8 (A:137-190, A:367-409)	0.37, 96	1KID (A)	0.36, 177
1KP8 (A:137, A:367-409)	0.37, 96	1KID (A)	0.36, 177	1HF2 (A:100-206)	0.32, 59
1KID (A)	0.36, 177	1HF2 (A:100-206)	0.32, 59	2IOJ (A:206-325)	0.32, 63
2IOJ (A:206-325)	0.32, 63	2IOJ (A:206-325)	0.32, 63	1M1H (A:5-50, A:132-186)	0.32, 58
1HF2 (A:100-206)	0.32, 59	1M1H (A:5-50,A:132-186)	0.32, 58	2HI6 (A:1-132)	0.31, 63
1M1H (A:5-50,A:132-186)	0.31, 58	2HI6 (A:1-132)	0.31, 63	2DST (A:2-123)	0.31, 61
2HI6 (A:1-132)	0.31, 63	2DST (A:2-123)	0.31, 61	1ASS (A)	0.30, 83

Table S4 (a): This lists candidate domains, with associated scores ((S_{AV}: Chimera score, S_{FE}: FOLD-EM score); see METHODS for score definitions), automatically picked by FOLD-EM for the simulated GroEL 10 Å map. Three domains were picked: equatorial (column 1&2; column 2 is the associated FOLD-EM generated score), apical (column 3&4), and the intermediate domain (column 5&6). The first row lists the three domains with best scores, which are finally chosen by FOLD-EM to build the Cα model of the simulated map.

Map resolution (Å)	Fitting Error (RMSD Å)	Fitting Error (RMSD Å)	Fitting Error (RMSD Å)
5	0.48	1.4	0.63
10	0.49	1.41	0.63
15	0.54	1.4	0.66
20	0.73	1.44	0.68

Table S4 (b): RMSD error in docking/fitting, using FOLD-EM

Column 2: Fitting errors for the intermediate domain of GroEL (size: 90 residues),

Column 3: Fitting errors for the apical domain of GroEL (size: 182 residues),

Column 4: Fitting errors for the equatorial domain of GroEL (size: 249 residues).

The fittings are done onto simulated GroEL cryo-EM maps with resolution ranging from 5-20 Å (column 1).

Top candidates for domain #1	Score (S_{AV}, S_{FE})
1A7A (A:190-352)	5.35, 66
1QY9 (A:130-297)	5.21, 65
2FS2 (A:1-131)	5.14, 66
1F00 (I:658-752)	4.84, 64
2DI4 (A:406-607)	4.51, 68

Table S4 (c): This lists candidate domains for the first domain, with associated scores ((S_{AV}: Chimera score, S_{FE}: FOLD-EM score); see METHODS for score definitions), automatically picked by FOLD-EM for building the Cα backbone of the ϕ29 map. The correct domain 1F00 is ranked #4. After this domain is picked, the final domain (2FT1) is picked as the domain with best score among those which occupied the whole input cryoEM map together with the first picked domain 1F00.

.....

Top candidates for domain #1	Score (S_{AV}, S_{FE})	Top candidates for domain #2	Score (S_{AV},S_{FE})
1UF2 (C:1-147, C:301-421)	0.18, 100	1UF2 (C:148-300)	0.15, 86
1UF2 (C:148-300)	0.15, 86	1WN0 (B:11-138)	0.12, 49
1WN0 (B:11-138)	0.12, 56	1RCU (A)	0.08, 55
1RCU (A)	0.08, 55	1SUM (B)	0.06, 51
1SUM (B)	0.06, 54	4AIG (A)	0.06, 49

Table S4 (d): This lists candidate domains, with associated scores ((S_{AV}: Chimera score, S_{FE}: FOLD-EM score); see METHODS for score definitions), automatically picked by FOLD-EM for building the Cα backbone of this RDV map. Two domains were picked: P8 lower domain

(column 1&2; column 2 lists the associated scores), P8 top domain (column 3&4). The first row lists the three domains with best scores, which are finally chosen by FOLD-EM to build the Cα model of the map.

.....

Top candidates for domain #1	Score (S_{AV}, S_{FE})	Top candidates for domain #2	Score (S_{AV},S_{FE})	Top candidates for domain #3	Score (S_{AV}, S_{FE})
1YAR (H:1-203)	3.29, 100	1YAR (H:1-203)	2.56, 70	1YAR (H:1-203)	2.48, 65
1HQY (A)	2.59, 74	1HQY (A)	2.50, 70	1HQY (A)	2.46, 70
1YAR (H:1-203)	2.56, 70	1YAR (H:1-203)	2.48, 65	1HQY (A)	2.41, 67
1HQY (A)	2.50, 70	1HQY (A)	2.46, 70	1IAZ (A)	2.00, 67
1YAR (H:1-203)	2.48, 70	1HQY (A)	2.41, 67	1RVV (A)	1.99, 61

Table S4 (e): This lists candidate domains, with associated scores ((S_{AV}: Chimera score, S_{FE}: FOLD-EM score); see METHODS for score definitions), automatically picked by FOLD-EM for building the Cα backbone of this 20S map. Three domains were picked: 1YAR (column 1&2; column 2 lists the associated scores), 1YAR (column 3&4), and 1YAR (column 5&6). The first row lists the three domains with best scores, which are finally chosen by FOLD-EM to build the Cα model of the map.

Top candidates for domain #1	Score (S_{AV}, S_{FE})	Top candidates for domain #2	Score (S_{AV},S_{FE})	Top candidates for domain #3	Score (S_{AV}, S_{FE})	Top candidates for domain #4	Score (S_{AV}, S_{FE})
1UF2 (C:1-147, C:301-421)	0.88, 262	1KID (A)	0.86, 193	1YAR (H:1-203)	0.85, 210	1UF2 (C:148-300)	0.87, 153
1KID (A)	0.86, 193	1YAR (H:1-203)	0.85, 210	1UF2 (C:148-300)	0.87, 153	3BZY (A:246-262, B:263-345)	0.36, 64
1YAR (H:1-203)	0.85, 210	1UF2 (C:148-300)	0.87, 153	1HQY (A)	0.59, 113	1AZC (A)	0.35, 68
1UF2 (C:148-300)	0.87, 153	1HQY (A)	0.59, 113	3BZK (A:325-473)	0.36, 77	1HPL (A:337-449)	0.35, 65
1HQY (A)	0.59, 113	1ASS (A)	0.47, 86	2F9Z (C:1-157)	0.35, 74	2G0Y (A:7-138)	0.27, 65

Table S4 (f): This lists candidate domains, with associated scores ((S_{AV}: Chimera score, S_{FE}: FOLD-EM score); see METHODS for score definitions), automatically picked by FOLD-EM for building the Cα backbone of the 5 Å map simulated from the four domain atomic resolution structure shown in Fig. S7. The first row lists the four domains with best scores, which are finally chosen by FOLD-EM to build the Cα model of the simulated map.

Top candidates for domain #1	Score (S_{AV}, S_{FE})	Top candidates for domain #2	Score (S_{AV},S_{FE})	Top candidates for domain #3	Score (S_{AV}, S_{FE})	Top candidates for domain #4	Score (S_{AV}, S_{FE})
1UF2 (C:1-147, C:301-421)	0.40, 261	1YAR (H:1-203)	0.40, 210	1KID (A)	0.40, 193	1UF2 (C:148-300)	0.39, 152
1YAR (H:1-203)	0.40, 210	1KID (A)	0.40, 193	1UF2 (C:148-300)	0.39, 152	1G6G (A)	0.28, 77
1KID (A)	0.40, 193	1UF2 (C:148-300)	0.39, 152	1ASS (A)	0.33, 84	1CZS (A)	0.28, 74
1UF2	0.39, 152	1HQY (A)	0.37, 103	1G6G (A)	0.28, 77	1PJZ (A)	0.26, 75

(C:148-300)							
1HQY (A)	0.37, 103	2VB1 (A:1-129)	0.30, 88	1R8S (E)	0.27, 87	1R8S (E)	0.27, 88

Table S4 (g): This lists candidate domains, with associated scores ((S_{AV}: Chimera score, S_{FE}: FOLD-EM score); see METHODS for score definitions), automatically picked by FOLD-EM for building the Cα backbone of the 10 Å map simulated from the four domain atomic resolution structure shown in Fig. S7. The first row lists the four domains with best scores, which are finally chosen by FOLD-EM to build the Cα model of the simulated map.

Top candidates for domain #1	Score (S_{AV}, S_{FE})	Top candidates for domain #2	Score (S_{AV},S_{FE})	Top candidates for domain #3	Score (S_{AV}, S_{FE})	Top candidates for domain #4	Score (S_{AV}, S_{FE})
1YAR (H:1-203)	0.32, 206	1UF2 (C:1-147, C:301-421)	0.30, 255	1KID (A)	0.30, 192	1UF2 (C:148-300)	0.29, 145
1UF2 (C:1-147, C:301.421)	0.30, 255	1KID (A)	0.30, 192	1UF2 (C:148-300)	0.29, 145	3BZY (A:246-262, B:263-345)	0.28, 74
1KID (A)	0.30, 192	1UF2 (C:148-300)	0.29, 145	3BZY (A:246-262, B:263-345)	0.28, 74	1CZS (A)	0.26, 75
1UF2 (C:148-300)	0.29, 145	3BZY (A:246-262, B:263-345)	0.28, 74	1CZS (A)	0.26, 75	2ISB (A:2-179)	0.25, 82
1HQY (A)	0.29, 91	1CZS (A)	0.26, 75	2ISB (A:2-179)	0.25, 82	1MG7 (A:14-187)	0.22, 75

Table S4 (h): This lists candidate domains, with associated scores ((S_{AV}: Chimera score, S_{FE}: FOLD-EM score); see METHODS for score definitions), automatically picked by FOLD-EM for building the Cα backbone of the 15 Å map simulated from the four domain atomic resolution

structure shown in Fig. S7. The first row lists the four domains with best scores, which are finally chosen by FOLD-EM to build the Cα model of the simulated map.

Map resolution (Å)	Fitting error (RMSD Å)	Fitting error (RMSD Å)	Fitting error (RMSD Å)	Fitting error (RMSD Å)
5	0.61	0.27	0.18	0.28
10	0.46	0.49	0.12	0.25
15	0.41	0.50	0.19	0.07

Table S4 (i): RMSD error in docking/fitting of the final selected domains reported in Tables S4 (f)-(h). Row #1 corresponds to Table S4 (f), and so on. Column 2: Fitting errors for selected domain #1. Column 3: Fitting errors for selected domain #2. Column 4: Fitting errors for selected domain #3. Column 5: Fitting errors for selected domain #4. Just like in Table S4 (b), a fitting error was obtained by comparing the corresponding docked domain with the domain that was used to simulate the corresponding region in the target map.

Table S5 | Evaluation of FOLD-EM generated fittings

Fitting (Figure #; Structure name; Domain name; Reference)	Cα RMSD error (Å)
Fig 2 (a); GroEL 6Å; equatorial domain; (Ludtke *et al*, 2004)	1.01
Fig 2 (a); GroEL 6Å; apical domain; (Ludtke *et al*, 2004)	1.51
Fig 2 (a); GroEL 6Å; intermediate domain; (Ludtke *et al*, 2004)	2.87
Fig 3 (a); ϕ29 7.9Å; HK97 domain; (Morais *et al*, 2005)	3.31
Fig 3 (b) ϕ29 7.9Å; BIG2 domain; (Morais *et al*, 2005)	2.17
Fig 5 (d); GroEL 6Å; apical domain; (Ludtke *et al*, 2004)	3.69
Fig 5 (e); GroEL 6Å; equatorial domain; (Ludtke *et al*, 2004)	1.75
Fig 8 (d); GroEL 4Å; equatorial domain; (Ludtke *et al*, 2008)	1.03
Fig 8 (d); GroEL 4Å; apical domain (Ludtke *et al*, 2008)	2.94
Fig 8 (d); GroEL 4Å; intermediate domain; (Ludtke *et al*, 2008)	1.19
Fig 8 (h); GroEL 6Å; equatorial domain; (Ludtke *et al*, 2004)	1.06
Fig 8 (h); GroEL 6Å; apical domain; (Ludtke *et al*, 2004)	3.12
Fig 8 (h); GroEL 6Å; intermediate domain; (Ludtke *et al*, 2004)	2.66
Fig 10 (a); GroEL 6Å; equatorial domain; (Ludtke *et al*, 2004)	1.38
Fig 10 (a); GroEL 6Å; apical domain; (Ludtke *et al*, 2004)	2.67
Fig 10 (a); GroEL 6Å; intermediate domain; (Ludtke *et al*, 2004)	2.95
Fig 10 (b); ϕ29 7.9Å; HK97 domain; (Morais *et al*, 2005)	3.05
Fig 10 (b); ϕ29 7.9Å; BIG2 domain; (Morais *et al*, 2005)	-(*)-
Fig 10 (c); RDV 6.8Å; upper domain; (Zhou *et al*, 2001)	0.86
Fig 10 (c); RDV 6.8Å; lower domain; (Zhou *et al*, 2001)	0.87
Fig 10 (d); 20S 6.8Å; upper domain; (Rabi *et al*, 2008)	0.4
Fig 10 (d); 20S 6.8Å; middle domain; (Rabi *et al*, 2008)	2.28
Fig 10 (d); 20S 6.8Å; bottom domain; (Rabi *et al*, 2008)	1.84

Table S5: For a FOLD-EM generated fitting, reported in this manuscript, the above table shows the associated Cα RMSD error with respect to the corresponding fitting proposed by the authors of the respective low resolution cryo-EM structure. The authors of a cryo-EM structure may not have deposited/released the respective fitted high resolution domains along with the structure. So, in many of our test-cases we had to re-generate the fittings, as directed in the respective publications, with which we compared our results. For instance, in the case of Fig. 2 (a), (Ludtke *et al*, 2004) indicated where in the GroEL cryo-EM structure the individual domain regions

(equatorial, apical, intermediate) occur. These domain regions from the map were manually segmented and SITUS was used to fit the respective high resolution domains (from PDB ID: 1OEL, as indicated in (Ludtke *et al*, 2004)). A fitting resulting from SITUS was then compared with the corresponding fitting from FOLD-EM and the respective RMSD deviation between them is reported in the 2nd column of the above table. In the case of Fig. 3 (a), the authors of the structure were able to make available the fitted HK97 and BIG2 domains.

-(*)-: As stated in the main text, FOLD-EM retrieved a different domain (PDB ID: 1F00; residues: 658-752) from SCOP for the, so called, BIG2 region in the case of ϕ29 model building, than what is reported (PDB ID: 1F00; residues: 658-752) in (Morais *et al*, 2005). However, both these domains have similar fold and hence they are in the same SCOP family called "Invasin fragments from Yersinia pseudotuberculosis". We concluded that FOLD-EM preferring a different domain is not surprising because the BIG2 region is ambiguous, as the co-author of (Morais *et al*, 2005) recently proposed another domain (PDB ID: 2L04; (Pell *et al*, 2010)), with a similar fold, for this region.

Text S1 | MOTIF-EM

MOTIF-EM (Saha *et al*, 2010) solves the structural comparison problem **P** defined as follows: compare a non-atomic resolution structure (*i.e.*, a cryoEM map from EMDB) with another structure (another map or an atomic resolution structure) and identify conserved structural domains or motifs or sub-map (if there is any) between the pair of input structures. The precise algorithm used by MOTIF-EM to solve **P** is outlined in the Figs. S1, S2, & S3. The technique used by MOTIF-EM to detect conserved sub-structures is inspired by a recent breakthrough in 2D object recognition, called "scale-invariant feature transform" or SIFT (Lowe *et al*, 2004). The input to MOTIF-EM is a pair of volumetric electron density maps. The program then uses geometric processing, statistical analysis, and graph theory to detect conserved regions between the input pair by executing the following six steps. In step 1 (Figs. S1 (step 1), S2, & S4 (a)), three-dimensional Cartesian reference frames are assigned to every grid point in each of the input maps. These reference frames are computed by examining the local density variations at each grid point, using singular value decomposition. For example, the primary axis of the reference frame points to the direction of largest local density variation. In step 2 (Figs. S1 (step 2), S3, S4(b & c)), for each grid point in the input maps, we construct a local region descriptor (LRD) - a rotationally-invariant, low-dimensional representation of electron density variation in the local region around the grid point. The LRD for a grid point *p* is essentially an orientation histogram of the local density variation vectors around *p* that were calculated in step 1, *i.e.*, the first axis of the reference frames for the neighboring grid points. LRD is a simple 3D version of the 2D local descriptor (known as "keypoint" in (Lowe, 2004)) that was invented for the feature detection algorithm, known as SIFT, in (Lowe, 2004). In step 3 (Figs. S1 (step 3), S4(d)), for a grid point *p* in input map 1, we find *k* potential matches in input map 2, *i.e.*, local regions in map 2 which are "similar" to the local region around *p*. These matches are essentially those grid points in input map 2, whose LRDs closely match the LRD for point *p* in map 1. In steps 4 and 5 (Figs. S1 (step 4 & 5), S4(e & f)), we cluster all the

matches obtained from step 3, based on the six degrees of freedom geometric transformation that maps a grid point (along with the local reference frame) onto its match in map 2. In step 6 (Figs. S1 (step 6), S4(g)), we choose the most prominent cluster obtained in the previous step. The matches in this prominent cluster form the potential conserved domain between the input maps. False positives occur due to two main reasons: (a) high noise in either/both of the input maps and (b) dimensionally reduced representations (LRDs) used to characterize local regions necessarily result in information loss. However, these false positives are removed using graph theory; a graph is constructed with the matches in the prominent cluster as its node. An edge is added between two nodes in the graph if the inter-point distance (between the two grid points of the same map in the two graph nodes) is preserved across the input map pair. Finally the largest clique in the graph (the sub-graph with an edge between every pair of nodes) is the final predicted domain region that is structurally conserved between the pair of input maps. Steps 1-3 are the ones that are inspired by the SIFT algorithm in (Lowe *et al*, 2004). In (Lowe *et al*, 2004), the corresponding situation was to identify correspondences between a pair of 2D photographic images using local descriptors called “keypoints”, which we call LRDs here.

Briefly, the input to MOTIF-EM is a pair of volumetric electron density maps obtained either experimentally from cryo-EM image reconstructions or calculated from atomic coordinates. The program uses geometric processing, statistical analysis, and graph theory to detect conserved regions between the input pair by executing six steps. In laymen’s terms, MOTIF-EM first tries to find small regions of similarity between two maps, and then determines how each small region in one map must be rotated and translated to superimpose it on the equivalent region in the other map. Next, MOTIF-EM groups regions that require similar rotations and translations, and if the members of a group also form contiguous regions in the maps then this region is considered a

structurally equivalent domain/motif. Since MOTIF-EM also gives the relative spatial orientation of the common sub-structures extracted, the fitted domain/motif is returned along with its corresponding transformation matrix. If no homologous structural motifs are identified between the input pair, then no fitted structure is returned. Hence, MOTIF-EM only outputs fitted coordinates if meaningful structural homology can be detected. In this regard, MOTIF-EM differs from other fitting programs which always return a fitted structure regardless of whether or not the determined fit is meaningful.

In essence, MOTIF-EM identifies conserved domains/motifs in large macromolecular assemblies. Because domain/motif correspondences are built from matching smaller structural units, no prior knowledge of the extent of homologous domain/motif structures is required and the program will thus work even if only portions of domains are similar. In contrast, other fold recognition/fitting algorithms require that the structures being compared are similar over the entirety of the search structure. Similarly, MOTIF-EM is inherently able to compare and fit structures that have undergone conformational changes; the ‘bottom-up’ approach of assembling structural correspondences in MOTIF-EM assures that discrete conserved structural units are automatically identified and fitted separately, thus providing a computationally objective approach for performing flexible fitting. As a result, MOTIF-EM automatically identifies domains/motifs in large macromolecular assemblies that remain conserved upon conformational rearrangement. As a by-product, non-conserved regions in structures are also revealed, which can point to potentially important molecular flexibility. Hence, MOTIF-EM has the potential to facilitate biomedical research and discovery by accelerating the rate at which structures of large macromolecular assemblies can be determined and analyzed.

Text S2 | Validation of P22 results (shown in Figs. 6 (a-h))

We claim that alignment obtained by FOLD-EM (Fig. 6 (g)) is better than in (Jiang *et al*, 2003) (Fig. 6 (h); obtained using FOLDHUNTER (Jiang *et al*, 2001)) using these two means:

(a) visual inspection: in Fig. 6 (h), we circled the regions where local alignment can be clearly seen (by eyes) as worse than in corresponding regions in Fig. 6 (g).

(b) automated scoring: FOLD-EM had a better alignment score (obtained using Chimera) of 0.91 compared to 0.87 obtained by FOLDHUNTER.

FOLD-EM improves the alignment obtained using FOLDHUNTER by RMSD of 2.8 Å.

FOLD-EM was able to improve the alignment of the P22 subunits done in (Jiang *et al*, 2003) (using FOLDHUNTER), because it is able to automatically separate the conserved base domain from the rest in the two subunit maps (as seen in Fig. 6 (c) & (d)). FOLDHUNTER has no means to the separate conserved (base) and non-conserved regions, and hence looses it's accuracy due to inclusion of non-conserved regions while trying align the subunits by their conserved base. Also, very importantly, FOLDHUNTER needed an initial approximate alignment guess, whereas FOLD-EM didn’t.

Text S3 | Fold recognition in ribosome 70S

The two domains 50S and 30S of the ribosome 70S do not exist in the SCOP database. So, the way we came up with the fold shown in Fig. 10 (e) is as follows. We extracted the two low resolution domains (50S and 30S) from the 70S conformation #1 (from (Valle *et al*, 2003)) by comparing it with 70S conformation #2 (from (Valle *et al*, 2003)), as described in Fig. S5. We included the extracted domains along with other domains in the SCOP to build the fold for the conformation #2 using FOLD-EM. As expected, conformation #2 scored the best against the the two extracted domains. The authors of (Valle *et al*, 2003) have also released the high resolution models for these domains, which we used to construct the final fold shown in Fig. 10 (e). The point of this testcase of building a fold for 70S was to show that FOLD-EM is applicable to real cryo-EM maps with resolutions as low as 13 Å. In the future, we would like to also test FOLD-EM on real cryo-EM maps with resolutions worse than that.