Defining the limits and reliability of rigid-body fitting in cryo-EM maps using multi-scale image pyramids

G.C.P. van Zundert[a] and A.M.J.J. Bonvin[a,*]

[a] Bijvoet Center for Biomolecular Research, Faculty of Science – Chemistry, Utrecht University, Utrecht, 3584 CH, the Netherlands
[*] Correspondence: a.m.j.j.bonvin@uu.nl; Tel: +31-30-2533859; Fax: +31-30-2537623

Keywords1
Cross correlation; PowerFit; Ribosome; Fisher z-transformation; Core-weighted; Modeling.

Abstract

Cryo-electron microscopy provides fascinating structural insight into large macromolecular machines at increasing detail. Despite significant advances in the field, the resolution of the resulting three-dimensional images is still typically insufficient for *de novo* model building. To bridge the resolution gap and give an atomic interpretation to the data, high-resolution models are typically placed into the density as rigid bodies. Unfortunately, this is often done manually using graphics software, a subjective method that can lead to over-interpretation of the data. A more objective approach is to perform an exhaustive cross-correlation-based search to fit subunits into the density. Here we show, using five experimental ribosome maps ranging in resolution from 5.5 to 6.9Å, that cross-correlation-based fitting is capable of successfully fitting subunits correctly in the density for over 90% of the cases. Importantly, we provide indicators for the reliability and ambiguity of a fit, using the Fisher z-transformation and its associated confidence intervals, giving a formal approach to identify over-interpreted regions in the density. In addition, we quantify the resolution requirement for a successful fit as a function of the subunit size. For larger subunits the resolution of the data can be down-filtered to 20Å while still retaining an unambiguous fit. We leverage this information through the use of multi-scale image pyramids to accelerate the search up to 30-fold on CPUs and 40-fold on GPUs at a negligible loss in success rate. We implemented this approach in our rigid-body fitting software PowerFit, which can be freely downloaded from https://github.com/haddocking/powerfit.

[1] **Abbreviations**
CW = Core-weighted
FFT = Fast Fourier Transform
LCC = Local cross-correlation
L = Laplace pre-filtered

Introduction

A structural understanding of large macromolecular complexes is of fundamental importance to rationalize and manipulate cellular processes. Cryo-electron microscopy (cryo-EM) is quickly becoming the method of choice for studying these macromolecular machines as recent advances are enabling unprecedented levels of detail to be visualized (Bai et al., 2015). Sub-nanometer resolution maps are no exception anymore, although the level of detail is usually still too low for *de novo* building of atomic structures. When possible, cryo-EM data are therefore combined with high-resolution atomic models of subunits for a proper structural understanding of the data. Typically, the first step in the modeling process is placing the subunits in the density as rigid bodies, after which the models can be refined using some flexible fitting procedure (Esquivel-Rodriguez and Kihara, 2013).

A variety of tools and software have been developed to help users with rigid body fitting, both for manual and automatic placement. Though manual placement is frequently performed, most notably using UCSF Chimera (Pettersen et al., 2004), it is subjective and can lead to over-interpretation of the data. The problem is exacerbated when flexible fitting is applied afterwards, as it requires an initial local cross-correlation minimum between the model and the density, else the model would drift away from its fitted location.

An automatic and objective method to determine the placement of the subunits is to perform a full-exhaustive systematic cross-correlation search of the three translational and three rotational degrees of freedom of the model in the density. Further optimization of the fit can then be tried using the Fit in Map routines in UCSF Chimera. Many advances have been made in both sensitivity and speed of cross-correlation based rigid body fitting (Bettadapura et al., 2015; Chacón and Wriggers, 2002; Derevyanko and Grudinin, 2014; Farabella et al., 2015; Garzón et al., 2007; Hoang et al., 2013; Hrabe et al., 2012; Roseman, 2000; Volkmann and Hanein, 1999; Volkmann, 2009; Wu et al., 2003). Recently, we introduced the core-weighted local cross-correlation scores in our rigid-body fitting package PowerFit (van Zundert and Bonvin, 2015). However, to our knowledge, no thorough investigation into the limits of rigid body fitting has been performed so far, nor has the resolution requirements to fit a subunit of a certain size or shape in the density been quantified. In addition, as the size of cryo-EM data has been steadily increasing as a result of the higher information content, the CPU requirements for an exhaustive search, which is usually performed using Fast Fourier Transform (FFT)-techniques for fast translational scans, are considerably increasing, which slows down the entire process.

Here we report on a comprehensive exploration of cross-correlation based rigid-body fitting into cryo-EM densities, using five high-resolution ribosome maps in the range of 5.5 to 6.9Å for which high-resolution models are available. We analyze the success rate of fitting all 379 subunits into these maps as a function of resolution using four different scoring functions. This is done by progressively lowering the resolution of the initial data down to 30Å. We show how the size and shape of the subunits influences the success rate of fitting. Further we demonstrate that the Fisher z-transformation (Fisher, 1921) and its confidence interval

(Volkmann, 2009) are proper indicators for the accuracy and confidence of a fit, which allow to identify over-interpreted regions of the map. Finally, we leverage all this information by using the concept of multi-scale image pyramids (Cyganek, 2009), well known in the field of image analysis, to significantly reduce the required computational resources and time to perform a fit by up to two orders of magnitude. This is implemented in an updated version of our PowerFit package for fast rigid body fitting in cryo-EM data, which can be freely downloaded from https://www.github.com/haddocking/powerfit.

Methods

Exploring the limits of rigid body fitting

To explore the resolution limit for successful rigid body fitting, we selected five high-resolution cryo-EM ribosome maps from the EMDatabank (Lawson et al., 2011), ranging in resolution from 5.5 to 6.9Å, for which structural models were deposited in the Protein Databank (Gutmanas et al., 2014) (Table 1). The ribosome is an excellent case study as it contains many chains of various sizes and types.

Table 1. The five ribosome cases with high-resolution data and deposited structural models used in this work.

EMDB-ID	Resolution (Å)	PDB-ID	Molecular weight (Mda)	Number ofsubunits
1780 (Armache et al., 2010)	5.5	4V7E	4.2	88
2620 (Budkevich et al., 2014)	6.9	4UJE	4.5	83
2845 (Aylett et al., 2015)	6.5	4UER	1.2	39
5591 (Anger et al., 2013)	6.0	4V6W	4.0	86
5976 (Svidritskly et al., 2014)	6.2	3J77	3.5	83

Subsequently, we tried to fit each separate chain independently in their respective density map with PowerFit, using four different scoring functions: the local cross-correlation (LCC), the core-weighted (CW-) LCC, and their Laplace pre-filtered versions the L-LCC and L-CW-LCC, respectively, given by the master equation

$$\mathrm{CC}(\boldsymbol{T}, \boldsymbol{R}) = \frac{1}{N} \frac{\sum_i^N (w_i A_i - \overline{A_w}) \cdot (w_i B_i - \overline{B_w})}{\sigma_A \sigma_B} \qquad \text{Eq. 1}$$

where the summation is over all N voxels that are within half a resolution distance of any atom of the search object indexed by i; w_i is a weight factor given to voxel i; A_i and B_i are the intensities of the search object and the cryo-EM density at voxel i, respectively; $\overline{A_w}$ and $\overline{B_w}$ are the weighted density average for the search object and the local cryo-EM data, respectively, given by $\overline{X_w} = 1/N \sum_i^N w_i X_i$, where X is either A or B. Finally, σ_A and σ_B are the weighted density standard deviations for the search object and EM-data. The LCC-score is defined by setting w_i to 1, while for the CW-LCC w_i is given by the core-index (Wu et al., 2003; van Zundert and Bonvin, 2015), a measure for how close the voxel is to the core of the search object. The Laplacian enhanced scoring functions are defined by mapping $X_i \rightarrow \nabla^2 X_i$ (Chacón and Wriggers, 2002). The goal is to optimize the cross-correlation score by finding the correct translation $\boldsymbol{T}$ and rotation $\boldsymbol{R}$ of the rigid body in the cryo-EM density.

After each round of fitting, the resolution of the cryo-EM data was lowered by 1Å using the following procedure. Assuming that the density is described by a collection of atoms with a spherical Gaussian density distribution, where the width of the Gaussian depends on the resolution of the data, the density at each point $\vec{r}$ in space is given by

$$\rho(\vec{r} \mid R) = \sum_{i}^{N} A_i \exp\left(-\frac{\left|\vec{r} - \vec{r}_i\right|^2}{2\sigma_R^2} \right) \qquad \text{Eq. 2}$$

Here the summation is over all N atoms indexed by i, where the amplitude of the density is given by the atom number of the element A_i, $\vec{r}_i$ is the position of atom i, and the spread σ_R is a function of the resolution R given by

$$\sigma_R = \frac{1}{\sqrt{2}\pi} R \qquad \text{Eq. 3}$$

This definition of the resolution ensures that the amplitude of the specified resolution is at $1/e$ of the maximum in Fourier space. In order to lower the resolution of the map to a lower target resolution, the density can simply be convoluted with a Gaussian kernel as

$$\rho_{\text{target}}(\vec{r}) = G_k * \rho_{\text{init}} \qquad \text{Eq. 4}$$

where ρ_{init} and ρ_{target} are the initial and target density, respectively, and G_k is the Gaussian kernel with standard deviation σ_k, and * is the convolution operator. The convolution of two Gaussians results in another Gaussian (Weisstein, 2015) as follows

$$G_1 * G_2 = A \cdot \exp\left[-\frac{\left|\vec{r} - \left(\vec{r}_1 + \vec{r}_2\right)\right|^2}{2\left(\sigma_1^2 + \sigma_2^2\right)} \right] \qquad \text{Eq. 5}$$

where G_1 and G_2 are two Gaussian functions with center $\vec{r}_1$ and $\vec{r}_2$, and width σ_1 and σ_2, and A a normalization constant of no interest here. Thus, σ_k is then simply

$$\sigma_k = \sqrt{\sigma_{\text{target}}^2 - \sigma_{\text{init}}^2} \qquad \text{Eq. 6}$$

This procedure gives a handle and tool to lower the resolution of a map to a specified target resolution. After lowering the resolution, the data were resampled such that the voxel spacing was 1/4th of the new resolution using simple tri-linear interpolation.

Furthermore, we also explored the shape dependence of a chain on the success rate of fitting. To classify the shape of a chain, we used a rotation invariant descriptor, the relative shape anisotropy κ^2 (Arkin and Janke, 2013), given by

$$\kappa^2 = 1 - 3\frac{\lambda_x\lambda_y + \lambda_y\lambda_z + \lambda_z\lambda_x}{\left(\lambda_x + \lambda_y + \lambda_z\right)^2} \qquad \text{Eq. 7}$$

where $\lambda_x, \lambda_y, \lambda_z$ are the principal moments of the chain along each axis. The value of the relative shape anisotropy falls in the closed interval [0, 1], with 0 for spherical and 1 for linear particles.

Finally, to determine the ambiguity and statistical significance of a specific fit, we used the Fisher z-transformation (Fisher, 1921), introduced by Volkmann (2009) in cryo-EM

$$z(\mathrm{CC}) = \frac{1}{2}\ln\frac{1 + \mathrm{CC}}{1 - \mathrm{CC}} \qquad \text{Eq. 8}$$

Where CC is the cross-correlation coefficient. The Fisher z-transformation allows the determination of confidence intervals (Volkmann, 2009), using a standard deviation given by

$$\sigma_z = \frac{1}{\frac{\mathrm{MV}}{\mathrm{FC}} - 3} \qquad \text{Eq. 9}$$

where MV and FC are the molecular volume of the chain and the resolution at which the Fourier shell correlation is 0.5. The molecular volume was calculated as follows: a binary mask was generated from the molecule, where the radius of each atom was set to the element's van der Waal radius. The mask was projected onto a grid equal to the cryo-EM data, after which the volume was calculated by counted occupied voxels.

Leveraging the limits of rigid body fitting

The rapid advancement of the cryo-EM field has resulted in an impressive increase in the number of high-resolution density maps and corresponding atomic models. The increase in the level of detail, however, also requires the number of voxels to represent the data to rise. Consequently, the time required for an exhaustive search can increase dramatically as fitting algorithms typically use the Fast Fourier Transform (FFT) for rapid translation correlation scans, which scale with $N \log N$ where N is the number of voxels along an axis.

The actual level of detail present in current high-resolution maps may, however, be far surpassing the minimal required information to successfully fit a subunit into the density. The superfluous amount of information can be leveraged by building a multi-scale image pyramid to speed up the search: by progressively lowering the resolution and subsampling the data, the size of the density is reduced, which subsequently results in lower computational resources and time requirements (see Figure 1). However, for the image pyramid concept to work effectively, the resolution boundaries to perform a successful fitting of a particular subunit must be established. Natural parameters to investigate are the size and shape of the subunit, expressed here simply as the number of residues, since larger chains carry more information and thus require a lower level of detail to be properly fitted in the density, and the relative shape anisotropy, as spherical particles might be easier to fit than linear subunits. Once the minimal resolution requirements to successfully fit a particularly sized or shaped chain have been explored (as described in the previous section), this information can be used to deduce the proper resolution levels for building an image pyramid. Thus, larger subunits will be fitted in lower resolution maps to enhance the speed of fitting.

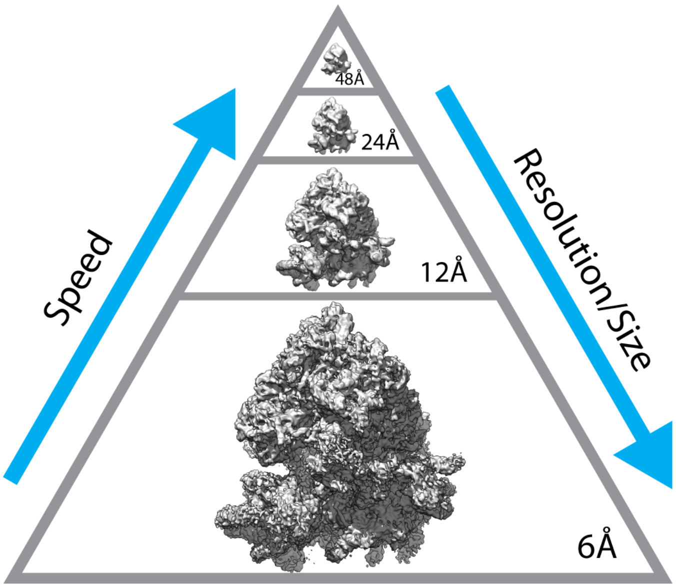

Figure 1. Example of a multi-scale image pyramid of the D. melanogaster ribosome (EMD-5591). The time required for an exhaustive search increases with increasing resolution.

Results and discussion

Translational, rotational and noise sensitivity of each scoring function

We first determined the best performing cross correlation score for fitting subunits in the experimental maps. As fine-grained 6-dimensional searches are computationally demanding, we investigated the translational and rotational performance of each scoring function separately. Thus, we performed a translational correlation scan with the correct orientation of each chain for all cases, and a fine rotational (6.6° interval, 27672 orientations) search for one case (EMD-1780) where the correct translation was used. The translational normalized success rate of fitting a subunit correctly is plotted against the resolution of the data in Figure 2A. A fit is considered successful if the subunit is placed within 2 voxels of the true solution, and we only considered the best-ranked fit, i.e. the fit with the highest correlation score. Congruent with an earlier analysis of noise-free simulated data, the L-CW-LCC score performs the best of the four, followed by the L-LCC, CW-LCC and LCC (van Zundert and Bonvin, 2015). Remarkably, the L-CW-LCC is capable of fitting about 90% of the 379 subunits down to a resolution of 13Å. Inclusion of the Laplace pre-filter has the biggest impact and increases the applicable resolution extent by about 5Å. The core-weighted approach has a smaller impact and extents the resolutions 1 to 2Å further.

The rotational analysis is shown in Figure 2B, where a successful fit is defined if the top solution is the identity rotation. In contrast to the translational analysis, here the CW-LCC is capable of fitting the most subunits correctly in the density, followed by the L-CW-LCC, L-LCC, and LCC score. Interestingly, the success rate of the Laplace pre-filtered solutions first increases up to a resolution of 12Å, after which it gradually decreases. The same can also be seen for the CW-LCC scoring function, where the success rate very slightly increases first. The LCC score is more regular, and steadily decreases, but its starting success rate is less than 30%.

Finally, we also investigated the noise robustness of each scoring function. Simulated 8Å cryo-EM data were generated using Eq. 2 from the fitted model (PDB-id 4V7E) of EMD-1780 that served as a noise-free ground truth, after which Gaussian noise was added until the Fourier shell correlation between the ground truth and the noisy data reached 0.5 for a specified resolution. We generated noisy data with a corresponding resolution between 8 and 15Å at a 1Å interval. Note that the data were not resampled, i.e. the voxel size was kept at 2Å. The Gaussian noise model may not fully describe real cryo-EM noise, but it at least gives an indication of each score's robustness. A fine 6-dimensional (6.6° interval, 27672 orientations) search was performed for each individual subunit, and a fit was deemed successful if the RMSD of the top solution was within 5Å. The success rate against the resolution is given in Figure 2C and the RMSD of each chain's best solution in the top 1, 10 and 100 in Figure S1. Even though the Laplace filter is known to increase noise, the L-LCC and L-CW-LCC are performing significantly better than the LCC and CW-LCC scores, and are capable of fitting all subunits at their correct position at 8Å resolution. Even at a resolution of 15Å, the Laplace scores fit approximately 60% of the 88 chains in the density, compared to 10% of the non-Laplace scores. Furthermore, it shows that including the core-weighted correlation function decreases the score's robustness at high noise levels, although this effect is marginal for the Laplace pre-filtered score.

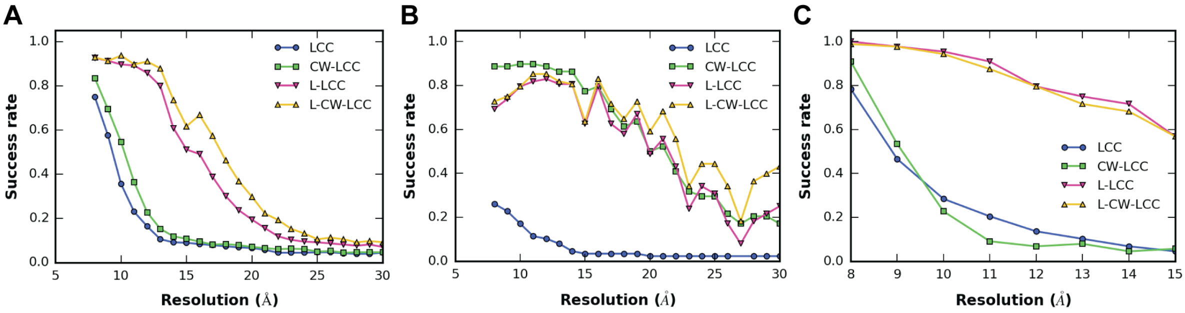

Figure 2. Translational, rotational, and noise-robustness success rate of different scoring functions. Normalized success rate of fitting a subunit at the correct position for four correlation scores as a function of the density map resolution for **(A)** a translational scan using the correct orientation, **(B)** a rotational search at the correct translation, and **(C)** a fine 6-dimensional search in noisy cryo-EM data.

Size and shape dependence of success rate

Based on the above analysis, we decided to quantify the success rate of the L-CW-LCC further as it showed the best balance between sensitivity and robustness. We performed a fine 6-dimensional search (6.6° interval, 27672 orientations) for each subunit for all 5 maps. We divided the chains in 7 size categories based on the number of residues (Table 2). The success rate of fitting each category of subunits is shown in Figure 3A for which we only considered the top 1 fit, where we considered a fit successful if the RMSD was smaller than half the resolution, i.e. 4Å for 8Å resolution data and 10Å when using 20Å resolution. The RMSD of the best solution in the top 1, 10, and 100 are shown individually in Figure S2. As expected, the smallest chains have the lowest success rate. Even when fitting in 8Å resolution data the success rate is less than 50% for chains consisting of 0 to 50 residues. This increases to around 80% already for subunits with a residue count of 50 to 100. The success rate stabilizes to 90% for larger sized chains and is stable down to 12Å resolution data. After the 12Å point, the success rate drops rapidly, though less strongly for larger chains. For subunits larger than 500 residues, which also include the rather large rRNA chains, the success rate remains stable down to 20Å. Thus, the bulk of the subunits can be properly fitted in the density down to 12Å resolution data.

Table 2. Size categories used in the size analysis and number of corresponding subunits.

Number of residues	Number of subunits	Average number of residues	Molecular Weight (kDa)
0 – 50	9	25	4.6
50 – 100	87	73	8.7
100 – 150	122	125	14.1
150 – 200	67	174	19.3
200 – 300	67	226	23.6
300 – 500	15	365	37.9
500+	12	2090	609.6

In addition to the size dependence, we also investigated the shape dependence on the success rate. To this end, we calculated the relative anisotropy κ^2 for subunits consisting of 100 – 250 residues to minimize the impact of the size, while keeping a decent population sample. The relative anisotropy is a rotation-invariant shape descriptor in the interval [0, 1], where 0 denotes spherical and 1 linear particles (Arkin and Janke, 2013). We divided the subunits in five bins, with a 0.2 step size (see Table 3 and Figure S3 for examples). For each bin, we again plotted the normalized success rate against the resolution (Figure 2B). Surprisingly, the influence of the shape is marginal on the success rate of fitting. The initial success rate starts between 0.9 and 1.0 for fitting in 8Å resolution data and starts to drop significantly at 14Å resolution. At the intermediate resolutions, the success rate is slightly lower for anisotropic chains, and might be caused by the fact that the average chain size is smaller for highly anisotropic chains. Regardless, the size influence has a markedly higher impact on the success rate of fitting than the chain's shape.

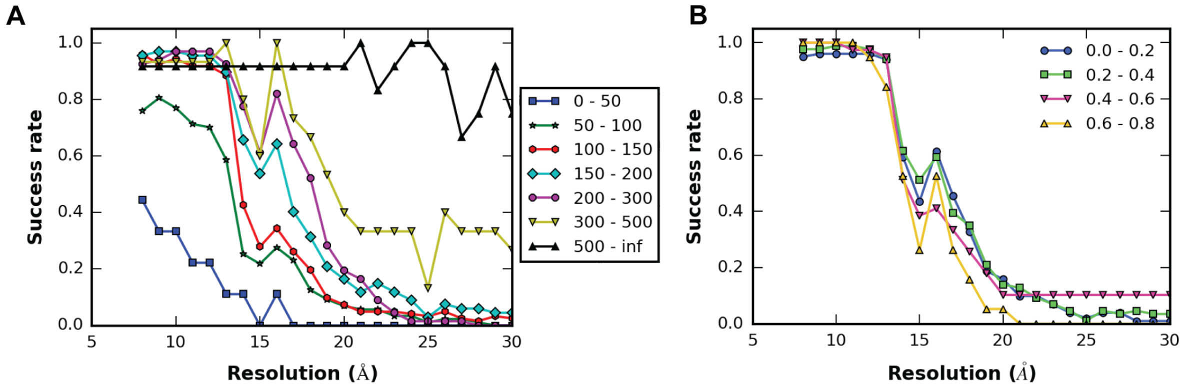

Figure 3. Size and shape dependence of fitting success rate. Normalized success rate for fitting subunits, binned in several size **(A)** and shape **(B)** categories based on their number of residues and relative shape anisotropy, respectively. For the shape dependence, only chains consisting of 100 to 250 residues were considered.

Interestingly, the success rate spikes locally at 16Å resolution, and is more pronounced for larger chains. This can also be observed in Figure 2A only for the L-CW-LCC score. The reason for this is not fully apparent. It might be an artifact of the core-weighted procedure: the core-indices of subunits consisting of multiple subunits may shift suddenly and coalesce, locally increasing the sensitivity of the score. Another reason might be that for those subunits the local resolution is significantly lower, and that fitting with a high-resolution template of the subunit results in too much noise entering the correlation score, which is remedied by filtering the template further down to lower resolutions. Or simply the resampling of the cryo-EM data results in better aligned solutions coincidentally at 16Å. Although this has no impact on the main finding of the fitting analysis, the observation is intriguing.

Table 3. Shape categories used in the shape analysis and number of corresponding subunits.

Relative anisotropy	Number of subunits	Average number of residues	Average molecular weight (kDa)
0.0 – 0.2	101	161	16.8
0.2 – 0.4	86	170	19.2
0.4 – 0.6	39	148	18.1
0.6 – 0.8	19	134	14.4
0.8 – 1.0	0	-	-

In addition, to determine whether the top scoring solution is an unambiguous fit, we used the Fisher z-transformation and its associated confidence intervals (Fisher, 1921; Volkmann, 2009) to establish the significance of the top fit compared to the second best scoring solution. The difference in standard deviations between the top 2 best scoring solutions is given for each chain individually in Figure S4. Remarkably, in 65% of the successfully fitted cases, the difference in Fisher's z-score was more than 3 standard deviations between the 2 top ranked solutions. This increases to 75 and 87% for at least 2 and 1 standard deviations difference. Moreover, in less than 0.01% of the failed cases is the difference in z-score higher than 2 sigma, and only 3% for 1 sigma. To quantify the reliability of a fit given a Fisher z-score difference, we binned all solutions in six bins starting from 0 to 3 sigma difference in steps of 0.5, and calculated the true positive rate (Figure 4A). If we were to trust blindly the z-score-based selection for all cases used in this study, our true-positive success rate is >99.4% for >3 sigma, dropping to 96.7% for 2 – 2.5 sigma, 59.7% for 1 – 1.5 sigma, and 5.3% for 0 to 0.5 sigma difference. Thus, the Fisher z-score together with its associated confidence intervals is an excellent indicator for formally evaluating the accuracy of a fit.

Finally, we investigated the accuracy of each successful fit, i.e. RMSD smaller than half the resolution, by calculating the average RMSD compared to the ground truth at a given resolution (Figure 4B). As expected, the accuracy of the fit decreases with decreasing resolution, but the drop is relatively linear and on the order of the used voxel spacing (1/4th of the resolution), which is very reasonable.

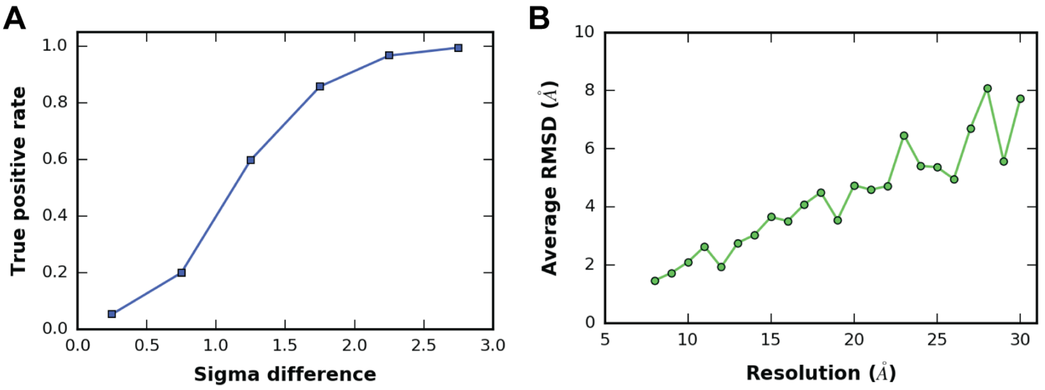

Figure 4. True positive rate and fitting accuracy. (A) The true-positive rate is given versus the difference in Fisher z-score standard deviations between the top 2 solutions. The fitting results were binned in 6 bins, starting from 0 to 3 sigma with a step size of 0.5. **(B)** The average RMSD of a successful fit compared to the correct solution versus the cryo-EM data's resolution.

Detecting over-interpreted regions of the density

The advantage of objectively fitting subunits in the density and characterizing the success rate is that it allows the identification of possibly over-interpreted regions of the density. For example, in the largest size category the eIF3c chain (543 residues) of EMD-2845 was placed incorrectly in the density. The unreliability of this solution was flagged by a very small difference in Fisher's z-score between the top 2 solutions at 8Å resolution (<0.35 sigma). When inspecting the current fit (Figure 5A), it shows that the global features of the chain are present in the density, although it is not of sufficient resolution to identify the secondary structure elements, and some parts are sticking outside the envelope. This was also implicated by the authors, as the local resolution of the density drops to around 10 to 15Å in that region (Aylett et al., 2015). Interestingly, the best scoring solution is fitted at the correct location in the density map starting from around 15Å resolution. The orientation, however, is highly ambiguous, as other solutions within 1 standard deviation are also placed at the same position (Figure 5B-D). As such, the density data here indicates that the fit is ambiguous, and that additional (experimental) data are required to validate the model, using integrative modeling methods such as offered by HADDOCK (van Zundert et al., 2015).

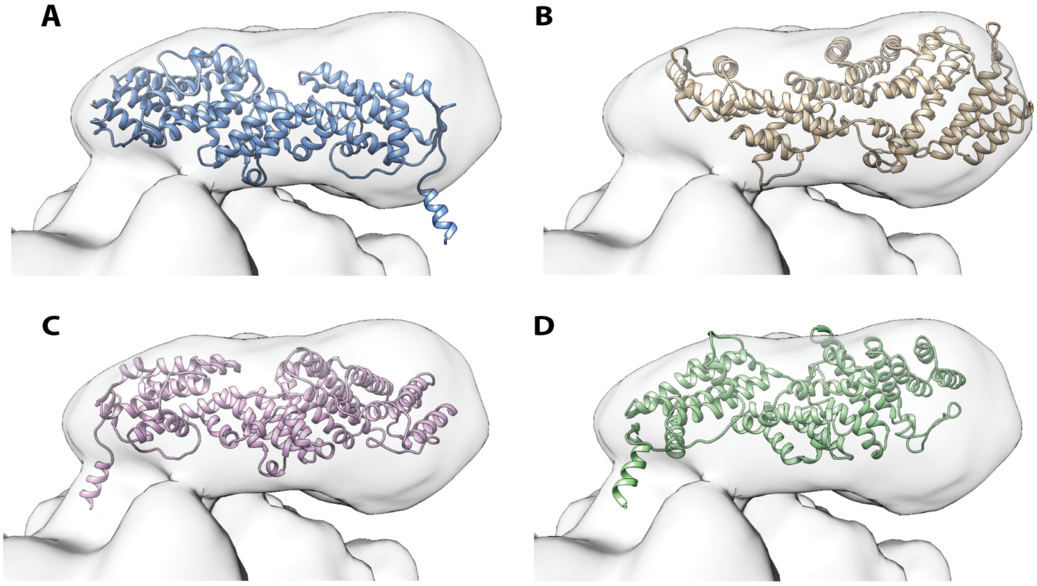

Figure 5. The eIF3c chain fitted in the cryo-EM density. Shown are the current fit **(A)**, and the 1st **(B)**, 3rd **(C)** and 4th **(D)** ranked solution found by PowerFit.

Fast fitting with multi-scale image pyramids

Now that resolution boundaries are defined for reliably fitting a particular sized chain into the density, this knowledge can be leveraged through building a multi-scale image pyramid to speed up the search. We discard the shape information here, as its influence was shown to be marginal. To demonstrate the speedup that can be achieved, we applied our approach to another ribosome case with a reported resolution of 5.7Å (EMD-2917, 5AKA) (von Loeffelholz et al., 2015). We constructed an image pyramid by filtering and resampling the original data down to 9, 12, 13 and 20Å resolution (Figure 6). We fitted only chains larger than 50 residues: chains consisting of 50 to 100 residues were fitted in the 9Å resolution density, chains consisting of 100 to 300 residues in the 12Å data, chains consisting of 300 to 500 residues in the 13Å map, and for subunits bigger than 500 residues the 20Å data was used. The chosen target resolutions in the image pyramid were based on a success rate of 90% for each size category based on Figure 3A. Even smaller maps can be attained by down filtering to lower resolutions, at the cost of a reduced success rate. For the fitting, we applied the L-CW-LCC score using a fine rotational search (6.6° interval, 27672 orientations).

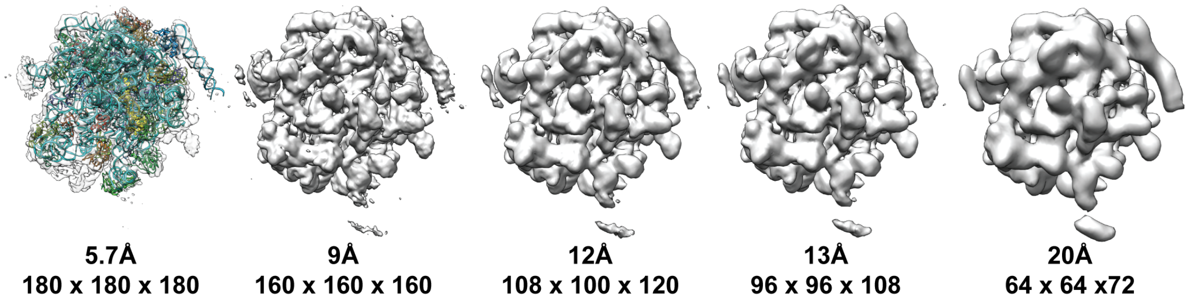

Figure 6. Cryo-EM data of E. coli ribosome (EMD-2917) at different resolutions with the deposited structure (5AKA) fitted into the original map (left). The resolution and the size of the data, the latter expressed in numbers of voxels, are indicated under each density.

All 31 chains were successfully fitted considering only the best scoring solution, with the exception of the 4.5S RNA consisting of 74 bases. A local cross correlation maximum can be found at the correct location, with the successful fit placed at rank 17. This is probably due to the local resolution of the data dropping to around 10 to 12Å in that region, indicating flexibility of the chain (von Loeffelholz et al., 2015). The time required to fit one subunit into the original map (180×180×180 voxels) is approximately 10 hours using a single AMD Opteron 6320 CPU processor and 40m for an NVIDIA GTX 680 GPU. This reduces to 6h and 29m for the 9Å resolution data (160×160×160, 2h and 7m for the 12Å data (108× 100×120, 1.5h and 5m for the 13Å data (96×96×108, and 20m and 1m for the 20Å data (64×64×72, respectively. Thus, the speed increase is up to 30 times for CPU and 40 times for GPU calculations for the larger subunits, at only a small cost in the success rate of fitting.

Conclusions

Here we have explored the resolution limits of rigid body fitting in high-resolution cryo-EM densities, ranging between 5.5 and 6.9Å resolution, using 5 different ribosome cases. We have shown that also for experimental data the L-CW-LCC score is the most sensitive of the 4 correlation-based scores tested and that it can successfully fit most chains in the density. In addition, we have quantified the success rate of fitting subunits based on their size and shape represented by their number of residues and relative shape anisotropy. As expected, larger subunits require a lower level of detail to be successfully fitted into the density, while the shape has an almost negligible impact. This phenomenon can be leveraged by building an image pyramid, i.e. representing the data at different resolutions, and subsequently fitting a subunit in the smaller, lower-resolution density dataset. The resulting speed increase can be up to 30-fold for CPUs and 40-fold for GPUs with virtually no loss in the success rate of fitting and only a small decrease in the fitting accuracy, typically on the order of one voxel spacing. We have implemented the use of image pyramids in an updated version of PowerFit for fast objective fitting of high-resolution structures in lower-resolution density maps (freely available from https://www.github.com/haddocking/powerfit). Furthermore, the use of the Fisher z-transformation in combination with its standard deviations is an excellent parameter to determine the reliability of a fit, and provides an indication for when additional data are required to validate a proposed model.

Acknowledgements

This work was supported by the Dutch Foundation for Scientific Research (NWO) [ECHO grant no.711.011.009] and by H2020 e-Infrastructure grants (West-Life, grant no. 675858; EGI-Engage, grant no. 654142 and INDICO-DataCloud, grant no. 653549). We also wish to thank the reviewers of this paper for their excellent suggestions which we have implemented.

References

Anger, A.M., Armache, J.-P., Berninghausen, O., Habeck, M., Subklewe, M., Wilson, D.N. and Beckmann, R., 2013. Structures of the human and Drosophila 80S ribosome. Nature 497, 80-85.

Arkin, H. and Janke, W., 2013. Gyration tensor based analysis of the shapes of polymer chains in an attractive spherical cage. J. Chem. Phys. 138, 054904.

Armache, J.-P., Jarasch, A., Anger, A.M., Villa, E., Becker, T., Bhushan, S., Jossinet, F., Habeck, M., Dindar, G., Franckenberg, S., Marquez, V., Mielke, T., Thomm, M., Berninghausen, O., Beatrix, B., Söding, J., Westhof, E., Wilson, D.N. and Beckmann, R., 2010. Localization of eukaryote-specific ribosomal proteins in a 5.5-Å cryo-EM map of the 80S eukaryotic ribosome. Proc. Natl. Acad. Sci. USA 107, 19754-19759.

Aylett, C.H.S., Boehringer, D., Erzberger, J.P., Schaefer, T. and Ban, N., 2015. Structure of a yeast 40S-eIF1-eIF1A-eIF3-eIF3j initiation complex. Nat. Struct. Mol. Biol. 22, 269-271.

Bai, X.-C., McMullan, G. and Scheres, S.H.W., 2015. How cryo-EM is revolutionizing structural biology. Trends Biochem. Sci. 40, 49-57.

Bettadapura, R., Rasheed, M., Vollrath, A. and Baja, C., 2015. PF2fit: Polar Fast Fourier matched alignment of atomistic structures with 3D electron microscopy maps. PLoS Comput. Biol. 11, e1004289.

Budkevich, T.V., Giesebrecht, J., Behrmann, E., Loerke, J., Ramrath, D.J.F., Mielke, T., Ismer, J., Hildebrand, P.W., Tung, C.-S., Nierhaus, K.H. Sanbonmatsu, K.Y. and Spahn, C.M.T., 2014. Regulation of the mammalian elongation cycle by subunit rolling: a eukaryotic-specific ribosome rearrangement. Cell 158, 121-131.

Chacón, P. and Wriggers, W., 2002. Multi-resolution contour-based fitting of macromolecular structures. J. Mol. Biol. 317, 375-384.

Cyganek, B. and Siebert, J.P., 2009. An introduction to 3D computer vision techniques and algorithms. John Wiley & Sons, Ltd., Chicester.

Derevyanko, G. and Grudinin, S., 2014. HermiteFit: fast-fitting atomic structures into a low-resolution density map using three-dimensional orthogonal Hermite functions. Acta Crystallogr. D. Biol. Crystallogr. 70, 2069-2084.

Esquivel-Rodriguez, J. and Kihara, D., 2013. Computational methods for constructing protein models from 3D electron microscopy maps. J. Struct. Biol. 184, 93-102.

Farabella, I., Vasishtan, D., Joseph, A.P., Pandurangan, A.P., Sahota, H. and Topf, M., 2015. TEMPy: a Python library for assessment of three-dimensional electron microscopy fits. J. Appl. Crystallogr. 48, 1314-1323.

Fisher, R.A., 1921. On the "probable error" of a coefficient of correlation deduced from a small sample. Metron 1, 1-32.

Garzón, J.I., Kovacs, J., Abagyan, R. and Chacón, P., 2007. ADP_EM: fast exhaustive multi-resolution docking for high-throughput coverage. Bioinformatics 23, 427-433.

Gutmanas, A., Alhroub, Y., Battle, G.M., Berrisford, J.M., Bochet, E., Conroy, M.J., Dana, J.M., Fernandez Montecelo, A., van Ginkel, G., Gore, S.P., Haslam, P., Hatherley, R., Hendrickx, P.M.S., Hirshberg, M., Lagerstedt, I., Mir, S., Mukhopadhyay, A., Oldfield, T.J., Patwardhan, A., Rinaldi, L., Sahni, G., Sanz-Garcia, E., Sen, S., Slowley, R.A., Velankar, S., Wainwright, M.E. and Kleywegt, G.J., 2014. PDBe: Protein Data Bank in Europe. Nucleic Acids Res. 42, D285-D291.

Hoang, T.V, Xavier, C. and Ritchie, D.W., 2013. gEMfitter: A highly parallel FFT-based 3D density fitting tool with GPU texture memory acceleration. J. Struct. Biol. 184, 348-354.

Hrabe, T., Chen, Y., Pfeffer, S., Cuellar, L.K., Mangold, A.-V. and Foerster, F., 2012. PyTom: a python-based toolbox for localization of macromolecules in cryo-electron tomograms and subtomogram analysis. J. Struct. Biol. 178, 177-188.

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. and Chiu, W., 2011. EMDatabank.org: unified data resource for Cryo-EM. Nucleic Acids Res. 39, D456-D464.

Von Loeffelholz, O., Jiang, Q., Ariosa, A., Karuppasamy, M., Huard, K., Berger, I., Shan, S.-o. and Schaffitzel, C., 2015. Ribosome-SRP-FtsY cotranslational targeting complex I the closed state. Proc. Natl. Acad. Sci. USA 112, 3943-3948.

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.

Roseman, A.M., 2000. Docking structures of domains into maps from cryo-electron microscopy using local correlation. Acta Crystallogr. D. Biol. Crystallogr. 56, 1332-1340.

Svidritskly, E., Brilot, A.F., Koh, C.S. Grigorieff, N. and Korostelev, A.A., 2014. Structures of yeast 80S ribosome-tRNA complexes in the rotated and nonrotated conformations. Structure 22, 1210-1218.

Volkmann, N. and Hanein, D., 1999. Quantitative fitting of atomic models into observed densities derived by electron microscopy. J. Struct. Biol. 125, 176-184.

Volkmann, N., 2009. Confidence intervals for fitting of atomic models into low-resolution densities. Acta Crystallogr. D Biol. Crystallogr. 65, 679-689.

Weisstein, E.W. MathWorld – A Wolfram Web Resource. June-July 2015. Web. 11 June 2015. http://mathworld.wolfram.com/Convolution.html.

Wu, X., Milne, J.L.S., Borgnia, M.J., Rostapshov, A.V., Subramaniam, S. and Brooks, B.R., 2003. A core-weighted fitting method for docking atomic structures into low-resolution maps: application to cryo-electron microscopy. J. Struct. Biol. 141, 63-76.

Van Zundert, G.C.P. and Bonvin A.M.J.J., 2015. Fast and sensitive rigid-body fitting into cryo-EM density maps with PowerFit. AIMS Biophysics 2, 73-87.

Van Zundert, G.C.P. and Bonvin A.M.J.J., 2015. Integrative modeling of biomolecular complexes: HADDOCKing with cryo-electron microscopy data. Structure 23, 949-960.