Europe PMC Funders Group
Author Manuscript
Nat Methods. Author manuscript; available in PMC 2016 June 17.

Published in final edited form as:
Nat Methods. 2012 September ; 9(9): 853–854. doi:10.1038/nmeth.2115.

Prevention of overfitting in cryo-EM structure determination

Sjors H.W. Scheres and **Shaoxia Chen**
MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 0QH, UK

In the field of single-particle analysis of electron cryo-microscopy (cryo-EM) data, a growing concern that some resolution claims might not be substantiated by the data has been one of the instigators of community-wide efforts to develop new validation tools[1]. A known issue with commonly used cryo-EM structure determination procedures is their liability to overfit the data. Most procedures counter overfitting by low-pass filtering, but the effective frequencies for these filters are often based on suboptimal Fourier Shell Correlation[2] (FSC) procedures. In the suboptimal procedure, FSC curves are calculated between reconstructions from two halves of the data, while a single model is used to determine the relative orientations of all particles. It is well known that bias towards noise in the single model may inflate the resulting resolution estimates. To illustrate this, we applied the suboptimal procedure to a simulated cryo-EM data set of 20,212 GroEL particles. Whereas the reported resolution was 4.6 Å, the true resolution of the map was only 7.8 Å. Also the presence of expected density features in the map does not necessarily provide sufficient evidence for a resolution claim: we could make convincingly looking figures of apparent side-chain density that in reality corresponded to overfitted noise (Supplementary Figure 1). Consequently, overfitting may remain undetected and interpretation of cryo-EM maps may be subject to errors.

The dangers of overfitting have been recognized, and refinement procedures with resolution-dependent weighting schemes to reduce overfitting have been proposed[3,4]. However, two known solutions to prevent it are not in common use. By refining two models independently (one for each half of the data), so-called gold-standard[1] FSC curves may be calculated that are free from spurious correlations. Alternatively, the data used for the orientation determination may be limited to a user-specified frequency, so that model bias beyond that frequency may be avoided. However, the argument that withholding part of the data from the refinement would substantially deteriorate the orientations and thereby the quality of the structure has prevented the wide-spread use of either of these solutions. In what follows, we prove this thesis to be false.

Analysis of simulated data with realistic signal-to-noise ratios (SNRs) indicates that the accuracy of the orientation determination is not affected by the exclusion of high-frequency terms, nor by the use of a model that is reconstructed from only half of the particles (Supplementary Figure 2). These simulations illustrate that only the low-medium frequency terms in the individual particles contain sufficiently high SNRs to contribute significantly to the orientation determination, which is in good agreement with experimental evidence that

scheres@mrc-lmb.cam.ac.uk

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

cryo-EM particles may be aligned accurately using only low-frequency data[5]. Because in most cryo-EM studies the low-medium frequencies of reconstructions from half of the particles are not expected to be significantly worse than those of reconstructions from all particles, we hypothesize that overfitting may be prevented without a notable loss of resolution using either frequency-limited refinement or refinement based on gold-standard FSCs. Since the former involves a decision by the user, i.e. choosing the frequency at which to limit the refinement, we favour gold-standard FSCs and implemented a procedure to independently refine two models as a script on top of the conventional projection matching protocol in the XMIPP package[6] (Supplementary Figure 3 & Supplementary Software).

We tested our hypothesis using three cryo-EM data sets: 5,053 GroEL particles that are distributed by the National Center for Macromolecular Imaging; an in-house collected data set of 50,330 β-galactosidase particles (Supplementary Methods); and 5,403 hepatitis B capsid particles from a previously published study[7]. High-resolution crystal structures are available for all three data sets, and these were used to assess the “true” resolution obtained using refinements based on either gold-standard or conventional FSC procedures (Figure 1). For all three cases, the conventional procedure reported apparently better FSC curves than the gold-standard procedure, but in no case did the gold-standard procedure actually result in a lower resolution map compared to the crystal structure. On the contrary, for the β-galactosidase data the gold-standard procedure yielded a structure that correlated up to higher frequencies with the crystal structure than the conventional procedure, which suffered from severe overfitting and gave rise to strong artefacts in the map. We also note that, in the absence of overfitting, the frequency at which the gold-standard FSC drops below 0.143 is a good indicator of the true resolution of the map (Supplementary Table 1), which is as expected from theory[8]. Finally, in the limit of very small data sets, division of the data into two halves might affect resolution. However, calculations with subsets of the GroEL particles suggest that this only becomes an issue for data sets that are much smaller than those typically used in cryo-EM reconstructions (Supplementary Figure 4).

The principal conclusion is therefore that overfitting of noise using suboptimal FSCs causes worse orientations and leads to a worse structure. In contrast, the use of gold-standard FSCs provides a realistic estimate of the true signal, which ultimately leads to a better map. The procedures proposed here are straightforward to implement in existing programs, and their application will eradicate the hazards of overfitting from cryo-EM structure determination procedures.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We are grateful to Tony Crowther and Richard Henderson for helpful discussions, and to Jake Grimmett for help with computing. Tony Crowther provided hepatitis B data, and the National Center for Macromolecular Imaging, which is funded by National Institutes of Health grant P41RR02250, provided GroEL data. This work was funded by the UK Medical Research Council (MRC) through grant MC_UP_A025_1013 to SHWS.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

References

[1]. Henderson R, et al. Structure. 2012; 20:205–214. [PubMed: 22325770]
[2]. Saxton WO, Baumeister W. J. Microscopy. 1982; 127:127–138. [PubMed: 7120365]
[3]. Stewart A, Grigorieff N. Ultramicroscopy. 2004; 102:67–84. [PubMed: 15556702]
[4]. Scheres SH. J. Mol. Biol. 2012; 415:406–418. [PubMed: 22100448]
[5]. Henderson R, et al. J. Mol. Biol. 2011; 413:1028–1046. [PubMed: 21939668]
[6]. Scheres SH, et al. Nat. Protoc. 2008; 3:977–90. [PubMed: 18536645]
[7]. Bottcher B, Wynne SA, Crowther RA. Nature. 1997; 386:88–91. [PubMed: 9052786]
[8]. Rosenthal PB, Henderson R. J. Mol. Biol. 2003; 333:721–745. [PubMed: 14568533]

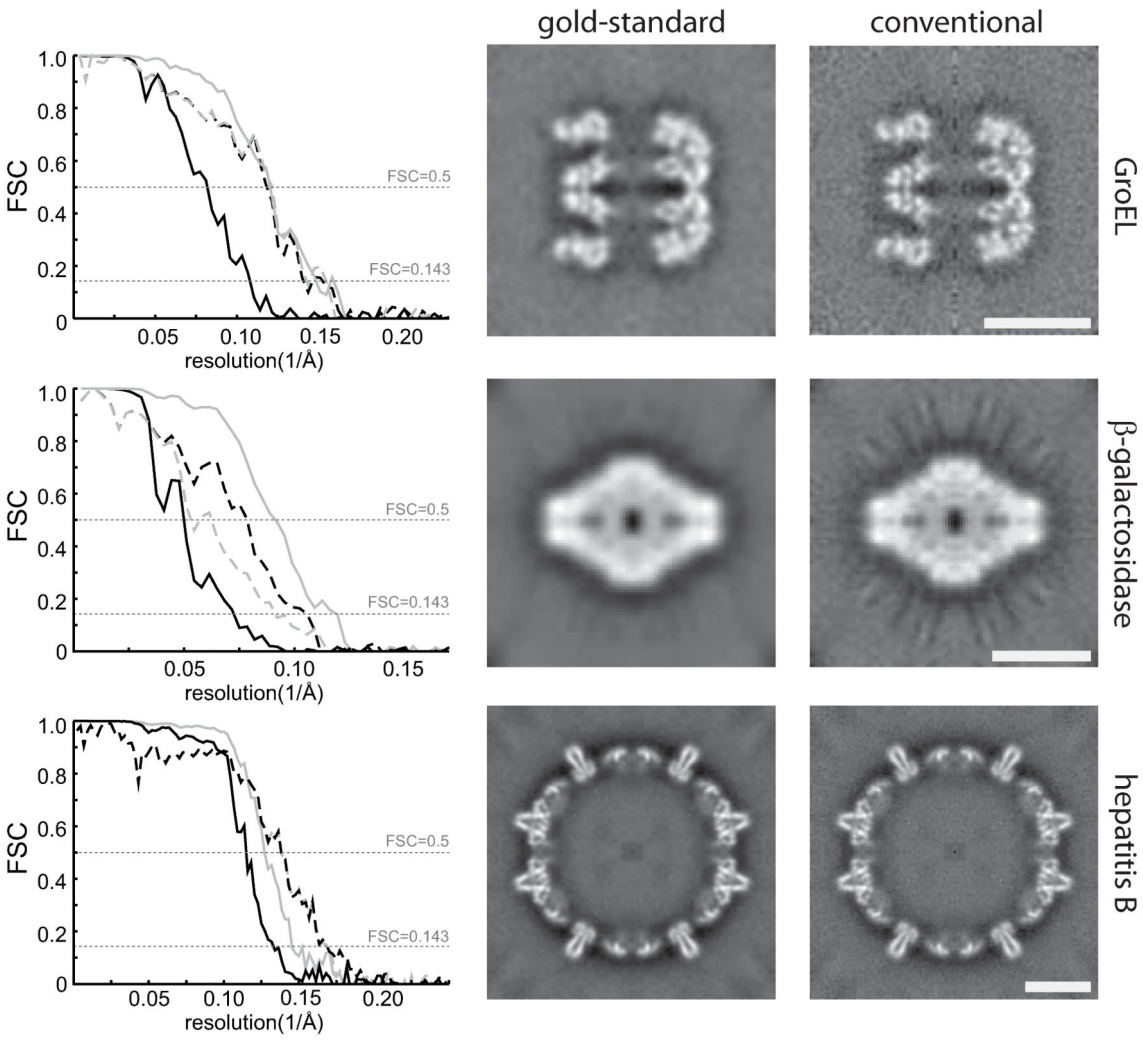

Figure 1. The prevention of overfitting

Tests with three cryo-EM data sets (GroEL, β-galactosidase and hepatitis B) illustrate that overfitting may be avoided without compromising resolution. FSCs between reconstructions from random halves of the data (solid lines) and FSCs between the reconstruction from all particles and the crystal structures (dashed lines) show that the gold-standard procedure (black) does not yield lower resolutions, while the conventional procedure (grey) over-estimates resolution (left column). The frequency where the dashed lines pass through FSC=0.5 indicate the true resolution of the reconstructions from all particles, whereas the frequency where the solid lines pass through FSC=0.143 indicate the reported resolution[8]. Corresponding values are given in Supplementary Table 1. Central slices through the reconstructions for the gold-standard procedure show less noisy maps with fewer artefacts (middle column) than for the conventional procedure (right column). Scale bars (white) correspond to 100 Å.

Europe PMC Funders Author Manuscripts
Europe PMC Funders Author Manuscripts