Journal of Structural Biology 204 (2018) 276–282

Contents lists available at ScienceDirect

Journal of Structural Biology

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

This article is part of the Special Issue on the 2016 CryoEM Challenges

Assessing the quality of single particle reconstructions by atomic model building

Joshua H. Mendez[a], Scott M. Stagg[b,c,*]

[a] Department of Physics, 77 Chieftan Way, Tallahassee, FL 32306, United States
[b] Institute of Molecular Biophysics, 91 Chieftain Way, Florida State University, Tallahassee, FL 32306, United States
[c] Department of Chemistry and Biochemistry, 95 Chieftain Way, Florida State University, Tallahassee, FL 32306, United States

ARTICLE INFO

Keywords:
cryo-EM map assessment
FSC
Map challenge
Best practices

ABSTRACT

The 2015/2016 Map Challenge challenged cryo-EM practitioners to process a series of publicly available cryo-EM datasets. As part of the challenge, metrics needed to be developed to assess and compare the quality of the different map submissions. The most common metric for assessing maps is determining the resolution by Fourier shell correlation (FSC), but there are well known instances where the resolution can be misleading. In this manuscript, we present a new approach for assessing the quality of a map by determining the map "modelability" rather than on resolution. We used the automated map tracing and modeling algorithms in Rosetta to generate populations of models, and then compared the populations between different map entries by the Rosetta score, RMSD to a reference model provided by the map challenge, and by pair-wise RMSDs between different models in the population. These metrics were used to determine statistically significant rankings for the map challengers for each dataset. The rankings revealed inconsistencies between the resolution by FSC, emphasized the importance of the interplay between number of particles contributing to a map and map quality, and revealed the importance of software familiarity on single particle reconstruction results. However, because multiple variables changed between map entries, it was challenging to derive best practices from the map challenge results.

1. Introduction

The 3DEM map challenge invited "challengers" to process several public cryo-EM datasets with the hopes of 1) establishing a benchmark set of datasets suitable for high resolution cryoEM, 2) encouraging developers and users of 3DEM software packages to analyze these datasets and come up with best practices, 3) evolve criteria for evaluation and validation of the results of the reconstruction and analysis, and 4) compare and contrast the various reconstruction approaches to achieve high efficiency and accuracy. Along with such a map challenge comes a need for metrics for comparing and assessing the resolution and quality of the submitted maps. The most common metric for assessing cryo-EM maps is the Fourier shell correlation method (FSC) (reviewed in (Sorzano et al., 2017)) where the particles contributing to a single particle reconstruction are split into two halves, reconstructed, and then the reconstructions compared in increasing frequency shells in Fourier space. The resolution for the overall reconstruction is determined as the frequency the Fourier shell correlation falls below some threshold, and many different cutoffs have been proposed including 0.5 (Harauz and van Heel, 1986) 0.143 (Rosenthal and Henderson, 2003), and a moving threshold based on information theory (van Heel and Schatz, 2005). Using FSC to assess map challengers is problematic because FSC doesn't necessarily measure map quality. For instance, it is possible that for any two maps, one might have a higher $FSC_{0.5}$ but a lower $FSC_{0.143}$. The map with higher $FSC_{0.5}$ may be higher quality and biological interpretability but be lower resolution by the $FSC_{0.143}$ metric. Another challenge with FSC-based assessment is that FSC values can be artificially inflated by overly aggressive masking or overfitting due to the alignment of noise during the single particle refinement (Scheres and Chen, 2012; Sousa and Grigorieff, 2007). Other metrics for determining resolution include local resolution determination (Kucukelbir et al., 2014) and Fourier neighbor correlation (Sousa and Grigorieff, 2007), but these methods can also be influenced by masking and/or correlated noise.

Here we developed an approach for comparing the quality of different reconstructions that does not rely on the FSC. The idea behind our analysis is that the cryo-EM map quality metrics that have been developed so-far are mostly directed at assessing the resolution,

* Corresponding author at: Institute of Molecular Biophysics, 91 Chieftain Way, Florida State University, Tallahassee, FL 32306, United States.
E-mail address: sstagg@fsu.edu (S.M. Stagg).

https://doi.org/10.1016/j.jsb.2018.09.004
Received 21 March 2018; Received in revised form 5 September 2018; Accepted 8 September 2018
Available online 10 September 2018

however the driving force for determining structures is what biological interpretations can be derived from the map. Most of the maps in the map challenge were determined to high enough resolution to build atomic models, so our metric for model quality was not resolution, but rather, how well atomic models could be built from the maps. We call this the 'modelability' of a map. In our approach, we generated families of models automatically using the EM modeling tools developed in Rosetta (DiMaio et al., 2015). We then used two metrics for assessing map quality; how closely the atomic models derived from the map match the published atomic structure and are scored in Rosetta, and how closely the atomic models in the family of models resemble each other. This is similar to a method that was proposed for assessing models built from EM maps (Herzik et al., 2017), but that method depended on a model being already available instead of building models *ab initio* from the maps like we are doing here.

Our modelability metrics proved quite powerful for comparing and ranking different maps. For instance, even when maps reported similar and/or identical resolutions by FSC-based metrics, our method was capable of discerning subtle but significant differences between maps. Most significantly, our analyses showed the maps that were the most modelable, and thus presumably the most interpretable, were not necessarily the ones with the highest resolutions by FSC. The purpose of this manuscript is to 1) describe our method and validate it, 2) report our rankings for the different map challengers, and 3) report on whether any best practices could be discerned from the results.

2. Approach

The Map Challenge 2015/2016 from EMDataBank was composed of seven targets, each with submissions from different researchers. The goal of the assessment stage was to devise a protocol capable of assessing and validating maps. Our protocol is driven by how much can be interpreted from the map, what we call "modelabillity", rather than the traditional approach of FSC resolution. Since we are focusing on modelability, we chose to use the sharpened maps that the challengers deposited for their submissions. Our view was that these represent the maps that would be interpreted biologically and presumably represent the best efforts of the challenger to produce a highly detailed map. We note that the sharpening could certainly influence the quality of the modeling, but we viewed our role as assessors to focus only on what the challengers submitted, so we processed the user-submitted sharpened maps.

The protocol for automated model building was divided in three major steps: initial model building, loop extension & refinement, and model comparison. The first step, initial model building, starts by aligning each map from a specific target. This was done by loading each map into Chimera (Petterson et al., 2004) for an initial visual inspection. The map which visually appeared to be the best quality was selected as the reference map. Subsequently, each map was automatically aligned using the Chimera tool *fit in map*. Modeling the entire atomic structures of all maps would be computationally untenable, so only representative segments of the maps were modeled. In order to segment the same region from all maps, the reference pdb model that was provided as part of the Map Challenge was loaded and aligned to the maps using the same procedure as the maps, first manual and then using *fit in map*. The next step was to identify a section of the map that would be a good representation of the entire map. For the symmetrical specimens, the extracted region represented one asymmetric unit. For the ribosome, we chose the protein uL15 to be representative of the entire map quality because it had regions on the outside of the ribosome and loops that extended into the interior. Once the representative segment was identified for the individual maps, the segment was extracted and was low pass filtered until the surface was smooth with no discernable features. Then the region was expanded by five shells with Gaussian decay and converted to a mask using *e2proc3d* from the software package EMAN2 (Tang et al., 2007). This mask was then used for the extraction of the selected region for each map without introducing hard edges or artifacts in the extracted map. Next, an initial atomic model was created using from the extracted segment using the default parameters of the Rosetta protocol denovo_density and the amino acid sequence. If the initial round of denovo_density did not cover seventy percent of the extracted map more rounds were performed using the previous initial model as the input for the new round until enough coverage was achieved or no improvement in the initial model was observed; in our case we had a limit of 5 extra rounds. For the GroEL maps, none of the maps produced realistic models, so this target was dropped from further analysis.

For the second step, the amino acid sequence, the extracted map, and the initial atomic model were used as inputs for the Rosetta function rosetta_scripts with the default parameters for loop extension and refinement. During this process, Rosetta completed the coverage of the extracted map while optimizing the structure simultaneously. We used this strategy to generate two thousand atomic models for each map in an attempt to populate as many conformations as were allowed by the given map.

In the third and final step, models were scored by two different metrics called Combined Score and Internal root mean square (RMSD) that assessed how consistently Rosetta was able to generate models from the given EM maps. The Internal RMSD evaluation was generated by calculating the pair-wise RMSDs between each generated atomic model for a given map, while the Combined Score was generated by dividing each atomic model's Rosetta energy score by its RMSD to the reference atomic model provided by the Map Challenge. The aim with the Internal RMSD was to score entries using no other information than how well the given map constrained the automatic modeling. The Combined Score combined two external pieces of information to generate a score: the RMSD to the Map Challenge reference structure and the Rosetta energy score. Assuming the Map Challenge reference structure is the ground truth, then models with a low RMSD would be the best, but it is possible that the reference structures were not necessarily the best that could be generated from the cryo-EM data and that Rosetta might do a better job. Presumably the best structure would have the lowest Rosetta energy, but then again this depends on the parameterization of the Rosetta force field and the relative weights of the map (Fig. 1A). Thus, the Combined Score combines both pieces of data by dividing a model's Rosetta score by its RMSD to the reference model. Assuming Rosetta was able to completely and correctly model a given map, then the Internal RMSD should be the best representation of the quality of a map and should agree with the Combined Score, but in cases where Rosetta was unable to completely and accurately model a map or was systematically off in its modeling, then the Combined Score is more reliable. For this reason, we ranked maps based on both scores. The score distributions were plotted using box and whisker plots to visualize the distribution of values (Fig. 1B,C). Good quality maps have a mean Internal RMSD value near zero and a Combined Score as negative as possible while maintaining the smallest possible spread (Fig. 1).

Finally, the mean Internal RMSD and Combined Scores were used to generate challenger rankings for each Map Challenge dataset (Fig. 1D). In order to test whether the differences between the means were statistically significant, Kruskal-Wallis one-way Analysis of Variance (Kruskal-Wallis) test (Kruskal and Wallis, 1952) and Dunn's Pairwise Comparison (Dunn's test) (Dunn, 1961) were performed. In the Kruskal-Wallis test, if the null hypothesis is rejected, at least one of the map's mean is different from the rest. To identify which differences were statistically significant, a post-hoc Dunn's test was used. Dunn's test is a pairwise comparison that identifies if the mean difference of a pair is significant at a specific p-value. In our case, all targets showed to have at least one map to be significant, which led us to proceed with the Dunn's test with a p-value of 0.05.

The full results and rankings for each map challenge sample, Rosetta score vs. RMSD plots, Combined Score distributions, Internal RMSD