Journal of Structural Biology 204 (2018) 351–359

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

De novo main-chain modeling with MAINMAST in 2015/2016 EM Model Challenge

Genki Terashi[a], Daisuke Kihara[a,b,*]

[a] Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA
[b] Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA

ARTICLE INFO

Keywords:
Cryo-EM
Electron microscopy
Protein structure modeling
CryoEM Model Challenge
Main-chain trace
Map interpretation
MAINMAST
Mean shifting algorithm
Minimum spanning tree
Rosetta
confidence score

ABSTRACT

Protein tertiary structure modeling is a critical step for the interpretation of three dimensional (3D) election microscopy density. Our group participated the 2015/2016 EM Model Challenge using the MAINMAST software for a de novo main chain modeling. The software generates local dense points using the mean shifting algorithm, and connects them into Cα models by calculating the minimum spanning tree and the longest path. Subsequently, full atom structure models are generated, which are subject to structural refinement. Here, we summarize the qualities of our submitted models and examine successful and unsuccessful models, including 3D models we did not submit to the Challenge. Our protocol using the MAINMAST software was sometimes able to build correct conformations with 3.4–5.1 Å RMSD. Unsuccessful models had failure of chain traces, however, their Cα positions and some local structures were quite correctly built. For evaluate the quality of the models, the MAINMAST software provides a confidence score for each Cα position from the consensus of top 100 scoring models.

1. Introduction

Recent technical improvements in cryo-electron microscopy (cryo-EM) have led to a rapid increase in macromolecular structures determined by cryo-EM (Frank, 2017), particularly those determined at a near atomic resolution (e.g. 4 Å or better). The statistics at EMDB (Patwardhan, 2017; Velankar et al., 2016) show that EM maps at 4 Å or better represent the fastest growing category among five resolution levels shown in the statistics (4, 6, 8, 10, 15 Å or worse) (https://www.ebi.ac.uk/pdbe/emdb/statistics_num_res.html/). From 2014 to 2017 this high resolution portion of the deposited maps in the EMDB increased its share of the total database by 92%, rising from 5.3% to 10.2%, nearly doubling in that time.

When an EM map is obtained, structure modeling of biomolecules, including proteins and nucleotides, in the map is a critical step for interpreting the map density. Various structure modeling techniques have been developed which are designed for maps of certain resolution ranges (Esquivel-Rodriguez and Kihara, 2013). Types of structure modeling tools include those used for atomic structure building originally developed for X-ray crystallography (Terwilliger et al., 2008), identifying main-chain conformations in a map (Baker et al., 2012a; Chen et al., 2016; Frenz et al., 2017; Wang et al., 2015), refining structure models (Afonine et al., 2018; DiMaio et al., 2009; DiMaio et al., 2015; Kirmizialtin et al., 2015; Trabuco et al., 2008), fitting known structures to density maps (Esquivel-Rodriguez and Kihara, 2012; Lopez-Blanco and Chacon, 2013; Miyashita et al., 2017; Woetzel et al., 2011; Wriggers and Birmanns, 2001), and identifying local structures in medium resolution (e.g. 6–10 Å) maps (Baker et al., 2007; Jiang et al., 2001). Although structure modeling tools have been improving to keep pace with the rapid progress in microscopy instrumentation on 3D map reconstruction techniques (Hohn et al., 2007; Punjani et al., 2017; Scheres, 2012; Tang et al., 2007), modeling tools still have substantial room for improvement.

To critically evaluate 3D map construction and protein structure modeling techniques, EMDataBank is hosting community-wide challenges for the EM community. Following the first challenge meeting in 2010 (Ludtke et al., 2012), EMDataBank hosted two challenges in 2015/2016, the Map Challenge and the Model Challenge, for evaluating and discussing protocols and results for single particle reconstructions and for methods and results of building protein structure models, respectively. In the Model Challenge, submitted models were evaluated in one of the four modeling categories: 1) optimization of the current models; 2) fitting of known structures to maps; 3) ab initio model building; and 4) other types. Our group participated in the third category, ab initio model building. The Model Challenge consisted of eight target macromolecules with maps of a reported resolution ranging from

* Corresponding author at: Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA.
E-mail address: dkihara@purdue.edu (D. Kihara).

https://doi.org/10.1016/j.jsb.2018.07.013
Received 27 February 2018; Received in revised form 13 July 2018; Accepted 19 July 2018
Available online 31 July 2018

2.2 to 4.5 Å. The targets were released on October 14, 2015 and the deadline for the model submission was on June 17, 2016. The subsequent evaluation meeting was held on October 6-8 2017 at Stanford University, California, USA.

Our group has submitted ten models each for four targets using a de novo main-chain tracing software, MAINMAST (MAIN chain Model trAcing from Spanning Tree), developed by us (Terashi and Kihara, 2018). Compared to other existing de novo modeling software (Chen et al., 2016; DiMaio et al., 2015; Frenz et al., 2017), MAINMAST is unique in that it does not refer to known structures, generates and ranks multiple structure models, and provides confidence levels of each residue positions by examining consensus among generated models. The modeling procedure using MAINMAST is fully automated and requires no manual parameter tuning or human intervention.

Here, we summarize and analyze the quality of the structure models of four target maps we submitted to the 2015/2016 Model Challenge. In addition to the submitted models, we also discuss models that were built for the other four target maps but not submitted to the Model Challenge. In addition to the protocol we used in 2016, we compare different components of structure refinements. At the end of this report, we also show the confidence score of predicted models, which correlated well with the accuracy of their Cα positions.

2. Materials and methods

2.1. Model Challenge targets

Eight EM maps from EMDB were specified as targets in the 2015/2016 Modelling Challenge (http://challenges.emdatabank.org/?q=model-challenge-targets) (Table 1). As indicated, the target EM maps were published in the literature and were released at EMDB with fitted structures by the authors. Although fitted structures by authors were available, we modeled protein structures only from the density maps and did not refer to the author-fitted structures during the modeling since we participated in the ab initio modeling category to test our software, MAINMAST. However, as a preprocessing of maps before applying MAINMAST, we segmented EM maps according to the fitted structures in each map so that a map region only include a single chain. This process was needed since the current version of MAINMAST assumes that there is only a single protein chain in a map. For each density map, a single subunit (the A chain) was manually segmented from a whole density map using UCSF Chimera's "zone tool" using the PDB structure as the reference.

2.2. The modeling protocol using MAINMAST

MAINMAST is a de novo main-chain structure modeling program for EM maps with resolutions of approximately 4–5 Å or better (Terashi and Kihara, 2018). Refer to the original paper for details of the algorithm. MAINMAST directly traces local dense regions of a map and does not refer to any known structures or structural fragments. MAINMAST consists of five steps (Fig. 1). In the first step, MAINMAST identifies local dense points (LDPs) in a density map using the mean shifting algorithm (Fukunaga and Hostetler, 1975). The implicit assumption is that a density observed in a map is the sum of Gaussian density functions that originate from atoms in the map. The density k of a position that originates from a grid point locating at a distance of d is defined as $k(d) = \exp\left(-1.5\|\frac{d}{\sigma}\|^2\right)$, where σ is set to 1.0. The total density of a position is the sum of the Gaussian-weighted densities from neighboring grid points. The mean shift algorithm starts from a set of grid points in the map that have a density value above a threshold value and iteratively move them toward local maxima until convergence is reached. The purpose of using mean shift is to perform local clustering to identify representative dense points. The number of LDPs is usually much more than the number of residues in the target protein. Typically, the number of clusters is about 40% of the number of heavy atoms of the underlined protein in the map.

In the second step, a minimum spanning tree (MST) is constructed that connects all LDPs. MST is a graph structure that connects all vertices with the minimal total weight of edges without forming cycles. It was found that the main-chain of the protein is well covered by the MST because the number of points is large enough so that neighboring points

Table 1
Summary of the models for 2015/2016 Modeling Challenge target maps.

Target	EMDB-ID[a]	Res. (Å)	PDB[b]	Model[c]	RMSD (Å)[d]	GDT-TS[e]	Unlabeled RMSD[f]	Recall d < 2/3Å[g]	Precision d < 2/3Å[h]
T0001	2842	3.3	4udv-A	1st	11.7	17.0	1.6	0.83/0.97	0.84/0.96
Tabacco Mosaic Virus				top10	11.4	19.6	1.6	0.84/0.97	0.85/0.98
T0002	5623	3.3	3j9i-A	1st	5.1	46.2	1.6	0.79/0.97	0.80/0.98
T20S Proteasome				top10	3.7	58.8	1.5	0.85/0.99	0.86/0.98
T0004	5778	3.3	3j5p-A (3j9j-A)	1st	9.2 (9.1)	15.6 (30.0)	2.1 (2.2)	0.34/0.48 (0.68/0.93)	0.64/0.87 (0.68/0.88)
TrpV1 Channel				top10	8.5 (8.3)	18.6 (36.0)	2.0 (2.1)	0.36/0.50 (0.68/0.94)	0.69/0.90 (0.68/0.89)
T0005	6000	3.8	3j7l-A	1st	3.4	60.7	1.6	0.79/0.99	0.79/0.97
Bromo Mosaic Virus				top10	3.4	60.7	1.6	0.81/0.99	0.81/0.99
T0006	5995	3.2	3j7h-A	1st	12.4	50.6	1.6	0.80/0.93	0.80/0.96
β-Galactosidase				top10	11.7	52.7	1.5	0.84/0.94	0.85/0.96
T0006	2984	2.2	5a1a-A	1st	29.9	2.6	2.1	0.64/0.81	0.65/0.87
β-Galactosidase				top10	27.7	3.6	2.0	0.66/0.82	0.67/0.90
T0007	2677	4.5	4upc-A	1st	27.2	4.5	2.7	0.38/0.67	0.39/0.74
γ-Secretase				top10	23.9	5.6	2.6	0.38/0.70	0.39/0.75
T0007	3061	3.4	5a63-A	1st	11.4	17.7	1.7	0.77/0.96	0.77/0.96
γ-Secretase				top10	8.8	37.3	1.6	0.82/0.96	0.82/0.96
Average				1st	13.3	27.2	1.9	0.67/0.86	0.71/0.91
				top10	11.9	32.5	1.8	0.70/0.87	0.73/0.92

[a] Density maps whose models were submitted to the official assessment are underlined.
[b] The reference PDB structure, against which models were compared. We only modeled the A chain of each complex. For EMD-5778, the map was segmented based on 3j9j-A and models were assessed with two reference structures, 3j5p-A, which was provided at the Map Challenge website, and 3j9j-A. See text for details. Residue numbers in MAINMAST models were renumbered based on the reference PDB structures when assessed.
[c] 1st, the top scoring model; top10, the best GDT-TS model among top 10 scoring models. All models were ranked by the Rosetta Free Energy.
[d] The RMSD of Cα atoms modelled by MAINMAST and the reference structure.
[e] Global distance test total score. The value ranges from 0 to 100 with 100 as the best score.
[f] The RMSD between nearest Cα atoms of MAINMAST and reference structure.
[g] The fraction of Cα atoms in the reference structure which are closer than a threshold distance (2.0 or 3.0 Å) to any Cα atoms in the model.
[h] The fraction of Cα atoms in the model which are closer than a threshold distance (2.0 or 3.0 Å) to any Cα atoms in the reference structure.