# Cylindrical Similarity Measurement for Helices in Medium-Resolution Cryo-Electron Microscopy Density Maps

Salim Sazzed[1], Peter Scheible[1], Maytha Alshammari[1], Willy Wriggers[2], Jing He[1]
[1]*Department of Computer Science, Old Dominion University, Norfolk, VA*
[2]*Department of Mechanical and Aerospace Engineering, Old Dominion University, Norfolk, VA*

## Abstract

Cryo-electron microscopy (cryo-EM) density maps at medium resolution (5-10 Å) reveal secondary structural features such as α-helices and β-sheets. However, they lack the side chain details that would enable a direct structure determination. Among the more than 800 entries in the Electron Microscopy Data Bank (EMDB) of medium-resolution density maps that are associated with atomic models, a wide variety of similarities exist between maps and models. To validate such atomic models and to classify structural features, a local similarity criterion, the $F_1$ score, is proposed and evaluated in this study. The $F_1$ score is normalized to a range from zero to one, providing a local measure of cylindrical agreement between the density and atomic model of a helix. A systematic scan of 30,994 helices (among 3,247 protein chains modeled into medium-resolution density maps) reveals a range of observed $F_1$ scores from 0.171 to 0.848. This range of $F_1$ scores suggests that the local similarity is quantified and differentiated as intended. The best (highest) $F_1$ scores tend to be associated with regions that exhibit high and spatially homogeneous local resolution (between 5 Å to 7.5 Å) in the helical density. The proposed $F_1$ scores can be used as a discriminative classifier for validation studies and as a ranking criterion for cryo-EM density features in databases.

## Introduction

The number of atomic structures derived from cryo-electron microscopy (cryo-EM) density maps has increased rapidly over the last five years. As of December 1, 2019, there were 824 structures modeled from cryo-EM density maps in the medium-resolution range (5-10 Å) compared to 3,144 structures derived from resolutions better than 5 Å. The quality of a map produced by an experimental cryo-EM laboratory improves as more data is collected. Consequently, medium-resolution maps are routinely created in the initial stages of a cryo-EM imaging project, before the specimen preparation protocols are tuned to perfection. Lower-resolution regions can also be present in overall high-resolution maps due to conformational flexibility or libration of the specimen. For these reasons, medium-resolution maps are often the first and only observations available for a new system. Early glimpses of an unknown system can be of significant biological importance, and there is pressure to interpret them at atomic detail. Understandably, investigators will attempt to build models despite the risks and inaccuracies. Our present work is concerned with assessing the accuracy of the model fit given the increasing number of deposited map model pairs in the challenging medium-resolution range. It remains quite difficult to model atomic structures de novo for most proteins. However, in many cases, secondary structure elements, such as α-helices and β-sheets, can be assigned with confidence in medium-resolution maps. Among secondary structural features, α-helices often appear as cylindrically shaped density regions, and β-sheets appear as thin layers of density in medium-resolution cryo-EM maps.

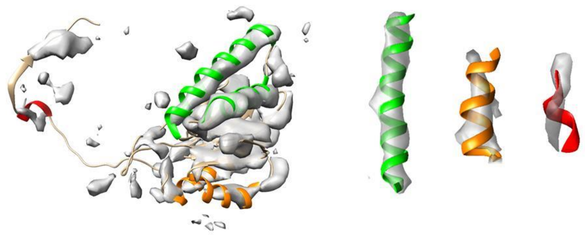

**Figure 1.** Illustration of different levels of map/model similarity exhibited by helices in the same map. The surface representation of the density map (EMDB ID 4089, gray, corresponding to Chain 2) is superimposed on Protein Data Bank (PDB) ID 5ln3 Chain 2 (ribbon). Helices with different levels of similarity are indicated by green, orange, and red, respectively, from strong similarity to poor similarity.

## Method

We downloaded 654 medium-resolution cryo-EM density maps with corresponding atomic models from the EMDB along with their atomic models. For a protein with multiple copies of the same chain sequence, only one was included to eliminate redundancies. The final data set consisted of 3,247 protein chains. The cryo-EM density region corresponding to each chain was extracted from the entire density map using UCSF Chimera. Because the method was designed to measure helices, chains without helices were excluded. Chains that lay entirely outside any molecular density were also excluded.

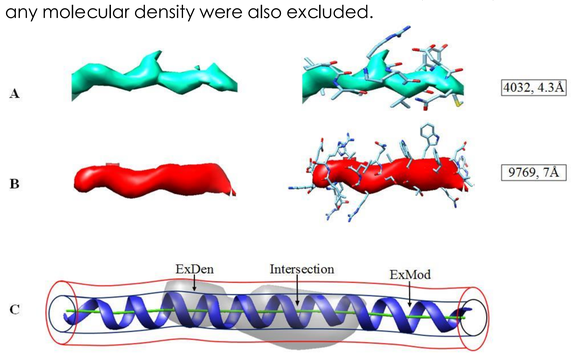


**Figure 2.** Examples of helix densities at different resolutions and their cylindrical similarity: (A) density for a helix in EMDB ID 4032 with 4.3 Å resolution superimposed on the atomic model; (B) density for a helix in EMDB ID 9769 with 7 Å resolution superimposed on the atomic model; (C) two template cylinders of 2.5 Å and 4 Å radii, respectively, were used to measure the cylindrical similarity (see section "Cylindrical similarity score"). Figure 2C shows the intersection of map and model, and two mismatch regions, ExDen and ExMod

The density distribution of a helix closely resembles a cylinder at medium resolution, with the highest densities found near the central axis of the helix. The similarity of an atomic model was quantified using two suitably chosen template cylinders, derived from the central line of a helix using Ca atoms (Figure 2C). The central line was produced using the AxisComparison tool from an atomic model in PDB format. Every four consecutive Ca atoms of a helix are averaged to generate initial central points, which are interpolated to produce a smooth line (Figure 2C). The radius was selected as 2.5 Å and 4 Å for the inner and outer cylinders, respectively, to approximate the radius of the helix backbone and a typical radial size of an α-helix. At each density threshold, the number of helix density voxels within the inner cylinder, *VxInner*, measured the volume of the intersection between the helix density and the model (Figure 2C). The number of helix density voxels between the inner and outer cylinder, *VxOut*, measured the volume of identifiable helix voxels outside of the helix backbone model (denoted ExDen in Figure 2C).

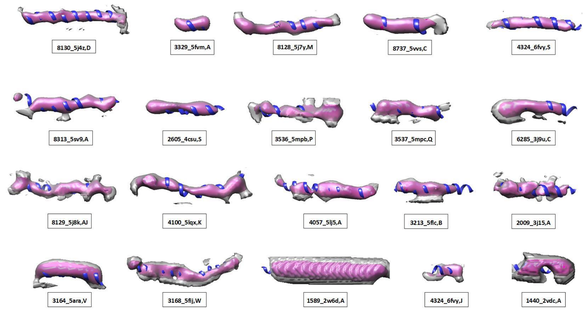


**Figure 3.** Twenty examples of helical map/model pairs with $F_1$ similarity scores. The atomic structure of a helix is superimposed on the density extracted using 4 Å radius and 7 Å radius, respectively, around the central axis of the helix. Each helix density is displayed using the threshold that optimizes the $F_1$ score. Panels are sorted by the $F_1$ score from top left to lower right.

$$F_1 = \frac{2 * Pden * Rmod}{Pden + Rmod} \quad (1)$$

$$Pden = \frac{Intersection}{Intersection + ExDen} = \frac{VxInnner}{VxInner + VxOut} \quad (2)$$

$$Rmod = \frac{Intersection}{Intersection + ExMod} = \frac{VxInner}{VxInner + \overline{VxInner}} \quad (3)$$

The $F_1$ score is a metric commonly used in machine learning. Our adaptation of the $F_1$ score (Eqn. 1) was adapted from the standard interpretation in statistics, where the $F_1$ score is the harmonic mean of precision and recall. the $F_1$ score in our implementation measured similarity between the density region above a threshold in the vicinity of a helix and the region of a helix backbone model. *Pden* (Eqn. 2) represents the accuracy of the helix density, i.e., the percentage of agreed map/model volume among the total volume of map relevant to the helix. *Rmod* (Eqn. 3) represents the accuracy of the model, i.e., the percentage of agreed map/model volume among the total volume relevant to the helix backbone model.

## Results

When the resolution is high, the spiral of the helix backbone starts to become visible in cryo-EM density maps (Figure 2A): the helix backbone exhibits a higher density than the side-chains. The higher spiral-shaped density of the helix backbone required a density threshold for visualization that obscured the expected side chain regions (Figure 2A). At medium resolution, however, instead of the spiral of the backbone, only a cylindrically shaped density is observed (Figure 2B): the highest density voxels are often located near the central axis of the cylinder. The inner radius was designed to capture the backbone density of a helix: a helix with good map/model similarity is expected to have a density threshold at which the density is primarily located within the inner cylinder. The cylindrical similarity of helices was evaluated for 30,994 helices from 3,247 protein chains corresponding to maps with resolutions between 5 Å to 10 Å. The observed $F_1$ scores varied considerably, suggesting an excellent discriminative ability of the measure. The $F_1$ score measures the cylindrical similarity between the density and the model at a helix region. This is reflected in the results, showing that the highest scoring densities, were cylindrical in shape and associated with higher local resolution that is also spatially homogeneous. Those with lower $F_1$ scores deviated from a cylinder shape and were mostly associated with lower or spatially fluctuating local resolution. $F_1$ scores collectively compare the density with the model of a helix. Local resolution varies from voxel to voxel, and for poor similarity map/model pairs, we observed that there could be as much as a 5 Å difference in local resolution within the same helix region.

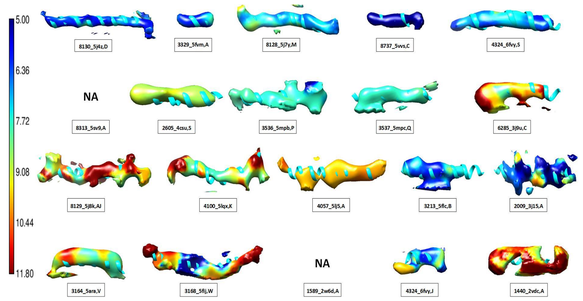


**Figure 4.** Local resolution of eighteen helix regions. Density regions near helices were extracted using a cylinder of 7 Å in radius from the central axis of the corresponding helix model (ribbon). Local resolutions produced using MonoRes were used to color the density according to the resolution bar. The EMDB ID, PDB ID, and chain ID are provided for each helix. The threshold that maximizes the $F_1$ score from top left to lower right.

## Acknowledgements

The work in this poster was supported, in part, by NSF DBI-1356621 and NIH R01-GM062968.