Matching of EM Map Segments to Structurally-Relevant Bio-molecular Regions

Manuel Zumbado-Corrales[1,2], Luis Castillo-Valverde[1], José Salas-Bonilla[1], Julio Víquez-Murillo[1], Daisuke Kihara[3], and Juan Esquivel-Rodríguez[1](✉)

[1] Instituto Tecnológico de Costa Rica, Escuela de Computación, Campus Cartago, Cartago, Costa Rica
manzumbado@ic-itcr.ac.cr, jesquivel@tec.ac.cr

[2] Advanced Computing Laboratory, National High Technology Center, San José, Costa Rica

[3] Department of Biological Sciences/Department of Computer Science, Purdue University, West Lafayette, IN, USA
dkihara@purdue.edu

Abstract. Electron microscopy is a technique used to determine the structure of bio-molecular machines via three-dimensional images (called maps). The state-of-the-art is able to determine structures at resolutions that allow us to identify up to secondary structural features, in some cases, but it is not widespread. Furthermore, because molecular interactions often require atomic-level details to be understood, it is still necessary to complement current maps with techniques that provide finer-grain structural details. We applied segmentation techniques to maps in the Electron Microscopy Data Bank (EMDB), the standard community repository for these data. We assessed the potential of these algorithms to match functionally relevant regions in their atomic-resolution image counterparts by comparing against three protein systems, each with multiple atomic-detailed domains. We found that at least 80% of amino acid residues in 7 out of 12 domains were assigned to single segments, suggesting there is potential to match the lower resolution segmented regions to the atomic counterparts. We also qualitatively analyzed the potential on other EMDB structures, as well as generating the raw segmentation information for the complete EMDB, for interested researchers to use. Results can be accessed online and the library developed is provided as part of an open-source project.

Keywords: Computational biology · Computational protein structures · Electron microscopy · 3DEM · Segmentation

1 Introduction

Structural biology has seen enormous progress in the 21st century, particularly with the rise of open databases that host three-dimensional models of biomolecular structures. On one hand, we have the Protein Data Bank (PDB)

J. L. Crespo-Mariño and E. Meneses-Rojas (Eds.): CARLA 2019, CCIS 1087, pp. 464–478, 2020.
https://doi.org/10.1007/978-3-030-41005-6_32

[7] that hosts over 150,000 atomic-detailed structures of proteins, DNA and RNA. Most of the structures deposited in PDB correspond to relatively small bio-molecular complexes. A second database, the Electron Microscopy Data Bank (EMDB) [16], focuses on three-dimensional models created from electron microscopy (3DEM), which can power the imaging of larger macro-molecular complexes that have been historically deposited in the PDB. Very significant structures have been identified thanks to 3DEM [19,20,32].

Because protein interactions actually happen at the atomic level, ideally we want EM maps to give us atomic details so that we can do functional analysis by just using this type of image. In [24] the authors were able to generate a reconstruction with a resolution of 3.5 Å that allowed them to create an all-atom model. Even when there have been steady improvements on attainable resolutions over the years, this level of detail is not widespread. A gamut of computational techniques are often applied to be able to obtain details that go beyond the density envelope that EM maps provide. Hybrid approaches have been used to extract finer-grain details out of EM maps up to 10 Å [18]. Techniques like these have been applied to shed light into the organization of proteomes, for instance [6]. The field of Electron Microscopy fitting deals with finding atomic-level details based on existing high-resolution structures that match EM maps [8,10,28].

Even if we are not able to identify all atoms in a map, other structural elements and annotations can also be useful, for functional analysis purposes. For example, the architecture and helical regions of 26S proteasome were determined this way in [5]. Annotations directly on density maps have shown previously unknown interactions in complexes [11]. A significant number of algorithms and tools have been developed to identify secondary structure elements [2,3,12–14]. More recently, de novo modeling of proteins has also been applied to EM maps [26].

Segmentation is another technique used to identify structural features in maps. The basic notion here is to divide EM maps into density regions that should match individual protein structures, or functionally relevant sections, like domains. Some automated techniques that assume the knowledge of the components, or the symmetry of the complexes have been previously developed [4,27,33]. Atomic models are not always available for the maps under study and we also need to deal with the added complexity of images in more complex, environments, that can lead to lower resolution images [21].

In this work, we study the potential to identify functionally-relevant regions in 3D Electron Microscopy maps by applying automated segmentation. Our goal is not only to approximate near-atomic features but, more in general, to identify structural hot-spots within maps that can later be mapped to larger images. Through the open-source library we have developed for this work, we aim to provide a way to both visually and analytically study EM maps. We apply these techniques to all the structures currently in the EMDB and show sample cases that highlight the potential of this type of method. As noted in [21], trying to bridge the gap between cellular and molecular structural data is key for the field to advance.

2 Methods

2.1 Watershed Segmentation

Our segmentation method uses the immersive watershed transform to generate region labels as a first step, then we perform region grouping with scale-space filtering as proposed in [22]. This approach is useful to reduce over-segmentation as reported by authors.

The watershed algorithm can be understood following the same analogy presented in [31]. Consider EM map densities as a topographic surface as seen in Fig. 1, where holes are pierced at surface local minima to let water flood basins. If each voxel located in a catchment basin would merge with water coming from different local minima, a dam is built to separate water from different regions. At the end, each resulting flooded region is separated by built dams, also called watershed lines, which coincide with surface local maxima.

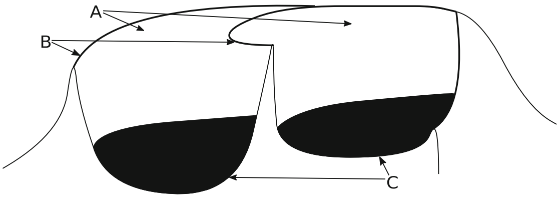


Fig. 1. Representation of watershed process with A as catchment basins, B as watershed lines and C as local minima (conceptual illustration inspired by [31]).

We take the additive inverse of an EM map as the topographic surface, regarding higher densities as surface local minima. Thus we get watershed regions surrounding higher density locations, separated by lowest surface densities. A fixed connectivity of 26 voxels is used in each dimension to connect neighbors in the process of assign adjoin voxels to the same region.

2.2 Scale-Space Grouping

Region grouping is performed by progressively smoothing the EM map using a Gaussian filter. This concept was introduced in [35] and is called scale-space filtering. Scale-space representation $L(x, y, z; \sigma) \in \mathbb{R}^3 \times \mathbb{R}^+$ of an EM map $f(x, y, z) \in \mathbb{R}^3$ is defined scale-space representation $L(x, y, z; \sigma) \in \mathbb{R}^3 \times$ scale-space representation $L(x, y, z; \sigma) \in \mathbb{R}^3 \times \mathbb{R}^+$ of an EM map $f(x, y, z) \in \mathbb{R}^3$

is defined as and is called scale-space filtering. The scale-space representation $L(x, y, z; \sigma) \in \mathbb{R}^3 \times \mathbb{R}^+$ of an EM maa$f(x, y, z) \in \mathbb{R}^3$ is defined as

$$L(x, y, z; \sigma) = (f * g)(x, y, z; \sigma), \tag{1}$$

where $\sigma \in \mathbb{R}^+$ controls the variance of the Gaussian kernel $g(x, y, z; \sigma) \in \mathbb{R}^3 \times \mathbb{R}^+$, defined as

$$g(x, y, z; \sigma) = \frac{1}{\sigma^3 (2\pi)^{\frac{3}{2}}} \exp\left(-\frac{x^2 + y^2 + z^2}{2\sigma^2}\right). \tag{2}$$

In order to group regions, an initial local maxima point set is computed from original EM map. Then, each initial local maxima point is successively moved up to the local maxima of a subsequent smoothed scale corresponding to the steepest ascent in terms of density intensity, as shown in Algorithm 1.

The process of Scale-Space filtering produces an EM map for each step with progressive attenuation of energy on higher density locations. Thus, computed local maxima of a succeeding step in the Scale-Space representation would replace several local maxima of a current step. After N number of smoothing steps, resulting local maxima points having the same coordinates in space would merge into a new region.

Parameters used for segmentation and grouping follow the same approach presented in [22]. The number of steps N controls how many steps of Scale-Space grouping are performed. Smoothing step size S regulates how much smoothing is achieved at each step. A density threshold level defines the structure contour to be segmented and also affects the isosurface generated by the Marching Cubes algorithm.

Algorithm 1. Space-scale grouping of watershed regions of segmented EM map

```
Input   : Watershed segmented map
Input   : Collection of successively smoothed maps
Input   : Steps
Output: Grouped regions
$N \leftarrow$ Steps;
$M \leftarrow$ Watershed segmented EM map;
$S \leftarrow$ Collection of successively smoothed maps;
$L \leftarrow$ Collection of local maxima of $M$;
for $i$ in $N$ :
    for $p$ in $L$ :
        $B \leftarrow$ Collection of local maxima of $S$ for corresponding $i$;
        Replace $p$ with the steepest local maxima in $B$ respect to $p$;
Find duplicates in $L$ and merge corresponding regions into new one;
```

2.3 Marching Cubes and Isosurface Generation

Marching cubes is a reference algorithm for isosurface reconstruction from sampling data. Several optimizations have been proposed to extend the basic approach, improve its performance and solve ambiguities. Our method relies on an efficient implementation of Marching Cubes algorithm proposed in [17]. Isosurface visualization of protein structures is essential to later identify segments enclosed in the three dimensional space of an EM map.

2.4 Library Design

The created library is composed of the following Python modules: `processing`, `visualizer`, `reader` and `molecule`. The `processing` module object contains watershed and space-scale implementations. Later, the `visualizer` module object implements main methods exposed to the user, namely, `segmentation`, `show` and `show_atom_matching`. The `reader` module object implements `read` function to read map files from disk and returns created `map` object.

Our library supports GPU accelerated visualization by using Glumpy [25] which is a fast and scalable open source library that takes advantage of the computational power of GPU through OpenGL.

In this work we show the effectiveness of open source scientific Python libraries such as Scikit-image, Biopandas and Numpy [23,30,34]. At the same time, we identify potential areas of improvement that will allow us, in the future, to augment them with custom features to scale up our system.

2.5 Validation Data Set

In order to determine the potential to identify structurally relevant regions through segmentation, we used three protein systems from a data set previously identified as suitable for the analysis of algorithms related Electron Microscopy map fitting [1]. The data set focuses on proteins for which we have both low-resolution EM maps but also there is an atomic level Protein Data Bank structure that matches the map. While we are not directly tackling the EM-fitting problem in this study, the data set is still very much valid for our purposes. In particular, the fact that the authors have divided each of the protein systems into regions, using PDB structures, allows us to compare the segments that our library generates with the annotated domains. Intuitively, if each of the segments that we generate has a high overlap with the domains identified in that study, then the structural correspondence that we propose is valid. Table 1 summarizes the characteristics of the data set.

In addition to testing against these controlled protein systems we have also run the segmentation over two larger macro molecular structures to illustrate how promising the methods are at identifying not only structural regions within isolated proteins, but also in large complexes. For this purpose we have analyzed **EMDB ID: 1048** and **EMDB ID:2596**.

Table 1. Validation data set metadata

EMDB ID	PDB ID	Units	Residues	Description
1010	1GQE	4	362	Release Factor (RF2)
1364	1FNM	5	655	Elongation Factor G (EFG)
5017	1N0U	3	654	Elongation Factor 2 (EF2)

3 Results

Our method validation is based on the comparison between computational and biological segments. While the computational ones are obtained applying the methods described in Sect. 2, the biological segments are more difficult to come by. As we have described in Sect. 2.5, we have used a previously derived definition of domains in a protein. In general, domains are regions within a protein that have been conserved through evolution *for a good reason*, be that structural, functional, etc. Our premise is that segmentation algorithms that are able to closely predict the matching between computational and biological segments can allow us to better understand the different sections in a macro-molecule.

3.1 Atomic-Detailed Validation

Figure 2 shows the structural baseline for our detailed analysis. The wireframe representation shows the density envelope identified using the author-recommended contour level to create isosurfaces that resemble the true volume of the protein. In bright green we can see the ribbon representation of the protein backbone, which is important to determine the rough high density regions that should be expected to be identified. However, the knowledge of the backbone does not tell us on its own what biological sections we are supposed to target. For that, we fitted each of the domains (as found in [1]) using a method developed by the authors that uses Markov Random Fields to generate candidate alignments[1].

The fitted structures, shown in separate colors for each domain become our validation targets. We assessed what fraction of the residues were assigned to different segments, per domain. The theoretical ideal result is for every residue to be assigned to a single segment. We tackled this problem from both quantitative and qualitative angles.

Quantitative Results. Table 2 shows our way of quantitatively determining how well the segments generated for **EMDB ID: 1010** matched the domains. In this particular case, the results mostly meet our expectations. Two of the four

[1] This method is based on the combination of physico-chemical, shape and cross correlation features between each of the domains and the EM map. This work is not part of a stand-alone article as of this writing.

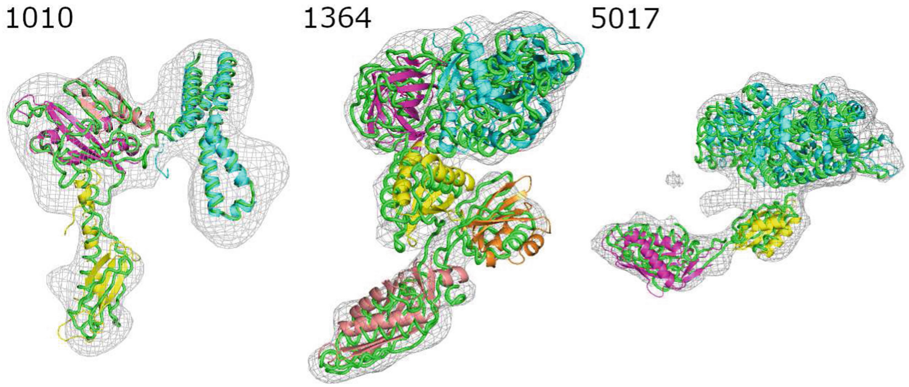


Fig. 2. EM maps in the data set aligned to the C-α trace (bright green) and a candidate fitting of the domains in each protein system (individually colored). Each label corresponds to the EMDB ID for each map. (Color figure online)

domains, **B** and **D**, are matched to a single segment, as well as 94.12% of residues in **A**. **C** has a slightly worse result since 16.48% of residues are not assigned to a single segment, but it can still be considered promising[2]. The drawback with **EMDB ID: 1010**'s results is that we should have identified 4 segments, as opposed to 3. That suggests that there is some density noise that we cannot overcome that yields two regions that should be separate to become a single one.

Table 2. Segment matches for EMDB ID 1010. The **All** row summarizes the overall assignment. The remaining rows show the per-domain assignment

Domain	Segment	Percentage (within domain)	Residues Assigned
All	1	31.22%	113/362
	3	46.96%	170/362
	2	21.82%	79/362
A	1	94.12%	112/119
	3	5.88%	7/119
	3	100.00%	99/99
C	3	16.48%	15/91
	2	83.52%	76/91
D	3	100.00%	45/45

[2] Note that for 1010 there are 8 missing residues, observed in the C-α trace but not the PDB with all atomic details. They are ommitted for analysis purposes.

Similarly, Table 3 summarizes the matches found for **EMDB ID: 1364**. In this case we can claim successful results for domains **A**, **D** and **E**, since they were mostly assigned to a single segment (82.87%, 95.52% and 98.61%, respectively). However, domains **B** and **C** are more evenly distributed across multiple segments, which is not the desirable outcome. As we will see in our qualitative analysis, there is a region where densities are more difficult to differentiate. We can also observe that we are identifying one less segment than we should. There are 5 domains in this protein but we are only generating 4. This can also explain the difficulty in assigning clear-cut segments.

Table 3. Segment matches for EMDB ID 1364. The **All** row summarizes the overall assignment. The remaining rows show the per-domain assignment

Domain	Segment	Percentage (within domain)	Residues Assigned
All	4	11.45%	75/655
	3	16.95%	111/655
	1	49.92%	327/655
	2	21.68%	142/655
A	4	1.20%	3/251
	3	15.94%	40/251
	1	82.87%	208/251
B	4	60.50%	72/119
	3	39.50%	47/119
C	3	30.38%	24/79
	1	53.16%	42/79
	2	16.46%	13/79
D	1	4.48%	6/134
	2	95.52%	128/134
E	1	98.61%	71/72
	2	1.39%	1/72

The last case analyzed was **EMDB ID: 5017**. As Table 4 reflects, this was the most challenging case from a quantitative point of view. The best match obtained corresponded to domain **C** with 64.79%, but **A** and **B** are generally split between two segments. On the flip side, this case correctly identified that 3 segments were needed to have a correct matching of domains. We will discuss in the qualitative analysis why this protein structure could have behaved this way.

Qualitative Results. The previous section had the purpose of providing a non-subjective metric that would shed light in terms of whether or not a large portion of residues were assigned to expected segments. We can argue that just looking at proportions is not enough to determine how good the assignment was.

Table 4. Segment matches for EMDB ID 5017. The **All** row summarizes the overall assignment. The remaining rows show the per-domain assignment

Domain	Segment	Percentage (within domain)	Residues Assigned
All	2	46.48%	304/654
	1	32.42%	212/654
	3	21.10%	138/654
A	2	44.35%	204/460
	1	40.65%	187/460
	3	15.00%	69/460
B	2	43.90%	54/123
	3	56.10%	69/123
C	2	64.79%	46/71
	1	35.21%	25/71

As we have stated in this work, the actual 3D structure of proteins is crucial to determine how well they function. Thus, a presumably good match of 80%+ that misses the key 20% of a protein is not necessarily the best result.

To complement the quantitative arguments made before, Fig. 3 shows the colored assignment of EM map regions to segments, made by our library. We contrast this against the fitted structures shown in Fig. 2. Based on the results obtained in the quantitative analysis, we assessed three elements. First, are the domains with majority single-segment assignments consistent with the expected structure? Second, are there clues as to why the algorithm identified one fewer segment for **EMDB ID: 1010** and **EMDB ID: 1364**? Finally, for the domains with unclear assignments, is there any structural reason that may explain them?

The general 3D structure of **EMDB ID: 1010** from Fig. 2 can be summarized as two separate domains on the left (yellow) and right (cyan) and two others that are tightly coupled between them (purple and orange). From that point of view, it is not unexpected that the algorithm identified only 3 segments, assuming that the main difficulty was separating the link domains. If we look at the segmentation from Fig. 3 we see that the overall left and right domains are captured by the blue and yellow segments. It appears as if the orange domain (in Fig. 2) corresponds roughly to the red segment in Fig. 3, which is encouraging. We do see that all segments over extend, which could be an artifact of the space scale filtering applied. We need to remember that the surfaces here are based on contour values that are suitable to convey the actual shape of the proteins, but the EM maps contain density in surrounding voxels too and there is no guarantee that at the contour level we used there is no noise. The two parts in the red segment are particularly interesting when compared to the fitted structure. The EM map, at the recommended contour, shows a gap not filled by the C-α trace in Fig. 2 which could back the idea that we're dealing with a noisy region.

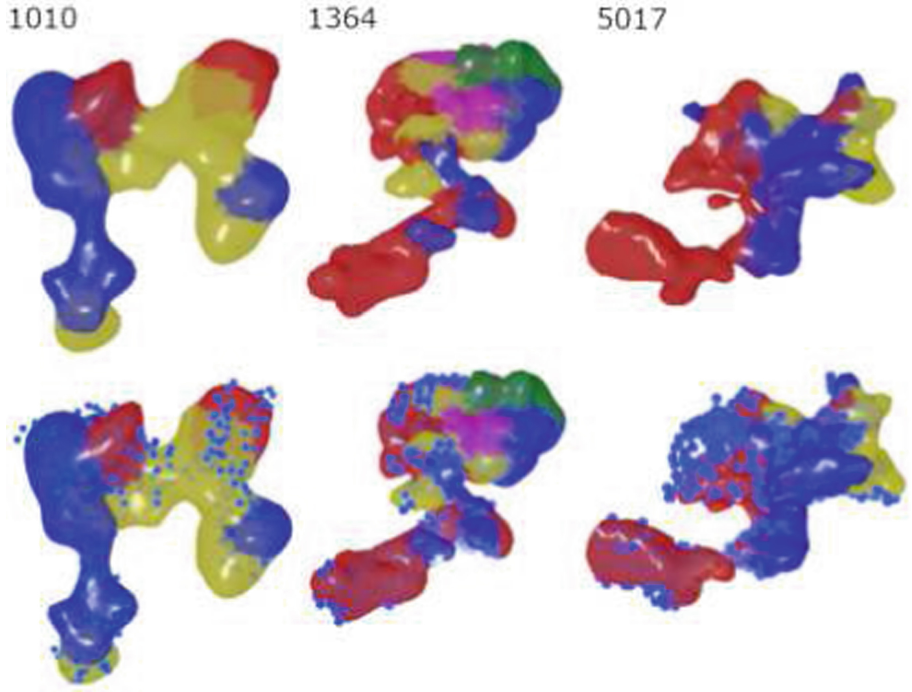


Fig. 3. Segmentation applied to EMDB ID 1010, 1364 and 5017 as detailed in Sect. 2. Every color represents an individual segment identified. The top row shows only the segments, for clarity, while the bottom row adds spheres to highlight the C-α atoms. Those atoms are expected to be slightly shifted due to small adjustments done to contour thresholds in the segmentation (Color figure online)

For **EMDB ID: 1364**, Fig. 2 shows a big domain on the top right corner of the structure that is segmented into multiple ones (as opposed to just a single one). This particular problem is less troublesome than some of the aspects found for **EMDB ID: 1010**. Refinement of segment assignments that are *supposed to be one* can be performed as a post-processing step. On a more critical note, there are 2 red segments in Fig. 3, but it is possible that the top one should have been colored yellow. Making that change should have mostly captured the structure, starting from the bottom of a red domain, followed by blue and then yellow (with some over extension of the red segment, though). Even though this case shows better metrics than **EMDB ID: 5017**, discussed below, it is arguably the most challenging structurally.

Finally, in the case of **EMDB ID: 5017** the overall coloring of the lower segments is not incompatible with the purple and yellow domains, in the fitted structure. We can argue that the lower left section should indeed have been colored red, and the lower center section should have been all blue, albeit with higher precision required to differentiate where the red section finished and the blue started. The main issue in this complex though comes from domain **A**, which is significantly larger than the other two. As it was the case for **EMDB ID: 1364**, the problem here is that a single domain was broken up into multiple segments. Post-processing could solve this in a later iteration of our algorithm. For the purposes of this study, we tried multiple thresholds for the parameters that could be tuned and the results were similar, in every case. Note that, as it was the case for **EMDB ID: 1010**, there is a region in the wire frame that does not correspond to our reference C-α trace, which could also be a factor in the less accurate segmentation.

There are two key takeaways from our qualitative analysis. First, even though we applied space scale filtering, that did not solve all the problems related to integrating multiple segments into one, when that was expected. Second, there is clear over extension of some segments into small areas that they should not, and it could be due to noise spreading from one region to the other. Even with these two areas to improve that we identified, the results are generally good. The segmentation of these types of density maps could generate a very large number of segments, which makes it very difficult to then map domains of the size that we are testing in this study. Furthermore, there are regions in each map where there is clear correspondence between both fitted structures and segments, which shows the promise of the approach.

3.2 Large Macro-molecule Segmentation

The three protein systems studied are useful for detailed analysis because there is atomic-level information throughout the structure. The more complex macromolecules do not necessarily have that type of information available in databases, to serve as a larger scale evaluation target.

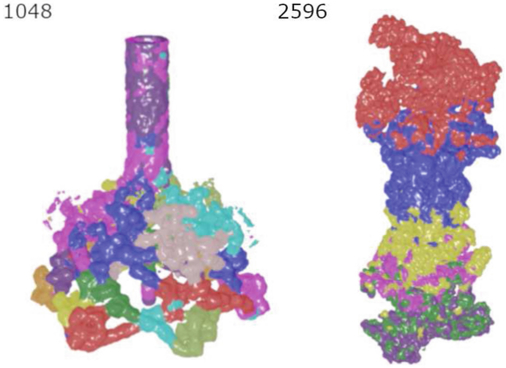
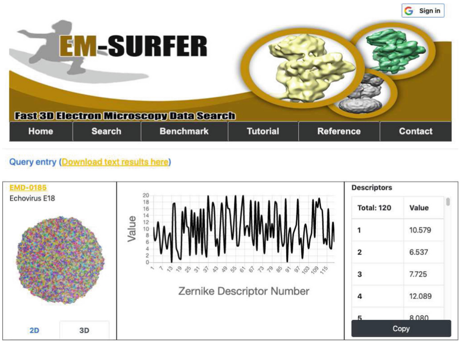


Fig. 4. Segmentation applied to EMDB ID 1048, 2596 as detailed in Sect. 2. Every color represents an individual segment identified. EMDB ID 1048 is an image of bacteriophage T4 baseplate while EMDB ID 2596 is a 26S proteasome structure (Color figure online)



Total: 120	Value
1	10.579
2	6.537
3	7.725
4	12.089
5	8.080

Fig. 5. Sample segmentation result from alpha release of EM-SURFER (http://emsurfer.tecdatalab.org/result/0185). The 3D section shows images generated by our library.

Even though we cannot provide rigorous analysis about the quality of the segmentation applied to large-scale macro-molecules, we applied the algorithm to two sample systems that are both interesting biologically but also have much larger scale. Figure 4 shows the segmentation results for **EMDB ID: 1048** and **EMDB ID: 2596**. The former is the structure of bateriophage T4 baseplate, which is a virus that infects *Escherichia coli* [15]. This structure is in the range of hundreds of nanometers. The latter structure, a 26S proteasome, is in charge of breaking down proteins [29].

The results obtained are sensible and resemble some fitted results referenced in the EM Data Bank. This path towards the validation of segmentation for larger structures is one that we want to explore further in the future.

3.3 Online Results

We have generated segmentation results for maps in the Electron Microscopy Data Bank, which can be accessed as part of an alpha release of the latest version of EM-SURFER [9], an EM map search engine that relies on the fast comparison of structural features. Figure 5 shows a screen shot of a sample result generated for **EMDB ID: 0185**[3].

4 Conclusions

In this work we have shown the potential to match biological domains to computationally derived segments using watershed segmentation with space-scale

[3] The production version of EM-SURFER is hosted at http://kiharalab.org/em-surfer. An example result from our alpha release of the latest version, that includes segmentation results, can be accessed at is available at http://emsurfer.tecdatalab.org/result/0185.

grouping. Our methods represent a valid approach to elucidate what regions in an EM map correspond to relevant regions in proteins. We have first evaluated this by analyzing three protein systems in detail, where we have both the atomic-details and the EM maps, which allowed us to do a thorough validation. We have also evaluated much larger macro-molecular structures to assess the potential to apply our methods to large scale problems.

As discussed in the Results section, we have identified areas where results can be refined. Those revolve mainly around the decision to integrate or break apart density clusters, but not to an extent that diminishes the positive results obtained.

As part of our work, we offer the community a library that is accessible as an open source project, which contains both the algorithms and visualization features to reproduce our results (github.com/tecdatalab/biostructure). Furthermore, we publish our segmentation results online through a new version of EM-SURFER.

Acknowledgements. Funded by the *Vicerrrectoría de Investigación y Extensión* at *Instituto Tecnológico de Costa Rica.*

References

1. Ahmed, A., Whitford, P.C., Sanbonmatsu, K.Y., Tama, F.: Consensus among flexible fitting approaches improves the interpretation of cryo-EM data. J. Struct. Biol. **177**(2), 561–570 (2012). https://doi.org/10.1016/j.jsb.2011.10.002
2. Baker, M.L., Baker, M.R., Hryc, C.F., Ju, T., Chiu, W.: Gorgon and pathwalking: macromolecular modeling tools for subnanometer resolution density maps. Biopolymers **97**(9), 655–668 (2012). https://doi.org/10.1002/bip.22065
3. Baker, M.L., Ju, T., Chiu, W.: Identification of secondary structure elements in intermediate-resolution density maps. Structure **15**(1), 7–19 (2007). https://doi.org/10.1016/j.str.2006.11.008
4. Baker, M.L., Yu, Z., Chiu, W., Bajaj, C.: Automated segmentation of molecular subunits in electron cryomicroscopy density maps. J. Struct. Biol. **156**(3), 432–441 (2006). https://doi.org/10.1016/j.jsb.2006.05.013
5. Beck, F., et al.: Near-atomic resolution structural model of the yeast 26S proteasome. Proc. Natl. Acad. Sci. U.S.A. **109**(37), 14870–14875 (2012). https://doi.org/10.1073/pnas.1213333109
6. Beck, M., et al.: Exploring the spatial and temporal organization of a cell's proteome. J. Struct. Biol. **173**(3), 483–496 (2011). https://doi.org/10.1016/j.jsb.2010.11.011
7. Burley, S.K., et al.: Protein data bank: the single global archive for 3D macromolecular structure data. Nucleic Acids Res. **47**(D1), D520–D528 (2019). https://doi.org/10.1093/nar/gky949
8. Dou, H., Burrows, D.W., Baker, M.L., Ju, T.: Flexible fitting of atomic models into cryo-EM density maps guided by helix correspondences. Biophys. J. **112**(12), 2479–2493 (2017). https://doi.org/10.1016/j.bpj.2017.04.054
9. Esquivel-Rodríguez, J., Xiong, Y., Han, X., Guang, S., Christoffer, C., Kihara, D.: Navigating 3D electron microscopy maps with EM-SURFER. BMC Bioinform. **16**, 181 (2015). https://doi.org/10.1186/s12859-015-0580-6

10. Fabiola, F., Chapman, M.S.: Fitting of high-resolution structures into electron microscopy reconstruction images. Structure **13**(3), 389–400 (2005). https://doi.org/10.1016/j.str.2005.01.007
11. Hryc, C.F., et al.: Accurate model annotation of a near-atomic resolution cryo-EM map. Proc. Natl. Acad. Sci. **114**(12), 3103–3108 (2017). https://doi.org/10.1073/PNAS.1621152114
12. Jiang, W., Baker, M.L., Ludtke, S.J., Chiu, W.: Bridging the information gap: computational tools for intermediate resolution structure interpretation. J. Mol. Biol. **308**(5), 1033–1044 (2001). https://doi.org/10.1006/jmbi.2001.4633
13. Kong, Y., Ma, J.: A structural-informatics approach for mining beta-sheets: locating sheets in intermediate-resolution density maps. J. Mol. Biol. **332**(2), 399–413 (2003)
14. Kong, Y., Zhang, X., Baker, T.S., Ma, J.: A structural-informatics approach for tracing beta-sheets: building pseudo-C(alpha) traces for beta-strands in intermediate-resolution density maps. J. Mol. Biol. **339**(1), 117–130 (2004). https://doi.org/10.1016/j.jmb.2004.03.038
15. Kostyuchenko, V.A., et al.: Three-dimensional structure of bacteriophage T4 baseplate. Nat. Struct. Biol. **10**(9), 688–693 (2003). https://doi.org/10.1038/nsb970
16. Lawson, C.L., et al.: EMDataBank unified data resource for 3DEM. Nucleic Acids Res. **44**(D1), D396–D403 (2016). https://doi.org/10.1093/nar/gkv1126
17. Lewiner, T., Lopes, H., Vieira, A.W., Tavares, G.: Efficient implementation of marching cubes' cases with topological guarantees. J. Graph.Tools **8**(2), 1–15 (2003). https://doi.org/10.1080/10867651.2003.10487582
18. Lindert, S., Stewart, P.L., Meiler, J.: Hybrid approaches: applying computational methods in cryo-electron microscopy. Curr. Opin. Struct. Biol. **19**(2), 218–225 (2009). https://doi.org/10.1016/j.sbi.2009.02.010
19. Ludtke, S.J., Chen, D.H., Song, J.L., Chuang, D.T., Chiu, W.: Seeing GroEL at 6 A resolution by single particle electron cryomicroscopy. Structure **12**(7), 1129–1136 (2004). https://doi.org/10.1016/j.str.2004.05.006
20. Mitra, K., et al.: Structure of the E. Coli protein-conducting channel bound to a translating ribosome. Nature **438**(7066), 318–324 (2005). https://doi.org/10.1038/nature04133
21. Patwardhan, A., et al.: Building bridges between cellular and molecular structural biology. eLife **6** (2017). https://doi.org/10.7554/eLife.25835
22. Pintilie, G.D., Zhang, J., Goddard, T.D., Chiu, W., Gossard, D.C.: Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. J. Struct. Biol. **170**(3), 427–438 (2010). https://doi.org/10.1016/j.jsb.2010.03.007
23. Raschka, S.: BioPandas: working with molecular structures in pandas dataframes. J. Open Source Softw. **2**(14) (2017). https://doi.org/10.21105/joss.00279
24. Roh, S.H., et al.: The 3.5-Å CryoEM structure of nanodisc-reconstituted yeast vacuolar ATPase Vo proton channel. Mol. Cell **69**(6), 993.e3–1004.e3 (2018). https://doi.org/10.1016/j.molcel.2018.02.006
25. Rougier, N.P.: Glumpy. In: EuroScipy (2015)
26. Terashi, G., Kihara, D.: De novo main-chain modeling with MAINMAST in 2015/2016 EM model challenge. J. Struct. Biol. **204**(2), 351–359 (2018). https://doi.org/10.1016/J.JSB.2018.07.013
27. Terwilliger, T.C., Adams, P.D., Afonine, P.V., Sobolev, O.V.: A fully automatic method yielding initial models from high-resolution cryo-electron microscopy maps. Nat. Methods **15**(11), 905–908 (2018). https://doi.org/10.1038/s41592-018-0173-1

28. Topf, M., Baker, M.L., John, B., Chiu, W., Sali, A.: Structural characterization of components of protein assemblies by comparative modeling and electron cryomicroscopy. J. Struct. Biol. **149**(2), 191–203 (2005). https://doi.org/10.1016/j.jsb.2004.11.004
29. Unverdorben, P., et al.: Deep classification of a large cryo-EM dataset defines the conformational landscape of the 26S proteasome. Proc. Natl. Acad. Sci. U.S.A. **111**(15), 5544–5549 (2014). https://doi.org/10.1073/pnas.1403409111
30. van der Walt, S., Colbert, S.C., Varoquaux, G.: The numpy array: a structure for efficient numerical computation. Comput. Sci. Eng. **13**(2), 22–30 (2011). https://doi.org/10.1109/MCSE.2011.37
31. Vincent, L., Soille, P.: Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Trans. Pattern Anal. Mach. Intell. **13**(6), 583–598 (1991). https://doi.org/10.1109/34.87344
32. Volkmann, N., Hanein, D., Ouyang, G., Trybus, K.M., DeRosier, D.J., Lowey, S.: Evidence for cleft closure in actomyosin upon ADP release. Nat. Struct. Biol. **7**(12), 1147–1155 (2000). https://doi.org/10.1038/82008
33. Volkmann, N.: A novel three-dimensional variant of the watershed transform for segmentation of electron density maps. J. Struct. Biol. **138**(1–2), 123–129 (2002). https://doi.org/10.1016/S1047-8477(02)00009-6
34. Van der Walt, S., et al.: Scikit-image: image processing in python. PeerJ **2**, e453 (2014)
35. Witkin, A.P.: Scale-space filtering. In: Readings in Computer Vision, pp. 329–332. Elsevier (1987). https://doi.org/10.1016/B978-0-08-051581-6.50036-2. https://linkinghub.elsevier.com/retrieve/pii/B9780080515816500362