Journal of Structural Biology xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Journal of Structural Biology

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

Automated tracing of helical assemblies from electron cryo-micrographs

Stefan T. Huber, Tanja Kuhm, Carsten Sachse*

European Molecular Biology Laboratory (EMBL), Structural and Computational Biology Unit, Meyerhofstraße 1, 69117 Heidelberg, Germany

ARTICLE INFO

Keywords:
Electron cryo-microscopy
Helical assemblies
Automated particle detection
Image pattern recognition
Persistence length

ABSTRACT

Structure determination of helical specimens commonly requires datasets from thousands of micrographs often obtained by automated cryo-EM data acquisition. Interactive tracing of helical assemblies from such a number of micrographs is labor-intense and time-consuming. Here, we introduce an automated tracing tool MicHelixTrace that precisely locates helix traces from micrographs of rigid as well as very flexible helical assemblies with small numbers of false positives. The computer program is fast and has low computational requirements. In addition to helix coordinates required for a subsequent helical reconstruction work-flow, we determine the persistence length of the polymer ensemble. This information provides a useful measure to characterize mechanical properties of helical assemblies and to evaluate the potential for high-resolution structure determination.

1. Introduction

Visualization of biological macromolecules by electron cryo-microscopy (cryo-EM) is one of the best suited methods to study the three-dimensional (3D) structure of large assemblies. Highly symmetrical assemblies, in particular helical structures, have been critical to establish methodology of 3D image reconstruction (DeRosier and Klug, 1968) as well as for further developments for high-resolution structure determination (Beroukhim and Unwin, 1997; Fromm et al., 2015; Ge and Zhou, 2011; Sachse et al., 2007; Yonekura et al., 2003). In addition, large helical assemblies constitute a fundamental architectural building principle in biology found in cytoskeletal proteins, viral capsids, enzymes, amyloid fibrils, membrane-remodeling and signaling complexes (Frost et al., 2008; Hirose et al., 1996; Lynch et al., 2017; Moore et al., 1970; Sachse et al., 2008; Wu et al., 2014), recently reviewed (Sachse, 2015). As new structures of helical assemblies are determined, new functional roles are being discovered in various processes of the cell (Ciuffa et al., 2015; McCullough et al., 2015). With improved hardware and software near-atomic resolution structures of these assemblies are increasingly common and critical to reveal the structural basis of the underlying assembly mechanism.

One of the main reasons for the improved performance of cryo-EM structure determination were due to hardware developments. First, direct electron detectors with improved detective quantum efficiency (DQE) and frame readout gave rise to images of much improved quality (McMullan et al., 2009; 2016). Second, microscopes of improved stability with software-assisted automation are now commonly used to generate 1000 s of micrographs from a single sample (Biyani et al., 2017; Mastronarde, 2005; Suloway et al., 2005). Another important advance comes from software improvements and high-performance computing that have reduced user interference and made large-scale image processing on high-performance computing architectures feasible over the last decades (Desfosses et al., 2014; Frank et al., 1996; Grigorieff, 2007; Ludtke et al., 1999; Scheres, 2012; Sorzano et al., 2004; Tang et al., 2007). In order to obtain near-atomic resolution in most cases 100,000 s of asymmetric molecular units need to be analyzed before averaging. Whereas for single particles a plenitude of semi-automated or fully automated procedures exist, for helical specimens most of the specimens are digitally excised by a user-guided interactive cropping procedure (Ludtke et al., 1999; Tang et al., 2007) before helices are being segmented into a stack of single particles. Although this process generates helical data sets of high confidence, it is labor-intense and time-consuming and can often take days for large data sets before image processing can be initiated.

The need for automated particle detection on micrographs from negative stain had been realized in the early 1980s (Frank and Wagenknecht, 1983) and since then many approaches for single particle detection have been introduced. Particle detection from cryo-micrographs is even more demanding due to the poorer signal-to-noise ratio. Multiple algorithms have been put forward to address this technical challenge. The proposed approaches rely on principles of template matching by reference-based cross-correlation (Roseman, 2003), pattern recognition (Zhu et al., 2003), edge detection (Harauz and Fong-Lochovsky, 1989) and other types of intensity measures reviewed by Glaeser and colleagues (Nicholson and Glaeser, 2001). More recently, deep learning algorithms based on neuronal networks have been added

* Corresponding author.
E-mail address: carsten.sachse@embl.de (C. Sachse).

https://doi.org/10.1016/j.jsb.2017.11.013
Received 29 September 2017; Received in revised form 24 November 2017; Accepted 26 November 2017

to the repertoire of techniques used for particle detection (Wang et al., 2016; Zhu et al., 2017). The principle goal of the particle picking workflow can be summarized as follows: to precisely locate the particle while avoiding to recognize noise or contaminants. In most cases, additional pruning is required either by a human operator or other means. For this reason, many programs are semi-automated such that they include a user GUI and require a final step of human intervention (Frank et al., 1996; Ludtke et al., 1999; Scheres, 2015; Tang et al., 2007).

For helical specimens, there are very few automated approaches known to date suited for fully automated or semi-automated selection of filamentous or elongated helical assemblies. One of the earliest approaches was developed by combining near and far-to-focus images to trace the central helical axis using the helix contours segmented by a Canny edge detector (Zhu et al., 2001). Although particularly useful for images of wide and rigid specimens such as tobacco mosaic virus (TMV), such approaches are operationally challenging due to dual image acquisition and it is more difficult to recognize thinner filaments with less molecular mass. In the past, the majority of helical assemblies were still excised interactively as relatively few high-quality images were sufficient to compute 3D maps of helical assemblies (Yonekura et al., 2003). With the introduction of single-particle image processing to helical reconstruction (Egelman, 2000; Jiménez et al., 1999), more flexible and heterogeneous helical structures became amenable to 3D image reconstruction. Therefore, classification methods originally developed for single-particle image processing algorithms could be directly applied to helical structures (Behrmann et al., 2012; Desfosses et al., 2014; Wang et al., 2006). More recently, the maximum-likelihood based RELION software (Scheres, 2012) has been adapted to work with helical structures. Based on the single-particle framework an adapted semi-automated helix detection workflow has become available (He and Scheres, 2017). Once the helical axis has been determined on the micrograph, the segments can be extracted and subjected to the respective image processing pipeline. In addition, traces of elongated specimens can also be used to assess material properties on the rigidity and flexibility of the examined structures (Sachse et al., 2010; Wang et al., 2002). For high-resolution and lower-level structural characterization a fast implementation is essential to reliably trace elongated assemblies while minimizing human intervention.

In the current manuscript, we introduce a robust micrograph-based helix tracing (MicHelixTrace) algorithm that automatically detects helical specimens based on a reference image. Computational cost is minimized by a reduced search separating estimation of rotational and translational parameters for helical specimens in Fourier and real space domain respectively. The resulting cross-correlation map is thresholded and helix coordinates are extracted from binarized skeletons. We demonstrate that the approach is successful for rigid assemblies of TMV, thin cytoskeletal ParM filaments as well as flexible p62-PB1 filaments. In addition to helix coordinates, the introduced automated tracing approach also determines fundamental material properties in the form of persistence length for helical structures. The automated helix detection algorithm MicHelixTrace provides a computationally fast implementation for faithful tracing of helical assemblies from electron cryo-micrographs with minimized human intervention.

2. Principle of the approach

One of the most successful approaches to automatically detect single particles from a reference image is based on the local cross-correlation function (Chen and Grigorieff, 2007; Roseman, 2003; Scheres, 2015), which yields the x and y positions of the particles on the micrograph. In contrast, helical assemblies are long entities that extend continuously for more than 500 Å in one direction. Consequently, their location description can be reduced to a trace with start and end coordinate pairs only when they resemble a line or by a set of equidistant segment coordinates when the helical axis deviates from a straight line (Fig. 1). Computing complete local correlation functions for every possible location on the micrograph is computationally demanding. In order to estimate end, start and segment coordinates of a continuous helix, computing a much smaller subset of the local correlation function is sufficient. For this purpose, we subdivided the micrograph into tiles with 80% overlap along x and y to match the provided reference image (Fig. 2A). Tiles of 350–500 Å dimension bear the advantage that the helical axis can be approximated as a line within this window even though for larger dimensions the polymer deviates from a straight line. The path of the helix within the tile can be explained by the in-plane angle θ and the normal distance or shift relative to the tile center Δ (Fig. 2B/C). To determine θ and Δ, an exhaustive search could be employed to further localize the helix within the window using a multi-dimensional cross-correlation (cc) function. In order to cut down on the computational cost of this search, we separate the in-plane rotation θ from the positional Δ search. Due to their symmetry, helical specimens possess layer lines in the Fourier domain. As the amplitude spectrum is invariant to translation of the object in real space, we can reliably determine the rotation θ by a rotational correlation function directly from the power spectra of the tile and the reference. According to the determined helix angle, we then rotate the tile and compute only one correlation function to determine the shift relative to the tile center Δ including a cc-score. As a result of the procedure, the θ/Δ pair and a normalized cc-score is obtained for every tile.

Based on the determined θ and Δ values of each tile, cc-scores can be mapped back on the micrograph with much higher precision than the initial coarse tiling (Fig. 3A). The resulting micrograph map exhibits high correlation for paths where the reference is present and shows small correlations where the reference is absent. In the next step, the cross-correlation map requires thresholding to reliably extract the helical paths. For this purpose, we evaluated the background correlation distribution in a histogram and found that it follows an exponential distribution for areas devoid of helices (Fig. 3B). Only few high correlation values represent the paths of the helices, which are above the null distribution visible in a tail of the histogram. After fitting the null distribution and estimation of the tail threshold, we binarize the correlation map. Subsequently, the paths are skeletonized to single pixels by a morphological thinning procedure (Fig. 3C) (Lam et al., 1992). Overlapping paths can be identified as branch points by a feature-specific response filter (Olsen et al., 2011). Once thinned to single pixel paths, branch points can only occur in a limited set of pixel configurations that can be located by a precomputed convolution response (Supplementary Fig. 1). Subsequently, overlapping helices are split into separate helices (Fig. 3D). Finally, helices shorter than a user-defined minimum helix length parameter are discarded as they often correspond to contaminations and non-helical protein, for example aggregates (Fig. 3E). Longer helices are broken up into pieces of comparable size and the remaining traces are fitted by a 1st to 3rd order polynomial function to yield a set of equidistant segment coordinates including the start and end points of the helix (Fig. 3F). In a final pruning step, we evaluate the determined traces with respect to straightness of the helix population from the micrograph set. As thermal helix trace fluctuation is determined by their inherent material properties, helices outside the expected distribution should not correspond to the targeted assemblies. In fact, they are often kinked helices unsuitable for helical reconstruction or represent false positive hits, e.g. contamination perhaps of elongated shape, which is erroneously detected as a helix (Fig. 3G). After the final straightness analysis, we can eliminate those helices that do not match the expected straightness range.

3. Implementation

3.1. Preprocessing of the micrograph into tiles

In order to locate the helical axis of the filament within the