ARTICLE IN PRESS

Journal of Structural Biology xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Journal of Structural Biology

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

De Novo modeling in cryo-EM density maps with Pathwalking

Muyuan Chen [a,c], Philip R. Baldwin [b,c], Steven J. Ludtke [c], Matthew L. Baker [c,*]

[a] *Program in Structural and Computational Biology and Molecular Biophysics, United States*
[b] *Department of Psychology, United States*
[c] *Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX 77030, United States*

ARTICLE INFO

Article history:
Received 7 December 2015
Received in revised form 6 June 2016
Accepted 7 June 2016
Available online xxxx

Keywords:
Protein structure
Backbone model

ABSTRACT

As electron cryo-microscopy (cryo-EM) can now frequently achieve near atomic resolution, accurate interpretation of these density maps in terms of atomistic detail has become paramount in deciphering macromolecular structure and function. However, there are few software tools for modeling protein structure from cryo-EM density maps in this resolution range. Here, we present an extension of our original Pathwalking protocol, which can automatically trace a protein backbone directly from a near-atomic resolution (3–6 Å) density map. The original Pathwalking approach utilized a Traveling Salesman Problem solver for backbone tracing, but manual adjustment was still required during modeling. In the new version, human intervention is minimized and we provide a more robust approach for backbone modeling. This includes iterative secondary structure identification, termini detection and the ability to model multiple subunits without prior segmentation. Overall, the new Pathwalking procedure provides a more complete and robust tool for annotating protein structure function in near-atomic resolution density maps.

1. Introduction

While cryo-Electron Microscopy(cryo-EM) has traditionally been limited to low resolution (Cheng, 2015), it is now possible to produce structures for large macromolecular complexes at better than 3 Å resolution (Fischer et al., 2015; Campbell et al., 2015; Bartesaghi et al., 2015; Jiang et al., 2015), largely due to technological advances in electron detector technology (Li et al., 2013). While such high resolution cryo-EM structures have been achieved in a handful of cases, the vast majority of cryo-EM reconstructions achieve lower resolution, making interpretation much more challenging (Esquivel-Rodriguez and Kihara, 2013; Baker et al., 2010a; Chiu et al., 2005). At these resolutions, the pitch of helices, separation of individual strands in β-sheets and even some bulky sidechains may be visible in the density map but often lack the detail for traditional protein structure modeling (Baker et al., 2010b).

As such, the ability to accurately interpret a density map in this resolution range is paramount to understanding structure and function in macromolecules. When available, high resolution structural models from X-ray crystallography and NMR are often fit to a cryo-EM density map in order to produce atomic level information about the complex (Rossmann et al., 2005). In some instances, the high resolution structural model is allowed to deform in order to best fit the density map (Villa and Lasker, 2014). Computational modeling tools have also been used to generate, score and select models based density constraints (DiMaio et al., 2009; Topf et al., 2005; Baker et al., 2003; Topf et al., 2006; Zhu et al., 2010). In addition, lower resolution annotations, such as secondary structure localization (Baker et al., 2007; Jiang et al., 2001) and domain segmentation (Baker et al., 2006; Pintilie and Chiu, 2012), can also be used gain insight into macromolecular structure and function despite the lack of atomic detail (Zhou et al., 2001; Jiang et al., 2003; Serysheva et al., 2008).

While lacking true high resolution detail, many of the cryo-EM density maps between 3–6 Å resolution contain enough information to build robust structural models for the individual components directly from a density map (Gipson et al., 2014; Baker et al., 2013; Zhang et al., 2011; Liu et al., 2010; Ludtke et al., 2008; Zhang et al., 2010; Liao et al., 2013; Yu et al., 2008; Yang et al., 2012). Under ideal circumstances, the path of the backbone through a density map is unambiguous. That is, a human examining the structure in detail could accurately trace a protein backbone through the density map without any branching, breaking or other ambiguities in the density map. However, in the majority of density maps at near atomic resolution this is not true, as some ambiguities, due to imaging and reconstruction errors and noise,

---

* Corresponding author.
E-mail address: mbaker@bcm.edu (M.L. Baker).

http://dx.doi.org/10.1016/j.jsb.2016.06.004

Please cite this article in press as: Chen, M., et al. *De Novo* modeling in cryo-EM density maps with Pathwalking. J. Struct. Biol. (2016), http://dx.doi.org/10.1016/j.jsb.2016.06.004

will likely exist. In addressing these issues, we developed a procedure, Pathwalking, to enumerate plausible backbone paths through a density map (Baker et al., 2012b,c).

Pathwalking is based on the Traveling Salesman Problem (TSP), in which possible cyclical paths (the protein fold) are calculated through a density map without using any sequence or structure constraints (Applegate et al., 2006; Helsgaun, 2009). Essentially, the only required input data for Pathwalking is a segmented density map better than 6 Å resolution and the number of amino acids in the protein. Briefly, the protocol starts with secondary structure identification in the density map using software such as SSEHunter (Baker et al., 2007). Pseudoatoms are then generated directly in the density map, representing the Cα atoms in the protein to be modeled. In the next step, the shortest path that visits each pseudoatom only once is assessed using the TSP solver. Essentially, the TSP solver finds the path through the pseudoatoms by optimizing the spatial distance between the pseudoatoms such that they are representative of Cα–Cα distances in consecutive amino acids in the protein structure. The path is then evaluated manually and the position of pseudoatoms adjusted. A new path is then calculated based on the updated position of pseudoatoms; this process is often iterated multiple times until an acceptable model is returned.

This approach has produced in a number of relatively accurate structural models from density maps at near-atomic resolutions (Fan et al., 2015; Baker et al., 2013). However, the Pathwalking procedure was not fully automated, and requires human input at specific stages to guide the overall process. Building on the original Pathwalking approach and motivated by the rapidly growing number of near atomic resolution structures, we present here an enhanced version of our original utilities capable of producing more accurate models with reduced user interaction. In addition to improved model building, Pathwalking now incorporates a secondary structure element (SSE) detection routine into the model building process that further improves the accuracy at which a protein fold is defined.

2. Methods

While the original implementation of the Pathwalking protocol was far better than purely manual model building, it did require a considerable amount of human interaction. This means the ultimate accuracy of the model is highly dependent on the skill of the person doing the modeling. The main interactive steps in the original version of Pathwalking includes identifying secondary structures using the SSEHunter (Baker et al., 2007), setting a threshold for the density map before pseudoatom placement, and finally, evaluating the identified paths and setting constraints. Even with the semi-automated approach of Pathwalking, these manual steps could be relatively time-consuming for large proteins. This is further complicated by the fact that multiple models may be required for large structures, making accurate map segmentation a critical issue. In the new version of Pathwalking, we have addressed these problems by introducing additional quantitative assessments and automation. A fully automated modeling process can be performed using the command:

```
e2pathwalker_auto.py density_map.mrc --natoms = N --denthr = T
```

Here `N` is the number of residues in the protein and `T` is the threshold of the isosurface value in the density map. The program will calculate a default threshold value if `--denthr = T` is not provided, but a carefully selected threshold value that separates features is still recommended especially at lower resolution. This script

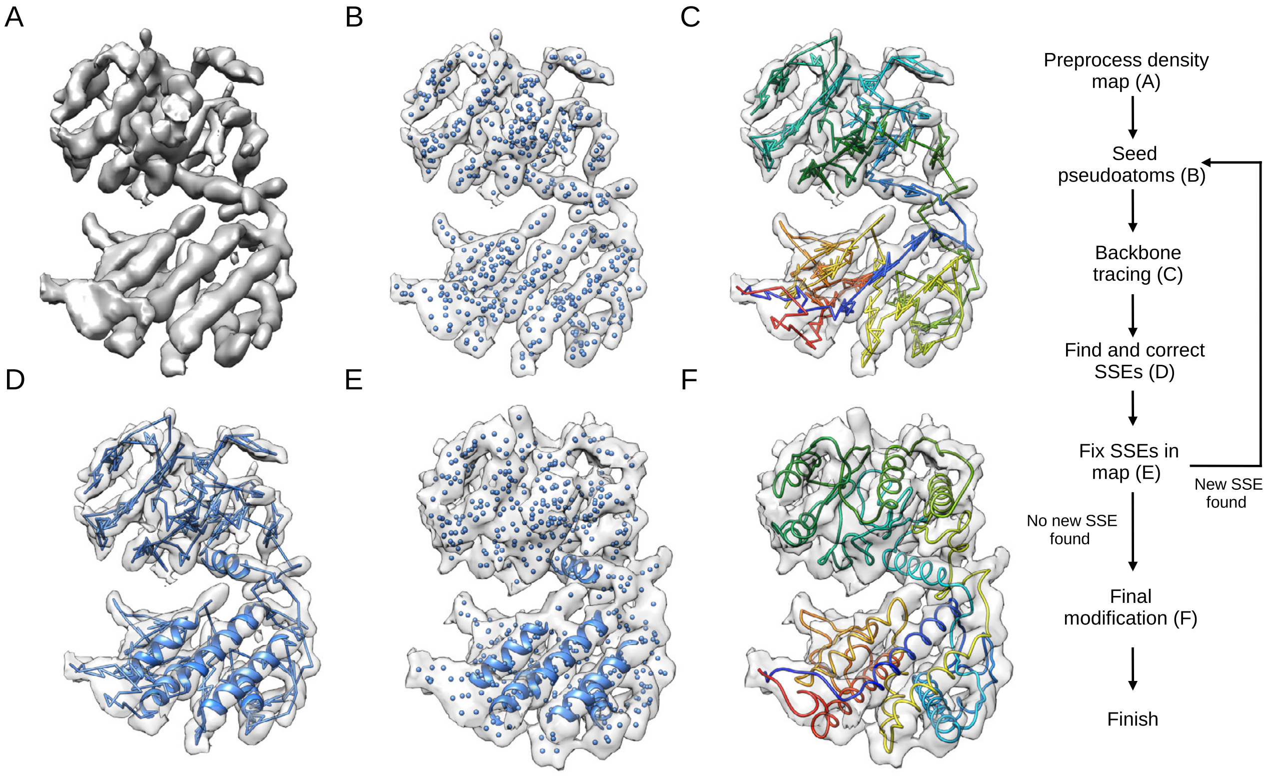

Fig. 1. Pathwalker work flow. (A) One subunit of the GroEL(EMDB ID: 5001, 4.2 Å resolution) is segmented out form the full GroEL density map. (B) Pseudoatoms are seeded in the density map. (C) Initial backbone tracing is performed using the pseudoatoms from B. (D) Secondary structure element are identified based on the density map and initial path. (E) The pseudoatoms are reseeded at with the fixed secondary structures. (F) Final result after multiple iterations of Pathwalking automated path modification.

Please cite this article in press as: Chen, M., et al. *De Novo* modeling in cryo-EM density maps with Pathwalking. J. Struct. Biol. (2016), http://dx.doi.org/10.1016/j.jsb.2016.06.004