

49. Discovery of Protein Substructures in EM Maps

Keren Lasker [1], Oranit Dror [1], Ruth Nussinov [2,3], Haim J. Wolfson [1]

1 School of Computer Science, Raymond and Beverly Sackler Faculty of Exact Sciences Tel Aviv University,

Tel Aviv 69978, Israel, e-mail: {oranit,wolfson}@post.tau.ac.il;

2 Sackler Inst. of Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel;

3 Basic Research Program, SAIC-Frederick, Inc, Lab. of Experimental and Computational Biology,

Bldg 469, Rm 151, Frederick, MD 21702, USA

Keywords: Structural bioinformatics, cryo EM, 3D alignment of secondary structures, macromolecular assemblies

Cryo-EM is a powerful technique for elucidating macromolecular structures that cannot be determined at atomic resolution. We present a highly efficient computational method for discovering atomic-resolution subunits of a complex in an EM map, without any prior knowledge about them. The method successfully recognized folds in EM maps from 887 SCOP representatives.

Structure determination of large macromolecular assemblies is one of the main challenges in structural genomics. To date, only 1.5% of the structures in the PDB are of large macromolecular complexes. The reason is that X-ray crystallography,

the most prolific and accurate technique for structure determination, has difficulties in the crystallization process of large and unstable assemblies such as membrane proteins and viruses. In the absence of crystals, cryo-electron microscopy (Cryo-EM) is a valuable source for studying both the structure and dynamics of large macromolecule assemblies. Although this method yields relatively low resolution of the data (6 to 30 ° Å) in which atoms cannot be discriminated, cryo-EM methods can be synergistically combined with atomic resolution methods for structure determination. For large flexible complexes

that cannot be crystallized, it is possible to fit the atomic structures of the individual subunits into an EM map of the entire complex. The resulting quasi atomic model of the complex may provide crucial information about the interactions of its subunits. Indeed, several hybrid approaches have been developed for fitting atomic resolution domains into cryo-EM data. A major drawback of these methods is the assumption that the searched domain and its conformation is present in the map. We propose a fully automated method for the discovery of subunits in intermediate resolution (7-9 ° Å) maps of complexes

without prior knowledge of their boundaries and content. We exploits the facts that (i) in intermediate resolution maps secondary structure elements (SSEs) become apparent and that (ii) the scaffold of a domain is defined by its SSE spatial arrangement. The method consists of two steps, helix extraction and fold alignment. The hybrid method detects helices in an EM map and uses the 3D arrangement of the identified helices to query a database of high resolution structures to find potential homologous folds. The method is highly efficient and can detect 'partial alignments' between the extracted set of helices and the folds of the database. Thus, it is tolerant to errors in the helix extraction stage and capable of detecting non predefined motifs. The method was tested successfully on several simulated 8.0 ° Å resolution maps. The obtained spatial helix arrangement was sufficient for the discovery of homologous folds from a dataset of 887 SCOP representatives. Figure 1 presents an example of identifying an $\alpha\alpha$ superhelix subunit of an AP2 core complex (PDB:1gw5). Despite the size of the

complex, a set of 80 out of 84 helices was detected with no false positives in less than ten minutes. This set of helices

was then used to query the SCOP database. The top-ranking was one of the $\alpha\alpha$ superhelix domain of the complex. The homologous fold was correctly aligned on the map with an RMSD of 3.8 °A between the midpoints of the helices' central axes.

Figure 1: The alignment between the detected set of intermediate-resolution helices for an endocytic AP2 core complex (PDB:1gw5, orange) and a high resolution structure of one of its $\alpha\alpha$ superhelix domain (PDB:1gw5, chain B, green)