Modeling large protein structures as graphs for automated analysis of their topology

J.N. Wolf, M. Zunker, J. Ackermann, I. Koch

Molecular Bioinformatics, Goethe-University, Frankfurt am Main, Germany

Wolf@bioinformatik.uni-frankfurt.de

The increasing number of protein structures and the increasing size of protein structures calls for automated methods. The Protein Topology Graph Library (PTGL) [1, 2] models the topology of protein structures as graphs. PTGL supports three levels of abstraction: amino acids, secondary structure elements (SSEs) and chains. For each level of abstraction, the vertices correspond to the level, i.e., vertices correspond to amino acids on amino acid-level and so on. On all abstraction levels, edges denote spatial neighborhoods (contacts). Contacts are based on the computation of Euclidean atom-atom distances. On SSE-level, vertices are labeled as helix or strand. Edges are labeled by the orientation of their SSEs as parallel, antiparallel or mixed. On chain-level, edges are weighted with the number of residue-residue contacts (see Fig 1.).

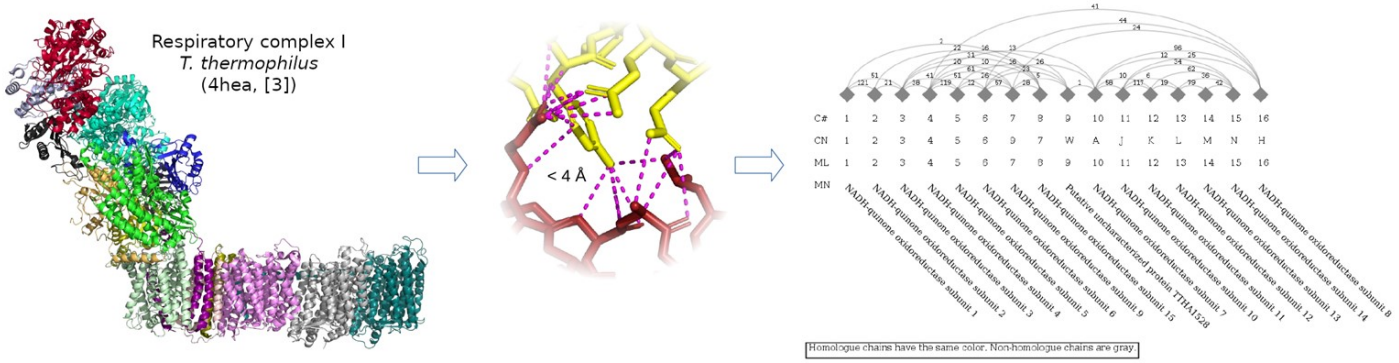

Figure 1. Schematic overview of Protein Topology Graph Library's modeling of the chain-level topology of respiratory complex I of *T. thermophilus* [3] as graph. On the left and in the middle, there are three-dimensional protein structures in cartoon and stick representation, respectively. Chains are colored individually. In the middle, atom distances below 4 Å are marked as dashed lines. On the right, a Complex Graph is visualized. Vertices denote protein chains and edges spatial neighborhoods. The edges are weighted with the number of residue-residue contacts. Below the graph, there is the number of each vertex (C#), its chain name (CN), its molecule ID (ML) and its molecule name (MN).

We used chain-level Complex Graphs (CGs) as a highly abstracted and meaningful view on respiratory complex I. We compared the CGs of the core subunits of complex I between *T. thermophilus* and *H. sapiens*. The Complex Graphs shared 29 edges. Each CG had one edge that the other has not. Therefore, the CGs were able to capture the topology of structurally conserved regions.

We applied hierarchical clustering to the edges of a CG. We compared the resulting dendrogram with an assembly process proposed in the literature [4]. Solely based on the CG, we found similarities between the dendrogram and the proposed assembly process. We also applied graph clustering to investigate whether complex I's modules could be extracted solely from the CG. We showed that CGs could identify modules and guide the finding of the assembly process for complexes.

Concluding, PTGL provides graphs modeling the topology of protein structures on different levels of abstraction for 151.837 PDB structures, including 921 *large structures*. The webserver provides a search for predefined motifs and user-defined arbitrary patterns. The computation is automated and the implementation publicly available. The representation of graphs enables the application of graph-theoretic methods, such as graph partitioning. This allows feasible analyses on the rapidly growing PDB.

[1] Wolf, J.N., Keßler, M., Ackermann, J. & Koch, I. (2020). *Bioinformatics*. **37(7)**, 1032-1034.

[2] May, P., Kreuschwig, A., Steinke, T. & Koch, I. (2009). *Nucleic Acids Res*. **38,** D326-D330.

[3] Baradaran, R., Berrisford, J.M., Minhas, G.S. & Sazanov, L.A. (2013). *Nature.* **494,** 443-448.

[4] Guerrero-Castillo, S., Baertling, F., Kownatzki, D., Wessels, H.J., Arnold, S., Brandt, U. & Nijtmans, L. (2017). *Cell Metabolism.* **25(1),** 128-139.

Keywords: protein topology; graph theory; large structures; structural bioinformatics