

Mining Protein Space: Discovering Frequent Topological Structures

Ruoming Jin^{*}, Helen Piontkivska[#] and Dong Wang^{*}

Department of Computer Science and [#]Department of Biological Sciences

Kent State University, Kent OH 44242

jin@cs.kent.edu, opiontki@kent.edu, dwang@cs.kent.edu

Graphs are a powerful representation for many complex biological entities, such as protein structures and cellular interactomes. Frequently occurring patterns in these graphs may provide useful insights into biological systems. Therefore, finding such patterns becomes an increasingly important task in bioinformatics research. Here, we propose a framework to mine frequent large-scale structures, formally defined as frequent *topological structures*, from graph datasets. This is driven by the need to uncover the hidden *large-scale structures* that represent high-level topological information in biological entities, e.g., the non-local tertiary substructures that are important in protein structure analysis. Key elements of our framework include fast algorithms for discovering frequent topological patterns based on the well known notion of a topological minor, algorithms for specifying and pushing constraints deep into the mining process for discovering constrained topological patterns, and mechanisms for specifying approximate matches when discovering frequent topological patterns in noisy datasets. We demonstrate the viability and scalability of the proposed algorithms on real and synthetic datasets and also discuss the use of the framework to discover meaningful topological structures from protein 3-D structure data.