295e A Generic Motif Discovery Algorithm for Diverse Biomolecular Data

Kyle L. Jensen, Mark Styczynski, and Gregory Stephanopoulos

Here we present a novel algorithm, GeMoDA (Generic Motif Discovery Algorithm), for discovering repeated patterns in real-valued, sequential data without requiring alignment or pattern enumeration. GeMoDA is closely related to the Teiresias algorithm, but borrows new concepts from frequent itemset mining and graph theory to extend its capabilities. GeMoDA is generic in the sense that it is applicable to almost any type of sequential data, e.g. DNA/protein sequences, protein structures, and time series data. Given a user-defined notion of similarity and identity, GeMoDA deterministically enumerates all repeated patterns of a minimum length and support. In the context of DNA sequences, for example, similarity may be the Hamming distance between two subsequences. These two subsequences may be considered identical if they have at least 5 bases in common over a window of 10. Given such a problem statement, GeMoDA will find all patterns that occur at least k times and are of at least size l---where k and l are user inputs---such that all the patterns are maximal in both support and length. These patterns can be modeled using any number of motif representations including position-weight matrices, hidden Markov models, and regular expressions.

We show applications of GeMoDA in three domains: protein and DNA sequence motif discovery, and the discovery of repeated substructure discovery in protein structures. Unlike other algorithms for these two applications---viz. Gibbs sampler, MEME, Dali, and others---GeMoDA's output is guaranteed to be both optimal and exhaustive.