Protein Secondary Structure Assignment (SSA) by Clustering Amino Acid Residues in the Space of Topological Descriptors

Ratna, Sumanth (School: Thomas Jefferson High School for Science and Technology)
Gupta, Sagar (School: Thomas Jefferson High School for Science and Technology)

Protein secondary structure is defined by the local interactions between amino acid residues and it is widely used in protein structure and function analysis. Secondary structure information plays an important role in many applications including antibody development, multiple sequence alignment, and homology modeling. Although many methods of protein secondary structure assignment (SSA) exist, inconsistencies frequently arise due to differences in parameters and assignment criteria. With the goal of creating an objective and parameter-free SSA method, we trained an unsupervised k-means clustering model on topological information in the form of a 64-dimension feature set calculated from the Delaunay tessellations of a large set of amino acid residues. The resulting clusters are highly similar to the assignments by the most commonly used SSA algorithm, DSSP, with an overall agreement of 83%. Through dimensionality reduction and statistical analyses, we identified the five topological descriptors that are most important in distinguishing between secondary structure types. Further, we conducted a Kruskal-Wallis rank-sum test and found that the average difference between the phi and psi angles was statistically significant across all three clusters (p-value < 2.2e-16). With no a priori definitions of protein secondary structure, the high accuracy of our unsupervised machine learning model proves that protein topology can be successfully used to distinguish between secondary structure elements. Such a method, due to its parameter-free nature, can also be used as an arbitrator where other SSA algorithms disagree and has applications in precision medicine and biomedical research.

Awards Won:

Third Award of $1,000