Heather Nelson, Computer Science

Year in School: Senior
Hometown: Wildwood, MO
Faculty Mentor: Dr. Chi-Ren Shyu, Computer Science
Funding Source: National Science Foundation, University of Missouri-Columbia Research Council

Efficiency and accuracy validation for incremental changes of a large-scale protein structure database

Proteins, the essential building blocks of organisms, have many important roles, from providing structure to aiding movement and digestion. The construction of proteins involves one or more polypeptide chains that fold into complicated 3D structures. Each protein has a unique shape and some specific functions, which are intricately and irrevocably connected. In order to aid the study of structure-function relationships, ProteinDBS developed at the University of Missouri-Columbia presents a fast structure retrieval system to find proteins with such structural similarities. To present the most accurate results, ProteinDBS features automatic weekly updates of its system from the Protein Data Bank (PDB) which has over 76,000 protein chains and continuously grows the database size at least linearly. This research focuses on the efficiency and accuracy of protein structure retrieval using the ProteinDBS system as the size of the dataset grows. The investigation examines changes in results arising from the addition of new proteins to the system and illuminates the reasons for differences among search results. First, the system automatically checks protein domains and folds after insertion of new proteins. Testing proteins collected from various plants, such as maize and soybean, are validated against both the original dataset and the new, larger dataset. The systems compares the results from both sets of data to determine the changes in the composition of the result set, including the proliferation of newly inserted proteins, and the relative ordering of proteins in the ranked results. The analysis provides a thorough investigation of the effect the dataset has on protein structure retrieval and suggests areas for future improvement of the algorithmic designs of ProteinDBS in feature extraction, database indexing, and result ranking.