Events

Current Events
Lectures
Events Archive

lecture image

CCT Colloquium Series

Protein Data Mining for Bioinformatics Applications

Sumeet Dua

Johnston Hall 338
May 05, 2006 - 03:00 pm

Abstract:
One of the daunting challenges facing Biology, and consequently multidisciplinary research in Computer Science, is to assign biochemical and cellular functions to the thousands of hitherto uncharacterized gene products discovered by several international gene-sequencing projects. These research endeavors are producing high dimensional, heterogeneously distributed data at an unprecedented rate, much more rapidly than the corresponding development of computational techniques capable of novel knowledge analysis and discovery. Data mining offers the promise of precise, objective, and accurate in-silico analysis of this emerging data using knowledge discovery routines that reveal embedded patterns, trends, and anomalies in order to create models for faster and more accurate physiological discovery. Protein structure classification and comparison has become a central area in the field of bioinformatics. Proteins serve as one of the major structural elements of living systems and their interactions determine most of the molecular and cellular operations within these systems. The quantity, complexity, and availability of protein structure databases has been increasing at a nearly exponential rate, leading to the demand for the development of automatic and expeditious techniques for protein structure comparison, classification, modeling, and functional prediction. Protein databases commonly suffer from the 'curse of dimensionality', necessitating the endeavor to reduce the dimensions and intricacy of ingrained information prior to protein classification. The determination of a protein's structure and function from its amino acid sequence has also provided an exciting challenge. Successful methods for the categorization of protein structural classes from sequence information involve multiple physio-chemical properties and machine learning algorithms. In this presentation, I will present a novel data mining algorithm for three dimensional (3D) structure-based classifications of proteins using orthogonal Euclidean distance-preserving transformations of the geometric shape descriptors derived from experimental observation of protein structures. An association rule-based and k-means, nearest neighbor-based supervised clustering approach is also employed to classify proteins. Experimental results demonstrate the accuracy and performance of the proposed techniques and their superiority over previous results. I will also present a unique computational methodology for protein structural classification using both physical and stereochemical properties of the protein sequence. While existing research in the area integrates information about amino acid composition, hydrophobicity, normalized van der waals volume, polarity, and polarizability to predict a protein structure, our research focuses on the development of an accurate feature vector based on determining points of spectral coherence between different hydrophobicity scales and exploiting this vector-space to achieve improved sensitivity and specificity in protein structural classification. Support vector machines and random forest classifiers are also developed for multi-class classification of proteins, achieving significant increases in precision accuracy. The results will demonstrate that hydrophobicity is a significantly contributory factor in evaluating discriminatory behavior of proteins, yielding a substantial 71% increase in precision accuracy compared to previous experimental results as executed on an independent dataset. The presentation will conclude with some directions for future investigation and improvement.

Speaker's Bio:
Dr. Sumeet Dua is an Upchurch assistant professor of computer science, coordinator of information technology research, and director of Data Mining Research Laboratory (DMRL) at Louisiana Tech University (Tech). He also serves as an adjunct assistant professor of research at LSU School of Medicine, New Orleans. Within his fields of research interests--data mining, bioinformatics, and distributed database integration--he has (co-) authored several publications and holds two patents. Additionally, he has served as a chair and as a program committee member for several special sessions and conferences in these areas, and frequently presents invited talks on data mining and bioinformatics, at institutions in both academia and industry. His research is funded by Louisiana Board of Regents, National Institutes of Health, and National Science Foundation.