Statement of Purpose

Applicant's Name: Abida Sanjana Shemonti

Research Experience: I was motivated by my professor Dr. M. Sohel Rahman, who possesses a diverse research interest that includes theory and algorithms, Stringology, Metaheuristics, Applied Informatics, Computational Biology and Bioinformatics. For my undergraduate thesis, I have worked under his supervision in the field of Bioinformatics. We designed a Secondary Structure Predictor for protein by incorporating an estimation of energy, PSEE (Position Specific Estimated Energy) on another Balanced Secondary Structure Predictor MetaSSPred. Secondary Structure of protein refers to a configuration of the local substructure of polypeptide backbone atoms that fold to generate three major kinds of configurations, Helix (H), Beta (E) and Coil (C). Helix and Beta residues are preferably located in the core of the protein, having favorable energy. However, Beta residues are more structured compared to the Helix residues. Coil residues stay in the surface areas of proteins and are highly flexible, having unfavorable energy. Our experiments found that the average PSEE values for H, E and C of a significant number of protein residues were reasonably similar to the theoretical concepts and thus validated the usefulness of PSEE in identifying Secondary Structure of protein residues. We used three independent binary SVMs (Support Vector Machines) with a sigmoid kernel to calculate the probability for each residue belonging to H, E or C structure. We combined the classification results from these three predictors using Genetic Algorithm with optimal weights of the individual class, to develop a new predictor PSEESSPred, that performed better than state-of-the-art predictor SPINE X and MetaSSPred.

A number of undergraduate courses have helped me to strengthen my foundation to pursue academic research. The most propelling factors have been the courses on Artificial Intelligence in junior and senior years. Different Machine Learning and Pattern Recognition concepts helped me a lot in my undergraduate thesis. Out of motivation, I have worked on species identification using partial DNA sequences where I have used DNA Barcode as partial DNA sequence. DNA Barcode is a short genetic marker in an organism’s DNA to identify it as belonging to a particular species. Specimens were chosen from phylogenetically diverse species belonging to the animal, plant and fungus kingdoms. For identification, I considered Simple Logistic Function, Random Forest, PART, K-Nearest Neighbor and Bagging from the Weka software suite. The results showed that the classification performances of the selected methods are at a comparable level, and even superior in some cases, to the well-established DNA Barcode classification methods- BLAST, BLOG, DNA-BAR, PAR, Nj and NN. The Technical Writing and Presentation course has been helpful in developing scientific writing skills. I have learned Image Enhancement, Filtering, Segmentation, Image Retrieval and Image Compression techniques in Digital Image Processing course, which I believe will help me to contribute in the area of security and surveillance, medical diagnostics, environmental monitoring and object recognition.

Long-term Degree Objective: My undergraduate thesis works introduced me to the possibilities of capturing biological systems into computational models and solving related biological problems computationally. This interest was elevated by an opportunity to meet Professor Costas Iliopoulos, King's College London, as he shared ideas and possibilities about research on Algorithms and Bioinformatics. The fact that there are opportunities to design remedies that might cure critical diseases like cancer or Alzheimer disease, and ensure a healthier life for people have fueled my ambition to look beyond the fixed academic curriculum and pursue research works further to discover the new in this discipline. Therefore I aim to steer my career towards the research-based direction to become a competent practitioner in the research field and contribute to the community.

Future Research Interest: My research interest is in Computational Biology and Bioinformatics, notably in applying Machine Learning techniques in Bioinformatics. I am intrigued by the research activity of Professor Daisuke Kihara focusing on developing computational methods to predict and analyze protein structure or function, as they align with my previous experiences. It would be a great opportunity to work with Professor Kihara on predicting protein-protein interactions and their applications in drug discovery. I am interested in the works of Professor Wojciech Szpankowski on Bioinformatics. His work with Professor Kihara on protein Superfamilies and Superfolds especially drew my attention. I am also interested in the works of Professor Alex Pothen on Flow cytometry to understand the immune system. In addition, I am open to working with Associate Professor David F. Gleich on biological data analysis. I am also happily willing to work with others, working in the fields of Bioinformatics and Computational Biology, Machine Learning and Data Mining, and Spatiotemporal Data Management and GIS, if opportunity permits.

I believe the strong research facilities at Purdue University will help me develop into a prominent researcher in this discipline. As I come from a culturally diverse community, I think I can blend in the student community here well. I am very much aware of the fact that getting admission in universities for a Ph.D. program is a competitive process. In spite of that, I believe my experiences from undergraduate studies and skills as a lecturer have prepared me for meaningful contribution to Purdue University as well as to the community.