Name of the Scholar: **Meenakshi Bhat**

Name of the Supervisor: **Dr. S. A. M. Rizvi**

Department: **Computer Science**

Title of the Thesis: **"*In silico* Comprehensive Sequence and Structure Analysis of Protease Family"**

ABSTRACT

Biologists are in need of information in digital form for correct and meaningful interpretation of their biological data and Bioinformatics is the computing response to the molecular revolution in biology. Bioinformaticians and molecular biologists often need molecular sequences like DNA, RNA & proteins to compare them with each other in order to determine the degree of similarity on the basis of which various conclusions can be derived regarding the features, structures, behavior and function of an organism or entire species as a whole.

The present ongoing research here is an attempt to develop a specific algorithm for identifying particular domains in the genome sequences of the protein enzymes, proteases. Proteases, also known as proteinases or peptidases are protein enzymes, which cut long sequences of amino acids and therefore, called by many as biology's version of "Swiss Army Knives". On the basis of their sequence analysis, one can identify their types and also can predict their secondary or tertiary structures. Besides, analysis based on phylogenetic relation of these proteases by constructing their phylogenetic trees in the light of evolution can be done. Storing all the information extracted from these sequences in a new database is another perspective of the present study.

To address these problems, a three-layered predictor, “Profinder”- proteases identifier, is shaped by collecting the protease domain families represented by alignments and Hidden Morkov Models (HMMs). Molecular biologists use hidden markov models as a popular tool to statistically describe biological sequence families. This statistical description can then be used for sensitive and selective database scanning, e.g., new protein sequences are compared with a set of HMMs to detect functional similarities. The first layer is for identifying the query protein sequence as protease or non protease; if it is a protease, the process will automatically go to the second layer to further identify it amongst the six types of proteases, and the third layer will be for structural analysis. User can test their sequences in fasta format for identification of proteases domain and the identification of domains that occur within proteins can therefore, provide insights into their functions.

The ongoing interface is currently the only program available for detection and classification of these proteases, since no existing database or annotation program is able to identify, classify and represent three dimensional structures of these proteases one after another in a particular search algorithm. These features make this program a valuable resource which can aid in identification of proteases and its type found in newly sequenced genomes. This tool represents a collection of profile Hidden Markov Models (HMMs) based on a rigorous analysis of six distinct domain families of protein enzyme, Protease, namely, (i) Aspartic acid domains, (ii) Cystein domains, (iii) Glutamic (iv) Domain of Metalloproteases, (v) Serine, and (vi) Threonine domains. Focus here is on offline automated, interactive and predictive search tool, the first of its kind dedicated to proteases domains.