Timothy Goya

Senior Project Proposal:
Using Uncertain Reasoning to Identify Peptide Sequences in Mass Spectra

Problem

Proteomics, the study of proteins, has potential applications in medical diagnosis and research. Using tandem mass spectrometry (MS/MS), researchers try to identify the protein content of a sample. The proteins are not directly measured by the mass spectrometer because the proteins are too massive. First, an enzyme digests proteins into smaller pieces called peptides. These peptides are often unique to a particular protein, so if a peptide is identified in a sample, that sample likely contains the associated protein. Many thousand mass spectra are generated by a sample, too many to manually process. Computers would be ideal for this task, but unfortunately no algorithm currently exists that identifies mass spectra accurately enough for pratical usage. Current algorithms are also implemented in frameworks that are not focused solely on peptide identification and contain useless features that only impede processing speed.

Solution

The identification engine consists of a chemistry library, a set of plugins, and the main glue executable. The chemistry library contains primitives to allow easy manipulation of the spectra and the sequences. The plugins provide different implementations for the identification process. Some types of these plugins include virtual enzymes, the data mining algorithm, and the scoring algorithm. The main executable glues together the functionality provided in various plugins as directed by a configuration file.

The data mining algorithm primarily takes as input a list of spectra and a list of proteins. Since the proteins in the sample undergo enzyme digestion, each protein in the list must also undergo virtual enzyme digestion. A real enzme cleaves the protein according to several chemical factors including enzyme shape. Virtual enzymes use a set of rules to cleave the virtual proteins at the same locations the associated real enzyme would to a real protein. The data miner must remember the protein a particular peptide originated from because otherwise the results would be worthless for tools that use those results. A scoring algorithm compares a spectrum and a peptide to generate a "goodness" measure on how likely the peptide caused the measured spectrum. A fixed number of the top matches for a single spectrum are then written out into a file for input into a variety of protein identification and analysis tools.

Schedule

In the first quarter I will analyze X!Tandem and OMSSA, two current solutions that have been released under an open-source license. During the analysis process I will reimplement the algorithms into a common plugin-based framewor for comparison and to show that my analysis is correct. For the second quarter I will develop my own algorithms using what I have learned by studying the current algorithms, as well as the underlying chemistry.

Meeting Minimum Criteria

Independence

The design and implementation of the identification engine is entirely my work.

Background Research

In order to successfully complete this project, I need to research the protein chemistry, various uncertain reasoning techniques, and the specific strengths and weaknesses of current frameworks and algorithms.

Creative

The data miner and virtual enzymes have well defined requirements and a limited number of valid solutions. However, the scoring algorithm is extremely open-ended and requires considerable creativity.