

## Peptide Mass Fingerprinting Using Machine Learning

Rachana Jain  
Department of Biomedical Engineering  
University of Cincinnati  
Cincinnati, OH 45221  
jainrn@email.uc.edu

Michael Wagner  
Division of Biomedical Informatics  
Cincinnati Children's Hospital Research Foundation  
Cincinnati, OH 45229  
mwagner@cchmc.org

The identification of proteins in complex biological samples is a critical task in many biochemical experiments. Peptide Mass Fingerprinting (PMF) involves purification of the protein using methods such as 2D-gel electrophoresis, followed by its digestion using a proteolytic enzyme such as trypsin. The peptide masses are then measured using a mass spectrometer and searched against a database of theoretical peptides obtained by *in-silico* digestion of proteins. Various factors such as the accuracy of mass spectrometers, the presence of contaminants in the sample, limited databases and post-translational modifications of proteins complicate the task for PMF and limit its success as a protein identification method.

The crucial ingredient in PMF methods is the definition of an appropriate scoring function that can accurately distinguish between random hits and true positives. We propose several innovations toward this goal. We first identify a number of new "features" (quality measures) to quantify the quality of matches between experimental and theoretical digests. Secondly we propose the use of machine learning techniques as a more principled method to combine the features for designing the scoring function.

Several factors can affect the efficacy of using machine learning as a method for generating the scoring function. Therefore, we have developed a training and testing framework which allows us to perform systematic studies of the robustness and sensitivity of the scoring function with respect to measurement errors, contamination and number of missed cleavages. Our initial investigations are based on simulated data for 860 non-redundant proteins. A constraint with using such large datasets is the time taken for training and testing. To alleviate this problem, we have parallelized the entire process and have been able to reduce the time from several days to less than an hour.

Our results on the simulated data as well as first studies on real data indicate high potential of our method to improve on the state of the art. The long term goal of the project is to develop a model that can predict the protein identity with significant accuracy when searched against public databases such as swissprot, while taking into account the deviations or errors induced by experimental conditions and limitations. Finally, we envisage making this tool available to the general public through a website.