

UCLA IPAM RIPS Project 2005

Project description

UCLA Faculty: Matteo Pellegrini
Sponsor Liaison: Roland Luethy, TimeLogic Corp.

Scoring functions for matching peptide tandem masspectra to protein sequence databases

Background

Masspec based proteomics is a modern technology used to identify proteins in complexes, cellular components or bodily fluids. It can not only identify the proteins, but also find posttranslational modifications. For a recent review see [1]. In a typical proteomics experiment a mixture of proteins is subjected to enzymatic proteolysis, most frequently trypsin is used for this digestion. The resulting mixture of tryptic peptides is fractionated by HPLC chromatography and the fractions are subjected to tandem masspectroscopy. The first mass separation isolates tryptic peptides according to their mass over charge ratio. Peptides within a mass range can then be selected and fragmented. The resulting fragments are then analyzed with the second masspec step producing the fragmentation spectrum. Fragmentation occurs most frequently by breaking of the peptide bonds, producing a series of fragments with mass differences corresponding to the amino acid masses, but other bonds break too and not all peptide bonds break with the same probability. This leads to more complicated and harder to interpret spectra than expected. There are attempts to derive the peptide sequence from the fragment spectra, but the more successful tools for interpretation rely on using theoretical spectra derived from the protein sequence database and matching these theoretical spectra to the observed ones. The most popular software tools for this kind of search are Sequest [2] and Mascot [3], but there are several others including two recent open source implementations [4, 5]. These tools work by calculating peptide molecular weights and theoretical fragment masses and then comparing them to the masses obtained from the masspec experiments. Sequest calculates a correlation score between the observed and calculated spectrum. Mascot is using a probability based scoring function. Both tools then rank the results by the scores and it is up to the user to decide which matches are significant. Significance depends on the size and composition of the database, the number of fragments in both the calculated and observed spectra and other factors. It would be useful to be able to estimate the significance of a match independent of database and spectra.

Goals

The main goal is to find a scoring method for comparing theoretical and observed spectra, that can report a probability for a match to be significant. This probability should

take into account database size, number of peaks in the spectra and possibly intensities of the peaks. The following questions should be investigated:

1. what is the probability of a peptide with mass $m(\text{calc})$ to be found in a sequence database of size N ?
2. what is the probability of a fragment with mass $f(\text{calc})$, given a peptide of mass $m(\text{calc})$?
3. what is the probability of a peptide mass match $m(\text{calc}) = m(\text{obs})$?
4. what is the overall probability of a match taking into account peptide mass and fragment masses?
5. can the scoring be improved by filtering the observed spectra, e.g. by removing masses with low intensities?
6. can the scoring be improved by taking into account the intensity values of the observed spectra?

Deliverables

- A function to calculate the significance of a match, by estimating the probability of the match to occur by chance.
- A computer program, written in C++, that calculates the significance, given a MS/MS spectrum and a list of candidate peptides.

References

1. Steen, H. and M. Mann, *The ABC's (AND XYZ's) of peptide sequencing*. Nat Rev Mol Cell Biol., 2004. **5**(9): p. 699-711.
2. Eng, J.K., A.L. McCormack, and J.R. Yates, III., *An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database*. J. Am. Soc. Mass Spectrom., 1994. **5**: p. 976-989.
3. Perkins, D.N., et al., *Probability-based protein identification by searching sequence databases using mass spectrometry data*. Electrophoresis, 1999. **20**(18): p. 3551-67.
4. Geer, L.Y., et al., *Open mass spectrometry search algorithm*. J Proteome Res, 2004. **3**(5): p. 958-64.
5. <http://www.thegpm.org/TANDEM/index.html>