Reducing false positive rates in MS/MS sequence searching and incorporating intensity into match based statistics

Lewis Y. Geer[1]; Dina L. Bai[2]; Jeffrey A. Kowalak[3]; An Chi[2]; Ming Xu[1]; Jeffrey Shabanowitz[2]; Sanford P. Markey[3]; Donald F. Hunt[2]; Stephen H. Bryant[1]
[1]National Library of Medicine, NIH, Bethesda, MD; [2]University of Virginia, Charlottesville, VA; [3]National Institute of Mental Health, NIH, Bethesda, MD

Overview and Introduction

Correlation Refinement

- False positives, a.k.a "one hit wonders," are a key issue in the interpretation of MS/MS sequence search results.
- Recent technological developments, including Electron Transfer Dissociation [ETD](1) and high resolution mass spectrometers such as the hybrid linear ion trap/electrostatic ion trap, benefit from algorithms that have low false positive rates. Search results, particularly those from high resolution mass spectrometers, can be biased by correlations between ions, especially if only a few ions are required for a match. ETD data can require searching many precursor charge states, each of which can generate false positives.
- We introduce a refinement to the Open Mass Spectrometry Search Algorithm [OMSSA](2) that addresses these correlations and helps reduce the number of false positives.

Intensity based scoring

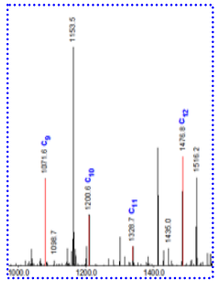
- Many sequence search algorithms do not directly incorporate intensity into peak match statistics. For example, previous versions of OMSSA only indirectly incorporated peak intensities by dynamically adjusting a peak intensity cutoff. We describe the direct addition of intensity based scoring to match based statistics.

Methods

- The OMSSA algorithm was written in C++ using the NCBI C++ toolkit.
- High mass resolution Collision Activated Dissociation [CAD] spectra were obtained from a commercial hybrid linear ion trap/electrostatic ion trap mass spectrometer, the Thermo LTQ Orbitrap (Thermo Electron, Waltham, MA). The 3844 spectra analyzed were generated from tryptically digested E. coli ribonucleoprotein complexes that contain approximately 55 ribosomal proteins and 3 RNA molecules.
- ETD spectra were obtained from a commercial quadrupole linear ion trap, the Finnigan LTQ mass spectrometer (Thermo Electron, Waltham, MA) equipped with a modified nanoflow electrospray ionization source. The LTQ was modified to accommodate a Finnigan 4500 CI source (Thermo Electron) placed at the rear of the instrument to facilitate ETD. The 1180 spectra generated were of 5 commercially available proteins digested with Endo Proteinase Lys-C.

Results

Correlation Refinement



To incorporate a correlation refinement into OMSSA, it was necessary to simplify the match statistics used in previous versions:

- The first step was to break the spectrum into bins centered on theoretical ions (see example to the left). We then model the probability of matching an ion as a Poisson distribution with

$$\mu_i = 2t_i / e_i$$

- where t_i is the mass tolerance of instrument and e_i is the number of experimental ions per unit mass in the bin i.
- The total distribution for all bins in the spectrum is a Poisson distribution with

$$\mu = \sum_i \mu_i$$

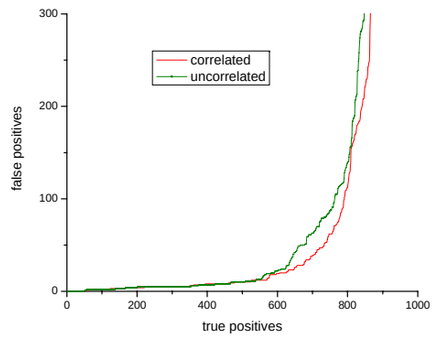
Note that the Poisson match statistics imply that neighboring ion peaks are uncorrelated. However, neighboring peaks are correlated as there are a finite number of unique amino acid masses, e.g. 19. Calculating the probability distribution of random matches must take into account this correlation. For simplicity, we use an approximation:

- the initial match in a series of consecutive matches is calculated as above.
- consecutive matches have a Poisson mean calculated as

$$\mu_i = v/u$$

- where u is the number of unique amino acid masses and v is the probability of a consecutive ion.
- the probability of a consecutive ion was measured on a sample data set as approximately v=0.4.
- the definition of consecutive ions was expanded to include complementary ions and ions that were consecutive but in different charge states.

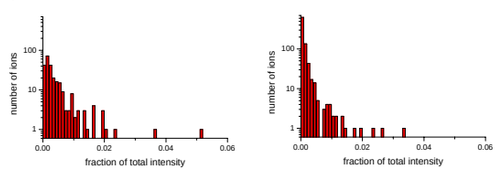
The model was tested on ribosomal proteins analyzed in an orbitrap. The analysis is summarized in the following ROC plot.



The correlation refinement significantly decreases the number of false positives for a wide range of true positive counts. It is likely that this effect is underestimated as it is possible that some of the false positives are in reality true positives since this sample was derived *in vivo*.

Intensity based scoring

Incorporation of ion intensities into probabilistic match based scoring has been impeded in part by wide variations in ion intensities from spectrum to spectrum. For example, note the differences in the intensity histograms for two typical spectra:



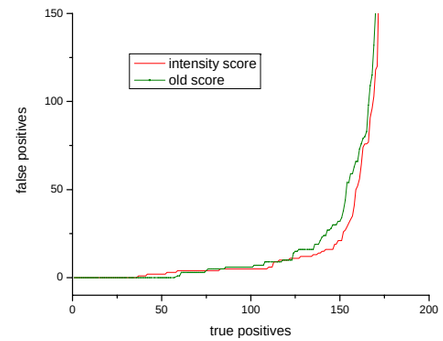
This observation is not surprising, given the complexity of some reaction pathways. This significantly impedes the ability to create a general statistical model of expected ion intensities. One solution is to use nonparametric statistics to transform the unknown distribution of peak intensities to a known distribution:

- Assign each experimental peak a rank. The sum of ranks of m randomly selected peaks from n total peaks is normally distributed with

$$\mu = (m(n+1))/2 \qquad \sigma = \sqrt{\mu(n-m)/6}$$

- Sum the ranks of all matched peaks and integrate the normal distribution from the value of the sum to infinity in order to compute a probability. Multiply this probability times the match probability.

The following is a ROC plot of false positives versus true positives for the 5 protein standard analyzed with ETD:



Note that there is an improvement in the false positive rate for a wide range of true positives found. The speed of the algorithm is significantly increased as it is no longer necessary to iteratively rescore spectra after adjusting a peak cutoff threshold, which was the indirect method for incorporating peak intensities.

Conclusions

- We have modified OMSSA to adjust for correlations between peaks in MS/MS spectra. This is particularly useful for high mass resolution mass spectrometers that only require a few matches for identification.
- Intensity based scoring has been probabilistically incorporated into the OMSSA match based statistics. This scoring can reduce false positive rates and decrease search time.
- A public search service and downloadable executables are available at http://pubchem.ncbi.nlm.nih.gov/omssa.

References

(1) Syka JE, Coon JJ, Schroeder MJ, Shabanowitz J, Hunt DF. Proc Natl Acad Sci U S A. 2004 Jun 29;101(26):9528-33.
(2) Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. J Proteome Res. 2004 Sep-Oct;3(5):958-64.