Modeling ETD Fragmentation with Bayesian Network for Peptide Identification

Xiaowen Liu[1]; Baozhen Shan[2]; Bin Ma[1]

[1]University of Waterloo, Waterloo, ON. [2]Bioinformatics Solutions Inc., Waterloo, ON.

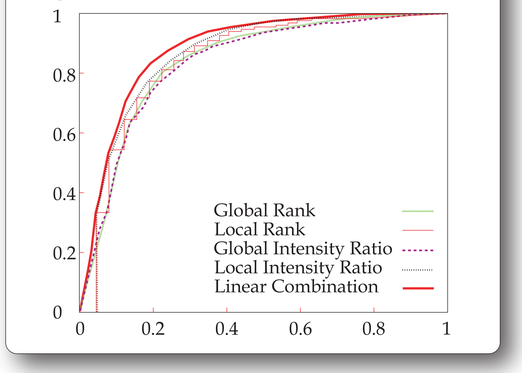
Introduction

In mass spectrometry–based protein identification and characterization, sequence assignment requires a successful MS/MS dissociation event, that is, production of a sufficient number of informative fragment ions. The recently introduced electron transfer dissociation (ETD) method has proven to be complementary to collision induced dissociation (CID) since it is better suited for sequencing larger, more basic peptides and is now becoming a more established technology. ETD spectra differ from CID spectra in the magnitude and the complexity of fragment ion signals; however, current software tools for peptide identification are usually better suited for CID data and give poorer results with ETD data. We present a Bayesian network model for peptide identification with ETD data. Preliminary data showed promising results of the model with PEAKS™ Studio software.

Methods

The absolute intensity of a peak does not directly reflect the likelihood of the peak being a signal. Before the training of the Bayesian network, each peak in the spectrum is assigned with a normalized intensity based on four features: (1) global rank, (2) local rank, (3) global intensity ratio, (4) local intensity ratio. The global rank of a peak is the number of peaks (in the same spectrum) that are higher than or equal to the current peak. The global intensity ratio is the intensity ratio between the highest peak (in the same spectrum) and the current peak. The local rank and local intensity ratio are defined similarly except that only the peaks within ±57 Da m/z difference to the current peak are examined. Then the final normalized intensity is equal to the linear combination of the logarithms of the four features. Note that a smaller value of the normalized intensity indicates a stronger peak. The coefficients of the linear combination were trained using approximately 1000 signal ion peaks and approximately 3000 randomly sampled background peaks. The ROC curves for z′-ion peaks of the four features and the linear combination are given in Figure 1. The figure illustrates that the linear combination improved the accuracy of distinguishing signal and noise peaks.

Figure 1. Comparison of ROC Curves



Depending on the fragment ion type, the distribution of the normalized intensity of the fragment ion peaks is different. Therefore, it is necessary to further convert the normalized intensity to a likelihood score for each ion type. For each fragment ion type, we used some training data to acquire the distribution of the normalized intensities of its peaks. Randomly generated positions were used to acquire the background distribution. Then we divided the ion distribution evenly into four intervals. On the centroid of each interval, the likelihood score was computed as log(signal probability/ background probability of the interval). The likelihood score for other normalized intensity values are computed with linear interpolation. (Figure 2).

Figure 2. Likelihood Scores

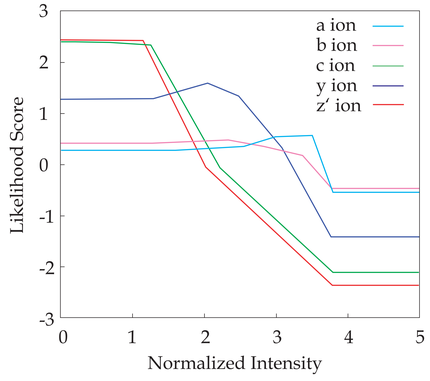
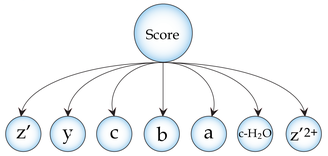


Figure 3. Bayesian Network



The probabilistic network is built based on seven common ions: z′, c, y, b, a, c-H_2O and z'^{2+}. The resulted network has a simple star topology (Figure 3). Consequently, for a cleaveage point, the likelihood scores of its corresponding ions are summed together to get its score. We also tried other Bayesian network topologies but no apparent improvement over this simple topology was achieved.

Data

The Bayesian network model was tested with an ETD dataset from a complex C.elegans protein mixture digested with trypsin followed by alkylation with iodoacetamide. The spectra were acquired on a LTQ Orbitrap XL ETD (Thermo, California). The peptide mixtures were separated with Surveyor LC equipped with MicroAS autosampler using a reversed phase peptide trap and a reversed phase analytical column at a flow rate of 250 nl/min. A gradient of 5~30% acetonitrile in 90 minutes was employed. PEAKS™ and Mascot software were used to identify peptides of the spectra from Swissprot database. We selected 259 spectra with confidently identified peptides. 130 of the spectrum-peptide pairs were used for training and the remaining 129 for testing.

Results

For each test spectrum-peptide, we randomly mutate the peptide sequence by replacing three consecutive residues with three other residues with the same total mass. If our model is good, then it should give the mutated sequence a lower score than the real sequence. By using the score function described above, 97.3% of the mutated sequences have scores lower than or equal to the real sequence.

We also compared our score function with that of PEAKS™ Studio 5.0. We used PEAKS™ Studio 5.0 to do de novo sequencing for each test spectrum. From the resulting peptide of PEAKS™ Studio 5.0, we use a local search method to find a better peptide based on our score function. PEAKS™ Studio 5.0 was able to correctly compute 40.6% of all the amino acids in the test peptides. Our strategy improved this to 48.4%.

Availability

The new score function will be included in a future version of the PEAKS™ software.