Optimization of a Model of Peptide Fragmentation for MALDI Q-TOF MS/MS Data

Helen Bruno[1], Thomas Eadsforth[2], Emmanuelle Claude[1], Marten Snel[1], Richard Denny[1], Keith Richardson[1], John Skilling[3], James Langridge[1], Therese McKenna[1] and Phillip Young[1]
[1] Waters Corporation MS Technologies Center, Float Road, Wythenshawe, Manchester M23 9LZ, UK [2] University of Manchester Institute of Science and Technology, Physical Methods For Bioanalysis and Post-Genome Science
[3] Maximum Entropy Data Consultants Ltd, Tresawsan, Killaha East, Kenmare, Ireland

Overview

The ProbSeq fragmentation model (Skilling, 2000a, 2000b) is used for databank searching and *de novo* sequencing MS/MS data produced by collision-induced dissociation (CID) of peptides. In a previous study, the model was successfully optimized for the peptide fragmentation data generated from multiply charged precursor ions using an ESI Q-Tof (Waters, Manchester, UK) instrument (Skilling *et al.*, 2004).

This poster describes the work carried out to optimize the model for MALDI Q-TOF MS/MS data. The model was tuned using the MS/MS data generated from tryptic peptides produced from known protein standards. The tuning algorithm optimized the prior probabilities of over a hundred different types of bond breakages.

The optimized ProbSeq model was tested against a distinct validation set of characterized fragmentation data. It is shown that for both databank searching and *de novo* sequencing the optimized model improved the likelihood of the correct sequence for the majority of peptides in the validation set.

Introduction

MALDI Q-TOF MS/MS
Traditionally, peptides are characterized via peptide mass fingerprinting using MS data from a MALDI TOF instrument. Ambiguous assignments of peptide masses may be investigated further using an ESI Q-TOF mass spectrometer to generate MS/MS data for fragment ion searches.

Using a MALDI Q-TOF instrument to carry out both MS and MS/MS experiments on peptides is becoming increasingly common. Using a single instrument requires only a single sample to be prepared and the MALDI ion source enables MS/MS experiments to be repeated several times for the same sample.

The MS/MS data produced by a MALDI Q-TOF mass spectrometer is considerably different from that produced by fragmenting multiply charged ions with an ESI Q-TOF. This can be attributed to the fact that a MALDI source primarily generates singly charged ions. Its MS/MS spectra contain a mixture of y''-, b- and a- ions compared to ESI MS/MS spectra, which predominantly contain y''-ions and some low mass b-ions (Wattenberg A *et al.*, 2002).

Due to the marked difference in its singly charged product ion spectra compared to that produced using an ESI source, we decided to tune ProbSeq for this type of configuration of ion source and mass spectrometer.

ProbSeq
We use the ProbSeq model to calculate the likelihood that a candidate sequence could give rise to a particular spectrum. For example, in *de novo* peptide sequencing, this likelihood function is used in a Bayesian calculation to determine the probability of a sequence for a given MS/MS spectrum:

$$\Pr(Sequence \mid Spectrum) = \Pr(Sequence) \times \Pr(Spectrum \mid Sequence) \;/\; \Pr(Spectrum)$$

Equation 1

Where

$\Pr(Sequence \mid Spectrum)$ is the posterior probability of a sequence for a given spectrum

$\Pr(Sequence)$ is the prior probability of a sequence without looking at the spectrum, which takes into account the natural abundance of amino acids

$\Pr(Spectrum \mid Sequence)$ is the likelihood of a spectrum for a given sequence, this is calculated using ProbSeq

$\Pr(Spectrum)$ is the evidence, which is a normalization factor

The ProbSeq model effectively sums the probabilities associated with each possible fragmentation pattern a sequence could theoretically produce. For each fragmentation pattern the product of the probability it could give rise to the observed spectrum and the probability a sequence would produce the pattern is calculated:

$$\Pr(Spectrum \mid Sequence, \psi) = \sum_{Frag} \Pr(Spectrum \mid Frag) \times \Pr(Frag \mid Sequence, \psi)$$

Equation 2

Where

$\Pr(Spectrum \mid Frag)$ is the probability of a spectrum for a given peptide fragmentation pattern

$\Pr(Frag \mid Sequence, \psi)$ is the probability of a fragmentation pattern for a given sequence and probabilistic information about peptide fragmentation

Instead of explicitly summing the probabilities for each fragmentation pattern, ProbSeq uses a Markov chain in order to reduce calculation times. The Markov chain allows the summation of all possible fragmentation pattern probabilities to be calculated in a time proportional to the number of amino acids in the sequence. For further details please see Skilling *et al*, 2004, where the use of the Markov chain by the ProbSeq model is discussed in greater detail.

In Equation 2, the symbol Ψ represents prior knowledge of peptide fragmentation. This is actually a database of over 100 probabilities. These probabilities consist of Markov chain parameters and the probabilities for different bond breakages for individual amino acids. It is these probabilities that were tuned using the MS/MS data generated by a MALDI Q-TOF mass spectrometer.

Methods

MS/MS Data Acquisition
The tuning exercise required MS/MS data for more than fifty peptides. The peptides used to generate the MS/MS data were obtained from the tryptic digest of five protein standards. These standards were:

Yeast Alcohol Dehydrogenase (SwissProt P00330), Bovine Serum Albumin (SwissProt P02769, Rabbit Glycogen Phosphorylase B (SwissProt P00489), Bovine Hemoglobin (SwissProt HBA P01966, HBB P02070), Yeast Enolase (SwissProt P00924)

Each standard was the product of a tryptic digest with subsequent purification to remove any by-products. The standards were obtained in a lyophilized form and were reconstituted prior to analysis at the appropriate concentration to provide good quality MS/MS product ion data.

A Waters Micromass Q-Tof Ultima™ MALDI (Waters, Manchester, UK) instrument was used to produce product ion data for the peptides in the five standards. The software application, MassLynx™ (Waters, Manchester, UK), was used to automate the collection of both the MS survey and MS/MS data. The Q-Tof was operated in the data directed analysis (DDA) mode where peptides identified in the MS survey had to pass a set of criteria before being analyzed in the product ion MS/MS mode.

Peptides were fragmented if the molecular ion was clearly visible in the MS survey spectrum and not within 3 m/z of any other peaks. It was ensured that a wide range of peptide masses were selected for fragmentation and the optimal collision energy for fragmentation, determined from the m/z, was used for the CID experiments.

All the MS/MS data acquired was processed using Masslynx. This consisted of the background subtraction of noise and deconvolution of the spectra into single monoisotopic peaks using the MaxEnt3 algorithm (Waters, Manchester, UK).

A theoretical amino acid sequence was predicted for each MS/MS spectrum, using an *in-silico* tryptic digestion of the theoretical protein sequence. The precursor ion masses selected for fragmentation were then matched to the theoretically derived mass and hence sequence information. Each MS/MS spectrum was validated against its expected sequence using BioLynx (Waters, Manchester, UK) and Mascot (Matrix Science, London, UK).

Tuning ProbSeq
Skilling *et al.*, 2004, developed an automated method that optimized ProbSeq for ESI Q-TOF MS/MS data. We were able to utilize this computer application to tune the model for a MALDI Q-TOF instrument. The only change required was to provide the program with MS/MS data from the MALDI Q-TOF mass spectrometer.

Two sets of data were used in this work; a tuning set and a validation set. The tuning set consisted of forty-five fully characterized MS/MS spectra from known peptides. The tuning program used this data to generate a new database of prior probabilities, Ψ.

The probabilities that were tuned included Markov chain parameters and the various bond breakages that can occur between amino acid residues and also their side-chain cleavages. For each amino acid, a maximum of seven probabilities were explored:

- Cleavage on the C-terminal side and N-terminal side.
- Loss of H_2O and NH_3 from the side chain.
- Loss of CO and NH_2 from the backbone.
- Immonium ion produced.

Results

The tuned values were used to update the ProbSeq model, which was then tested against the validation data set. Testing the ProbSeq model consisted of performing databank searches and *de novo* sequencing for every MS/MS spectra in the validation set. For each search the difference in the log likelihood for the top candidate and the correct sequence was calculated. These values were compared for the ProbSeq model before and after tuning.

Figures 1 and 2 illustrate the difference between the log likelihood's of the top candidate sequence and the correct sequence for each of the peptides in the validation set. The charts compare the results using the original and the tuned ProbSeq model.

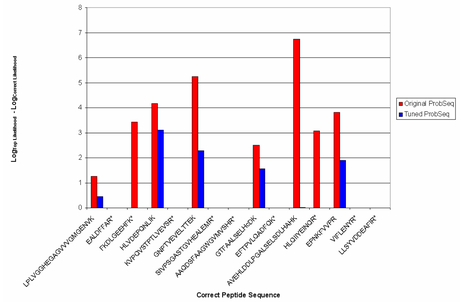

Figure 1: Difference in the log likelihood of the top candidate peptide sequence and correct peptide sequence for databank search results.

The ideal result is for the difference in the log likelihood's to be zero. That is, the correct sequence has the highest likelihood of all those explored. If the correct sequence is the top candidate generated by a search using the tuned ProbSeq model it is denoted by *.

Figure 1 displays the databank search results for the validation set. On the whole we observe an improvement in the likelihood of the correct sequence when using the tuned model. The most significant improvement is for the peptide AVEHLDDLPGALSELSDLHAHK (residue 22 from Bovine Hemoglobin), where the difference in its log likelihood compared to the top candidate decreases from 6.74 to 0.02.

Improvements in the likelihood of the correct sequence for the *de novo* results were not as dramatic. MALDI QTOF MS/MS spectra can contain prominent y''-, b- and a- ions (Wattenberg A *et al.*, 2002). In principle this could aid the discrimination of the correct sequence from the other candidate sequences. However, in practice, the abundance of data may confuse the *de novo* algorithm as they can be accounted for in many different ways by competing peptide sequences.

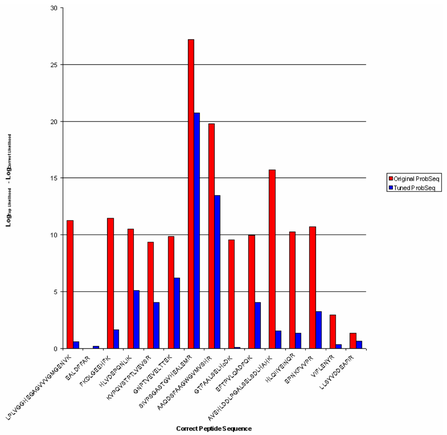

Figure 2: Difference in the log likelihood of the top candidate peptide sequence and correct peptide sequence for de novo sequencing results.

Conclusions

- The purpose of this work was to optimize the peptide fragmentation model, ProbSeq, for the interpretation of MS/MS product ion data from a MALDI Q-TOF mass spectrometer.
- Fully characterized MS/MS data from a MALDI Q-TOF instrument were used to tune the model using an automated method developed by Skilling *et al.*, 2004.
- It has been shown that the likelihood of the correct sequence in relation to the top candidate improved when using the tuned model for both databank searching and *de novo* sequencing MALDI MS/MS spectra.
- This will have a significant impact on the utility of MALDI MS/MS spectra in the core Proteomics laboratory.
- Further improvements could be made to the ProbSeq model by tuning for other types of mass spectrometer, whose fragmentation patterns differs significantly from MALDI Q-TOF MS/MS and ESI Q-TOF MS/MS data - for example MALDI PSD
- Finally, it would be advantageous to tune the model using data generated using digest reagents other than trypsin.

REFERENCE
1. Skilling J. 2000. US Patent No 6489608.
2. Skilling J. 2000. US Patent No 6489121.
3. Skilling J., Denny R., Richardson K., Young P., McKenna T., Campuzano I., Ritchie M. 2004. Comparative and Functional Genomics **5**: 61-68.
4. Wattenberg A., Organ A.J., Schneider K., Tyldesley R., Bordoli R., Bateman R.H. 2002. American Society for Mass Spectrometry **13**: 772–783.