MSMS Peak Identification and its Applications

Deep Jaitly[1*], Rachel Page Belanger[1], Denis Faubert[1], Pierre Thibault[1], Paul Kearney[1]

[1]Caprion Pharmaceuticals Inc, 7150 Alexander –Fleming, Montreal, QC, H4S 2C8

Received line

ABSTRACT

A peak detection algorithm for Tandem Mass Spectra is presented that scores a fragment using intensity and isotopic distribution. It classifies each fragment in a spectrum as noise or signal based on a maximum likelihood estimate derived from the distribution observed in a training set of 12,000 validated spectra. This is the largest such database known to the authors. We present three tools which apply this algorithm: the Quality Filter removes noisy spectra, Mod-Pro profiles modifications and amino acids in a sample and Spectrimilarity scores similarity of two spectra.

Contact: njaitly@caprion.com

Keywords: Mass Spectra, Isotope Distribution, Peak Scoring, Quality Filter.

INTRODUCTION

Tandem Mass Spectrometry has become a major analytical method in Proteomics. By analyzing tandem mass spectra, peptides and post-translational modifications (PTMs) of proteins can be identified. For example, the Human Proteome Organization's Proteome Project is dependent upon the acquisition and interpretation of MSMS data (Human Proteome Organization, 2004).

Proteins in a sample are digested with an enzyme such as trypsin, giving peptides which are amino acids joined together in series by peptide bonds. A tandem mass spectrometer fragments peptides stochastically along peptide bonds resulting in different types of ions (Eng *et al.*,1994). It then detects the mass to charge ratios (m/z) and intensities of the fragments, giving a mass spectrum for the peptide. The spectra can be interpreted by looking for peaks separated by the mass of an amino acid and/or PTMs. The peptide can be inferred by determining a ladder of amino acids in the spectra (figure 1).

Interpretation is complicated by several factors. The presence of significant amount of electrical and chemical noise, side-chain fragmentation, isotope peaks, etc, makes it difficult to determine which peaks correspond to fragment ions. Intensity can be used to classify peaks. Previous approaches used the intensity of a peak (relative to local maximum of intensity) to determine if a peak was significant or not (Eng *et al.*, 1994; Sadygov *et al.*, 2002). However, relative intensity varies over m/z and by itself is not a good enough indicator of significance of a peak. Isotope peaks are another complicating factor. Peptide fragments are usually detected not just as one peak but as two or three isotope peaks separated by an m/z of 1/*ch* Da in the spectra, where *ch* is the charge of the fragment. The relative intensities of these peaks depend on the composition of the fragment. Gay et. al. used a theoretical multinomial distribution to score the isotopes for peptides in peptide mass fingerprinting (Gay *et al.* 1999). However, the fragments in tandem mass spectra are of low intensity which makes it harder to observe the correct theoretical ratios of the isotopes. Moreover the model does not provide a score for noise.

We present a peak detection algorithm using a peak model based on intensity, isotopic ratios and m/z. We then present three MSMS data analysis tools that use this engine. Not only does the peak detection algorithm significantly improve the accuracy of these tools, but these tools also provide the foundation for large scale, high throughput proteomics.

PEAK IDENTIFICATION THROUGH A MAXIMUM LIKELIHOOD ESTIMATE

Let $\boldsymbol{r} \in Z^n$, be the ordered set of intensities of the first *n* isotopes of a fragment. For our purposes we used *n = 3*. Let the intensity, *int*, be the intensity of the first isotope, *ch* be the charge, *mz* be the *m/z* ratio. Let *type=peak*, denote the event that a given fragment is a real peptidic ion and let *type=non-peak*, denote the event that a given fragment is not a real peptidic ion.

Then, P(***r**,int*|*type=peak,mz,ch*) represents the probability that a real peak of *mz=m/z* and charge=*ch* would have an intensity, *int*, and isotopic ratio ***r***. Correspondingly, P(***r**,int*|*type=non-peak,mz,ch*) represents the probability that a non-peak fragment of *m/z=mz* and charge=*ch* would have an intensity, *int*, and isotopic ratio ***r***.

Thus, given *m/z*, ***r***, and *int*, we can get the charge and type of the fragment from the maximum likelihood estimate as follows:

$\text{type}^* = \arg\max_{\text{type}} P(\mathbf{r},\text{int}|\text{type},\text{mz},\text{ch})$ and $\text{charge}^* = \arg\max_{\text{ch}} P(\mathbf{r},\text{int}|\text{type},\text{mz},\text{ch})$.

We define Peak Intensity & Isotopic Ratio Significance or PIIRS (pronounced *peers*) as:

[*] To whom correspondence should be addressed.

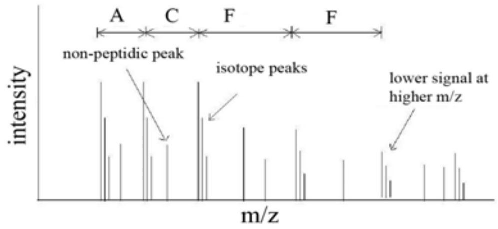

Figure 1. A peptide can be sequenced by inspecting the difference in mass between peaks.

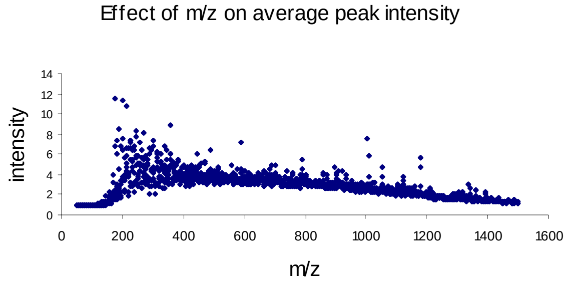

Figure 2. Interpretation of spectra is complicated by the effect of m/z on background noise.

PIIRS(***r*,***int,mz,ch*)=

$$\log \frac{P(\vec{r}, \text{int} \mid type = peak, mz, ch)}{P(\vec{r}, \text{int} \mid type = non - peak, mz, ch)}$$

TRAINING AND RESULTS

We used an internal database of over 12,000 manually validated spectra for training and a different set of 1200 manually validated spectra for testing. From the sequence assigned to each spectrum we were able to label peaks as *peak* or *non-peak*, and use this to train our model. The intensities of fragments in spectra were normalized before training. We developed two normalization procedures, *Background-Average (BA)*, and *Background-Median (Bmed).* In *BA* and *Bmed* peaks were normalized against the local average intensity and local median intensity respectively. We compared them against a commonly used normalization, which we are calling *Background-Maximum (BMax)*, where peaks are normalized relative to the local maximum intensity (Eng *et al.*, 1994; Sadygov *et al.*, 2002). Over 90% sensitivity, and 90% specificity was achieved in classifying peaks in a test set of 1200 spectra. Figure 3 presents a comparison of the Receiver-Operator-Characteristic (ROCs) using the three different normalization procedures. While *BMax* seems to perform as well as *BA*, *BMed* on clean data, it lacks robustness when applied to noisy data (see the section 4).

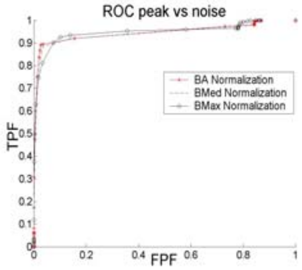

Figure 3. ROC for PIIR peak detection on a testing set of 1200 spectra.

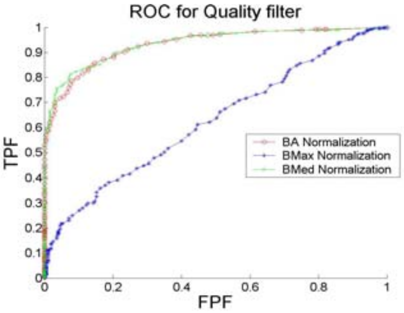

Figure 4. ROCs of Quality Filter using different normalizations. *BMax* is not very robust as a normalization procedure with PIIR.

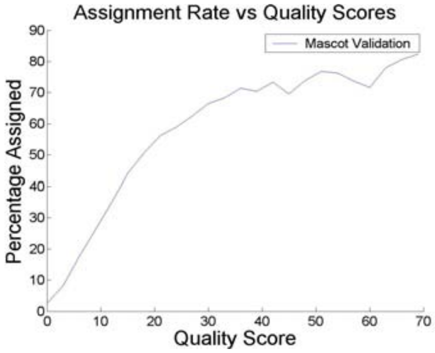

Figure 5. Assignment Rate vs Quality. 26,180 spectra were assigned sequences by MASCOT and semi-automatic validation performed using Caprion Filter. Assignment rate over the entire set (all quality score bins) with Mascot was approximately 40%.

4. MSMS TOOLS USING PIIRS

QUALITY FILTER

In a large scale project where thousands of spectra are acquired, spectra need to be prioritized for analysis based on their quality. Let *AAMass* be the set of amino acid masses (enumerated with and without PTMs), *fragments(sp)* be the set of fragments in spectrum, *sp*, m_i, p_i be the *m/z*, and the PIIR score of the *i*th peak in *sp*. Then, we define the quality of a spectrum, *sp* as

Quality(sp)=

$$\sum_{(\mathrm{i,j}) \in \{(i,j) \mid i,j \in fragments(sp), (m_i - m_j) \in AAMass\}} (p_i + p_j + continuity\_adjustment)$$

Informally, we define quality of a spectrum as the sum of PIIRS scores of peaks that correspond to amino acid tags. Continuous tags were rewarded more than discontinuous tags and overlapping tags were filtered out. Figure 4 shows ROC curves for the performance of the quality filter in classifying a set of 1000 manually labelled spectra. Sensitivity of over 99% was achieved with specificity of 75%. We can also see how *BMax* lacks the robustness of

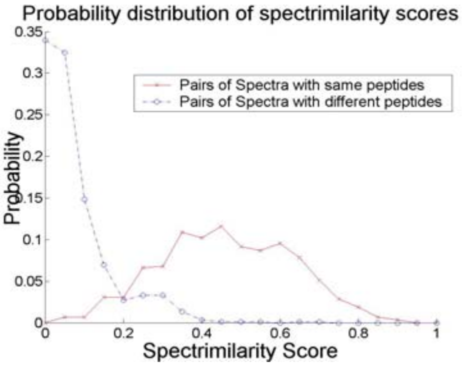

Figure 6. *Spectrimilarity* Score distribution between pairs of peptides from the same or different spectra.

BMed and *BA* for this tool. Figure 5 shows the relationship between assignment rates and *Quality*.

SPECTRIMILARITY

In an experiment, the same spectra might be acquired several times. It is useful to group redundant spectra to shorten interpretation time and also to choose a highest quality representative spectrum to interpret. A similarity score between spectra serves as a starting point. *NoDupe* defined a similarity metric which was the dot product of the intensities of the peaks of the two spectra (after preprocessing) (Tabb *et al.*, 2003). We define similar metrics,*spectrimilarity*, and *spectrimilarity_prob*, using *PIIR* scores rather than intensities. Given two spectra, sp_1 and sp_2, with fragments fr_{1i}, fr_{2i} , *spectrimilarity* is defined as follows:

spectrimilarity$(sp_1, sp_2) =$

$$\frac{\sum\limits_{mz(fr_{1i})=mz(fr_{2j})} score(fr_{1i})*score(fr_{2j})}{\sqrt{\sum score(fr_{1i})^2}\sqrt{\sum score(fr_{2j})^2}}$$

spectrimilarity_prob$(sp_1, sp_2)=$

$$\frac{\sum\limits_{mz(fr_{1i})=mz(fr_{2j})} score(fr_{1i})*score(fr_{2j})/\,p(mz(fr_{2j}))}{\sqrt{\sum score(fr_{1i})^2/\,p(mz(fr_{1i}))}\sqrt{\sum score(fr_{2j})^2/\,p(mz(fr_{2j}))}}$$

Figure 6 compares the *spectrimilarity* scores between pairs of spectra from the same peptide against the *spectrimilarity* scores between pairs of spectra from different peptides. Sensitivity of 95% and specificity greater than 92% was acheived at a threshold value of 0.2.

MOD-PRO

Global profiling of spectra can be done by looking for frequently observed mass differences in the spectra. We are also able to compare profiles from different samples to observe global differences such as modifications and amino acid compositions. For this, the spectra are preprocessed to remove peaks with low PIIR scores. The mass difference between each pair of peaks is tracked and a frequency diagram is constructed from these mass differences. For example, figure 7 compares the profile from a set of phospho-peptides against the profile from a set of colon cancer cell lines. By studying a Mod-Profile we are able to know which modifications to expect before conducting any protein identification searches.

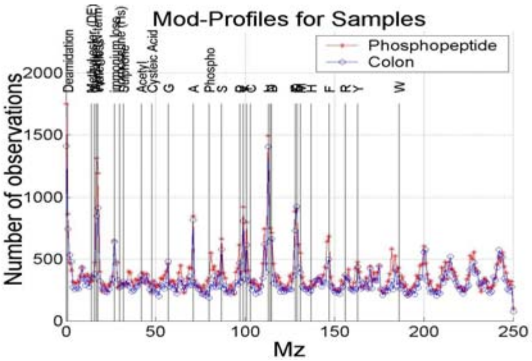

Figure 7. Mod-Profiles of a colon cancer sample, and a phosphopeptide sample

5. REFERENCES

Eng,J.K., McCormack,A.L., Yates III,J.R. (1994) An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. J. Am. Soc. Mass Spectrom, **5**, 976-989.

Human Proteome Organisation. HUPO–The Human Proteome Organisation. Jan. 2004. <http://www.hupo.org/hpp/hppp.htm>

Gay,S., Binz,P., Hochstrasser,D.F., Appel,R.D. (1999) Modeling peptide mass fingerprinting data using the atomic composition of peptides. Electrophoresis, **20**, 3527-3534.

Perkins,D.N., Pappin,D.J.C., Creasy,D.M., Cottrell,J. S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, **20**, 3551-3567.

Sadygov,R.G., Eng,J.K., Durr,E., Saraf,A., McDonald, H., MacCoss,M.J., Yates III,J.R. (2002) Code Developments to Improve the Efficiency of Automated MS/MS Spectra Interpretation. J. Proteome Res., **1**, 211-215.

Tabb,D.L., MacCoss,M.J., Wu,C.C., Anderson,S.D., Yates III,J.R. (2003) Similarity among Tandem Mass Spectra from Proteomic Experiments: Detection, Significance and Utility. Anal. Chem., **75**, 2470-24