Multi-scale Data Reduction Algorithm of Proteomic Mass Spectrum
M Nafati, M Samson, B Rossi

Abstract
The proteomic is a field which makes it possible to connect the sequence of the genome and the cellular behaviour. The proteomic analysis can be done in various stages: preparation of the samples, separation of proteins, analysis by mass spectrometry, preprocessing (data mining) and interrogation of the data banks. The mass spectrometry measures the mass of peptides (typically obtained by tryptic digestion). These masses are then compared to those theoretical in Databases in order to identify the protein name. Electronic and chemical noise is often the source of bad identification. In this article, we propose an original data reduction algorithm with the aim of removing the spectra baseline, then removing parasitic mass peaks and amplifying those useful. The algorithm principle uses the dyadic multi-resolution technique (bio-orthogonal decomposition/reconstruction) coupled to the fuzzy logic thresholding. In order to evaluate the quality of this algorithm, we present a comparison of the results obtained by our algorithm and those obtained using the data reduction software of MALDI-TOF spectrometer (Matrix-Assisted Laser Desorption/ionization).

INTRODUCTION
Proteomic analysis is done primarily by the use of the two-dimensional electrophoresis (2-DE) technique coupled with the Mass Spectrometry (MS) analysis. The first technique helped by the proteomic imaging leads to the localization of the candidates proteins for mass spectrometry analysis. The comparison between the spectra of masses obtained and those theoretical of DataBase leads to the identification of proteins of interest in term of peptides or amino acids [6,10].

In this paper, one propose an robust data reduction algorithm of mass spectra based on the multi-resolution technique [2,5,7,9] and the fuzzy set theory [3,4,11,12,13]. The idea is to separate the mass peaks into groups of dyadic sub-bands and then thresholding the high frequencies sub-band. The computation of the optimal threshold is done by minimizing the fuzzy Shannon entropy. The result is then amplified in an adaptive way. At the end of the process, the mass spectra is reconstructed and corrected by removing the baseline signal[1].

PROBLEM FORMULATION
The currently most common method to identify proteins is to first enzymatically digest the proteins, then determine the masses of result peptides by peak detection on a MALDI-TOF spectrum [10], and finally use the peptide mass fingerprints to research protein sequences. The found theoretical protein is that which gives a maximum rate of covering. It is clear that this result depends mainly on the quality of the mass spectrum. Consequently the data reduction processing is a primodial stage since the presence of (electronic and/or chemical) parasitic, or the absence of useful mass peaks distorts the result of the protein identification. As sometimes, only a few experimental peptide masses in the fingerprint match the theoretical masses in a database, failure to detect one peak can hinder the correct identification of a protein. The standard data reduction software (DataExplorer Voyager) provided with MALDI spectrometer, is often unoptimal and nonadaptive, it consists of doing these processes: denoising, baseline correction, thresholding, peak detection, protein identification. Here, our data reduction algorithm aim to optimize the denoising and baseline correction processes, and to improve the SNR ratio in an adaptive way.

OBJECTIVE DATA REDUCTION ALGORITHM
The global architecture of the proposed algorithm is:

Step 1: Dyadic sub-band decomposition
Step 2: High frequencies (HF) optimal thresholding
Step 3: Enhancement of the thresholded HF.
Step 3: Spectra Reconstruction
Step 4: Optimal Baseline correction
Step 5: Peak detection
Step 6: Protein identification

The dyadic sub-band decomposition is made in the following way:

**Figure 1**
Figure 1: sub-band decomposition.

The denoising difficulty resides in the fact that the noise is present in the upper and lower sub-band. It is the case of the MALDI spectra. In addition to the electronic noise, one finds the chemical noise. This is why, each sub-band on a given level is decomposed to HF and LF sub-band [2,6,7,8]. At each pyramid level, high frequency sub-band is thresholded by minimizing the fuzzy Shannon entropy. Then the spectra are reconstructed. It’s clear that the decomposition/reconstruction process should be perfect. To answer to this question, we have chosen a bio-orthogonal filter bank.

The optimal threshold computation process is found by first, defining a membership function is:

**Figure 2**

The second step is to determine a measure of the fuzziness at a given threshold t. One method for measuring fuzziness is based on the idea of Shannon Entropy [3,4,5,11]:

**Figure 3**

\[ \mu_0 (t) = \frac{\sum_{i = \text{min}}^{\text{max}} \mu_i (t) h(i)}{\sum_{i = \text{min}}^{\text{max}} h(i)} \]  
and  
\[ \mu_1 (t) = \frac{\sum_{i = \text{min}}^{\text{max}} \mu_{i+1} (t) h(i)}{\sum_{i = \text{min}}^{\text{max}} h(i)} \]

Where t is a given threshold level, C is a constant that represents the difference between the maximum (fmax) and minimum (fmin) high frequencies, \( \mu_0 \) and \( \mu_1 \) are the mean values of the upper and lower classes and h being the histogram.

The optimal threshold value is that minimizes E(t). Then the useful high frequencies are amplified by a factor G such as:

**Figure 4**

\[ H_t (x) = -x \ln(x) - (1-x) \ln(1-x) \]

The Shannon Entropy of the entire spectra is:

**Figure 5**

\[ E(t) = \frac{1}{N} \sum_{i = \text{min}}^{\text{max}} H_t (\mu (i)) . h(i) \]

The optimal threshold value is that minimizes E(t). Then the useful high frequencies are amplified by a factor G such as:

**Figure 6**

\[ G = \frac{\sigma_{\text{total}}}{\sigma_{\text{local}}} \]

Where, \( \sigma_{\text{total}} \) is the HF sub-band standard deviation (std), and \( \sigma_{\text{local}} \) is the current window std of the HF sub-band. After the reconstruction process, the baseline spectra are removed according to the concept provided by Golotvin [1]. Among N points the minimal and maximal values are found. If their difference does not exceed the noise std multiplied by a definite factor n (\( Y_{\text{max}} - Y_{\text{min}} < n \sigma_{\text{noise}} \)), the i-th point is considered to belong to baseline.

**ALGORITHM PRINCIPLE**
RESULTS

DATAEXPLORER REDUCTION SOFTWARE RESULTS

The raw masses spectrum given in Fig. 1, is that of a known protein coming from the rat species. It has been identified as “Acyl-CoA dhydrogenase” protein.

Figure 8

Figure 2: “Acyl-CoA dhydrogenase” protein raw spectrum protein coming from the rat species.

This latter spectra preprocessed with DataExplorer Software leads to this following result

The found masses compared to those theoretical contained in the SwissProt Database lead us to the protein identification given in fig. 4.

Figure 10

Figure 4: The MsFit software identification results.

One notices that the protein candidate is identified with a score of 3.586.10^4, a rate of covering (cov) of 36%, a mass precision of 29.7 ppm.

OUR DATA REDUCTION ALGORITHM RESULTS

The obtained preprocessed mass spectrum is:

Figure 11

Figure 5: Mass spectrum obtained with our algorithm.

The optimal threshold computation is done block by block.
The size of each block is 40. The optimal threshold values (block per block) calculated for the HF sub-band at level one are given in the following figure:

**Figure 12**
Figure 6: HF sub-band threshold values at level one.

![HF sub-band threshold values at level one](image)

**Figure 13**
Figure 7: Amplification gain corresponding to Fig.5.

![Amplification gain](image)

The matching result between experimental and theoretical masses is given in Fig.8. The database used is SwissProt.

![Theoretical masses matched in SwissProt Database](image)

One notices that the protein candidate is identified with a score of $1.63 \times 10^9$, an overlapping rate (cov) of 51%, a mass precision of 22.4 ppm.

**CONCLUSION**

Protein identification and characterization is one of the most essential tasks performed in proteome research. The precise determination of the peptide masses in the spectra, and highly discriminating mass comparison algorithm are therefore the keys to accurate identification of proteins. We have developed a precise and objective preprocessing algorithm. Often, the thresholds analysis associated with the peak detection is revealed that is preferable to be little selective in the choice of peaks in the mass spectrum, this is in order to avoid the loss of apparently fictitious peaks that might eventually appear to be useful. Our algorithm preprocessing is done to be more selective and precise regarding the masses peak determination. Indeed, the multiscale fuzzy thresholding is revealed as an objective tool regarding the peak selection. Obtained results confirm this, the score and the cov coeff are improved significantly. Introducing the fuzzy Shannon Entropy in multiscale concept is therefore an interesting idea.

**References**

3. H. Haußecker, H. R. Tizhoosh, Fuzzy Image Processing,
Multi-scale Data Reduction Algorithm of Proteomic Mass Spectrum

Author Information

M. Nafati
IFR50, Proteom Plate-Form, Medical University, Av Valombrose

M. Samson
IFR50, Proteom Plate-Form, Medical University, Av Valombrose

B. Rossi
IFR50, Proteom Plate-Form, Medical University, Av Valombrose