Proceedings of the 5th WSEAS Int. Conf. on Signal Processing, Computational Geometry & Artificial Vision, Malta, September 15-17, 2005 (pp222-226)

Objective Data Reduction Algorithm of Proteomic Mass Spectrum

M. Nafati, M. Samson, and B. Rossi
IFR50, Proteom Plate-Form
Medical University, Av Valombrose, 06172, Nice, France
phone: (33) 493377691
web: www.unice.fr

Abstract: Proteomic analysis is done primarily by the use of the two-dimensional electrophoreses (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. The presence of parasitic and/or the absence of useful mass peaks distort(s) the result of the identification process. 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 muli-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).

Key-Words: Protemic. Spectra of masses. Data reduction. Multi-resolution. Fuzzy logic thresholding.

1 Introduction

The proteomic [6,10] is a field which makes it possible to connect the sequence of the genome and the cellular behavior. 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) [10]. These masses are then compared to those theoretical in Databases in order to identify the protein name. Electronic and chemical noise are often the source of bad identification [6].

In this document, we propose an objective data reduction algorithm of mass spectra based on the multi-resolution technique [2,6,7,9] 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. This is 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].

2 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 [14], and finaly use the peptide mass fingerprints to reseach 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 databases, 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.

3 Objective data reduction algorithm

The global architecture of the proposed algorithm is:

Step1: Dyadic sub-band decomposition
Step2: High frequencies (HF) optimal thresholding
Step3: Enhancement of the thresholded HF.
Step3: Spectra Reconstruction

Step4: Optimal Baseline correction
Step5: Peak detection
Step6: Protein identification

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

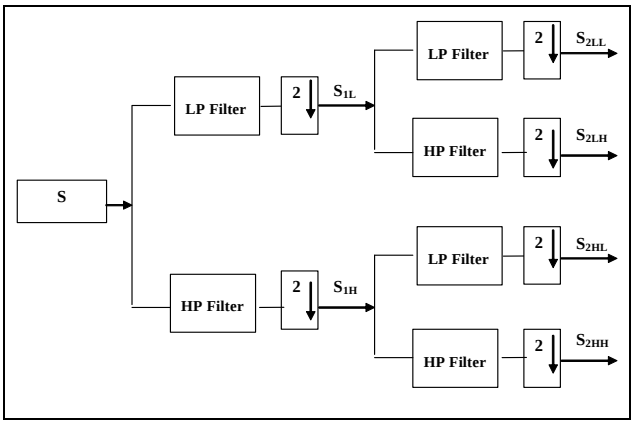

Fig.1 : sub-band decomposition.

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, ones finds the chemical noise. Ihis 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 is 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 :

$$\mu(f) = \begin{cases} \dfrac{1}{1+|f-\mu_0|/c} & if \quad f \le t \\ \dfrac{1}{1+|f-\mu_1|/c} & if \quad f > t \end{cases}$$

With

$$\mu_0(t) = \frac{\sum_{i=i\min}^{t} i.h(i)}{\sum_{i=i\min}^{t} h(i)} \qquad and \qquad \mu_1(t) = \frac{\sum_{i=t+1}^{i\max} i.h(i)}{\sum_{i=t+1}^{i\max} h(i)}$$

Where t is a given threshold level, C is a constant that represents the difference between the maximum (fmax) and minimum (fmin) high frequencies, μ_0 and μ_1 are the mean values of the upper and lower classes and h being the histogram .

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]:

$$H_t(x) = -x\ln(x) - (1-x)\ln(1-x)$$

The Shannon Entropy of the entire spectra is:

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

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

$$G = \frac{\sigma_{loctal}}{\sigma_{local}}$$

Where, σ_{total} is the HF sub-band standard deviation (std), and σ_{local} is the current window std of the HF sub-band.

After the reconstruction process, the baseline spectra is 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 ($Y\max - Y\min \le n.\sigma_{noise}$), the *i-th* point is considered to belong to baseline.

4. Results

4.1 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.

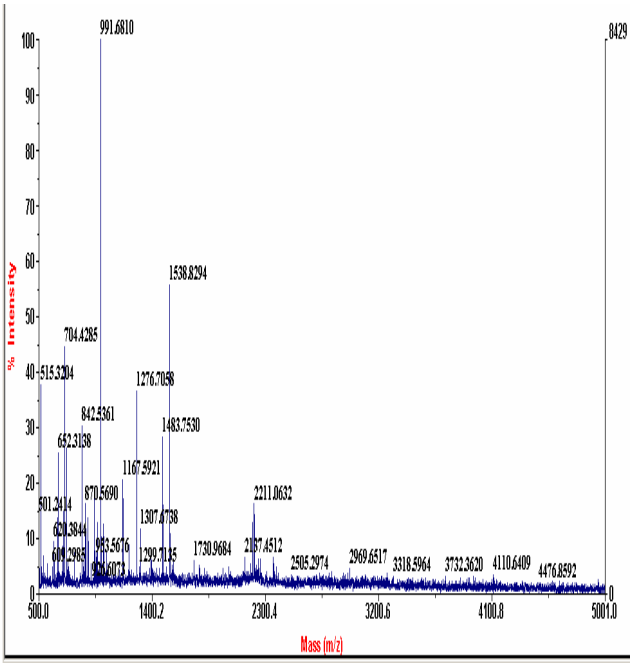

Fig.2 "Acyl-CoA dhydrogenase" protein raw spectrum protein coming from the rat species.

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

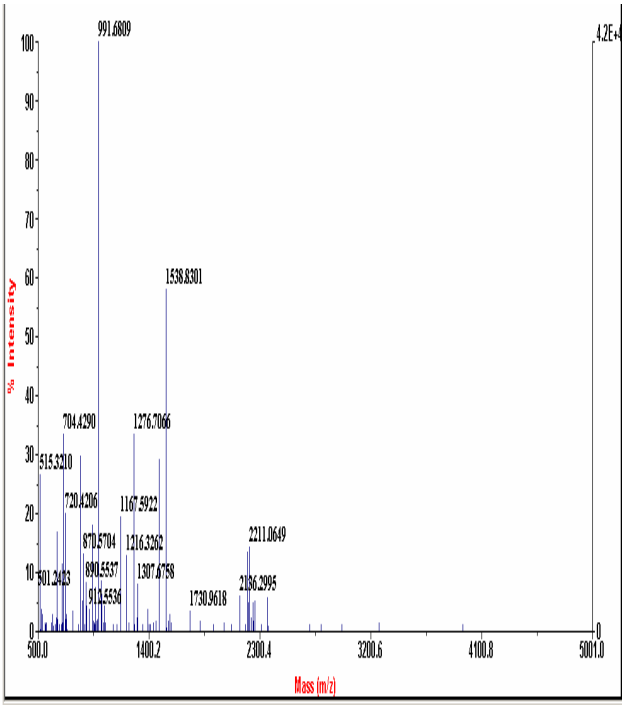

Fig.3: mass peak results obtained with DataExplorer.

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

	MOWSE Score	#/63(%) Masses Matched	% Cov	% TIC	Mean Err ppm	Data Tol ppm	MS-Digest Index #	Protein MW (Da)/pI	Accession #	Species	Protein Name
1	3.586e+04	13 (20)	36.0	20.6	29.7	31.6	324	44766/8.5	P15651	RAT	Acyl-CoA dehydrogenase, short-chain specific, mitochondrial precursor (SCAD) (Butyryl-CoA dehydrogenase)
2	1.598e+04	7 (11)	12.0	11.1	-15.2	55.7	48355	55265/9.1	Q9PUB4	CHICK	Cytochrome P450 26 (Retinoic acid degrading enzyme CYP26)
3	9997	4 (6)	13.0	6.3	0.200	52.3	123481	19914/5.2	Q07440	MOUSE	Bcl-2-related protein A1 (BFL-1 protein) (Hemopoietic-specific early response protein) (A1-A)
4	6900	4 (6)	20.0	6.3	3.99	58.0	112604	14094/6.5	Q58143	METJA	Hypothetical protein MJ0733
5	4017	5 (7)	6.0	7.9	1.02	52.4	66674	70889/8.4	Q8IUH4	HUMAN	Zinc finger DHHC domain containing protein 13 (Huntingtin interacting protein 14 related protein) (HIP14-related protein) (Huntingtin interacting protein HIP3RP)
6	3084	5 (7)	12.0	7.9	4.40	44.4	105708	28323/5.4	P42171	CAEEL	Hypothetical protein C03C10.4 in chromosome III
7	3074	4 (6)	4.0	6.3	-9.10	57.7	27739	80472/8.4	P46977	HUMAN	Oligosaccharyl transferase STT3 subunit homolog (B5) (Integral membrane protein 1) (TMC)
8	3069	4 (6)	4.0	6.3	-9.10	57.7	37585	80598/8.3	P46978	MOUSE	Oligosaccharyl transferase STT3 subunit homolog (B5) (Integral membrane protein 1)
9	3003	6 (9)	9.0	9.5	3.79	53.2	9808	55904/9.6	P12250	BACTB	Transposase for insertion sequence

Fig.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.

4.2 Our Data Reduction algorithm Results

The obtained preprocessed mass spectrum is:

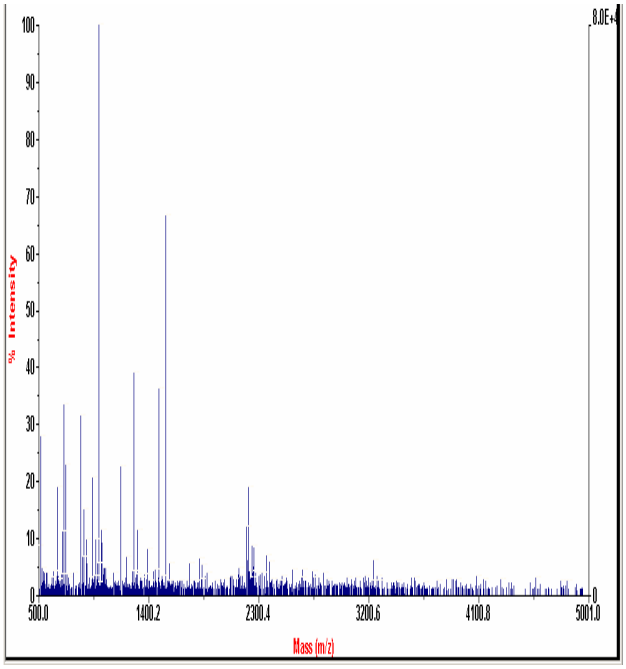

Fig.5 : Mass spectrum obtained with our algorithm.

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:

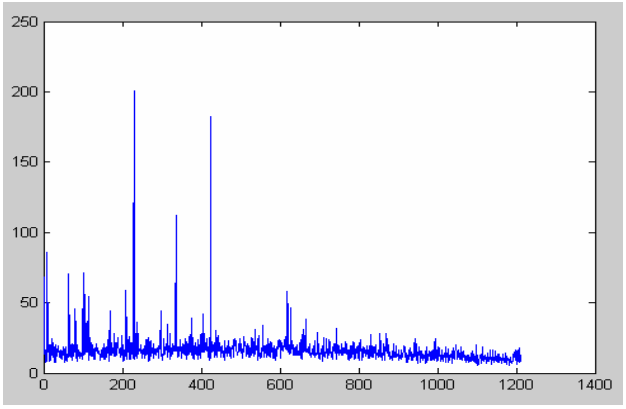

Fig.6: HF sub-band threshold values at level one.

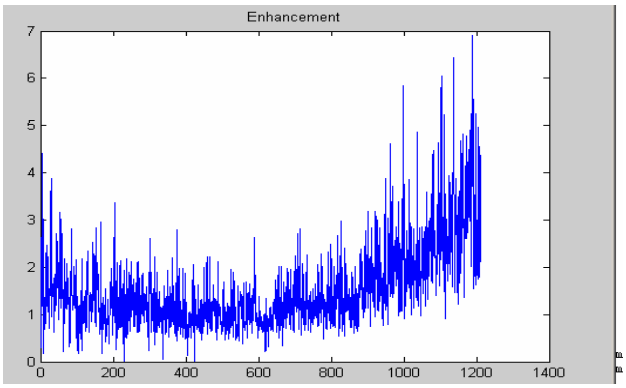

Fig.7: Amplification gain corresponding to Fig.7.

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

	MOWSE Score	#/305 (%) Masses Matched	% Cov	% TIC	Mean Err ppm	Data Tol ppm	MS-Digest Index #	Protein MW (Da)/pI	Accession #	Species	Protein Name
1	1.630e+09	22 (7)	51.0	7.2	22.4	40.0	324	44766/8.5	P15651	RAT	Acyl-CoA dehydrogenase, short-chain specific, mitochondrial precursor (SCAD) (Butyryl-CoA dehydrogenase)
2	7.950e+07	16 (5)	20.0	5.2	1.74	55.2	106063	95728/5.3	P34651	CAEEL	Hypothetical protein ZK632.5 in chromosome III
3	5.840e+07	13 (4)	26.0	4.3	-10.0	57.3	156784	69873/7.8	P43403	HUMAN	Tyrosine-protein kinase ZAP-70 (70 kDa zeta-associated protein) (Syk-related tyrosine kinase)
4	2.346e+07	17 (5)	33.0	5.6	19.5	48.2	123404	44947/9.0	Q07417	MOUSE	Acyl-CoA dehydrogenase, short-chain specific, mitochondrial precursor (SCAD) (Butyryl-CoA dehydrogenase)
5	5.107e+06	15 (4)	23.0	4.9	-5.41	54.2	83076	96391/5.0	Q9CEW0	LACLA	Alanyl-tRNA synthetase (Alanine--tRNA ligase) (AlaRS)
6	4.165e+06	14 (4)	18.0	4.6	-1.19	57.6	59871	65512/6.0	Q28422	ARCFU	Hypothetical protein AF1856
7	4.139e+06	15 (4)	22.0	4.9	-19.4	46.3	54505	92419/9.1	P59057	BUCAP	Phenylalanyl-tRNA synthetase beta chain (Phenylalanine--tRNA ligase beta chain) (PheRS)
8	4.129e+06	14 (4)	18.0	4.6	0.222	46.9	62462	94978/5.8	Q9HY08	PSEAE	DNA mismatch repair protein mutS
9	2.925e+06	11 (3)	13.0	3.6	-4.93	51.6	153694	76929/6.1	Q09057	NEIMB	Transferrin-binding protein 2 precursor (TBP-2)

Fig.8: Theoretical masses matched in SwissProt DataBase.

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

5. Conclusion

Protein identification and characterization is one of the most essential tasks performed in proteome research. The precise detemination of the peptide masses in the spectra , and highly discriminating mass comparaison 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:

[1] S. Golotvin, A. Williams, Improved Baseline Correction of FT NMR Spectra, *Advanced Chemistry Development,NMR Newsletter Advanced Chemistry Development,1999*

[2] S. Grace, B. Yu, M. Vetterli, Adaptive wavelet Thresholding for Denoising and Compression, *IEEE Transactions on image processing*, Vol. 9, NO.9, 2000, pp. 1532-1546.

[3] H. Haussecker, H. R. Tizhoosh, Fuzzy Image Processing, I. *Handbook of Computer Vision Application, Edited by B. Jagne, H. Haussecker, and P; Geisster, Academic Press 1999.*

[4] E. D. Jansing, T. A. Albert, D. L. Chenoweth, Two Dimensional Entropy Segmentation. *Pattern Recognition Letters 20*, Letters 20, 1999, pp. 329-336.

[5] A. Lindegren, Analysis of Proteomic Patterns for Detection of Prostate Cancer. *Master Thesis*. 2004.

[6] P. Lio, Wavelets in bioinformatics and computational biology : state of and perspectives, *Bioinformatics*, Vol. 19, 2003, pp 2-9

[7] B. Liu, Y. Sera, N. Matsubara, K. Otsuka, S. Terabe, Signal Denoising by Wavelets for Microchip Electrophoresis, *Chromatograpghy*

Tokyo- Society for Chromatograpghic sciences, Vol.23, Part Supp, 2002, pp. 59-60

[8] D.I. Malyarenko, W.E. Cooke, B.L. Adam, G. Malik, H. Chen, E.R. Tracy, M.W. Trosset, M. Sasinowski, O.J Semmes, and D. M. Manos, Enhancement of sensitivity and resolution of Surface-Enhanced Laser Desorption/Ionisation Time-of-Flight Mass Spectrometric Records for Serum Peptides Using Times Series Analysis Technique, *Proteomics and Protein Markers, Clinical Chemistry*, 2005, pp. 65–74

[9] N. Nafati.. Synthèse itérative et simultanée de banc de filtres biorthogonaux de reconstruction parfaite avec des critères adaptés aux applications de codage de la parole et de l'image, *GRETSI Grenoble France*. 1997, pp. 1085-1088.. Septembre.

[10] P. Nugues. Interprétation de Gels d'Electrophorèses 2D, *Thèse de Doctorat, Université de Nancy*, 1989.

[11] T. D. Pham, A New Approach for Calculating Implications of Fuzzy Rules, *IEEE International Conference on Artificial Intelligence Systems*, 2002, pp. 71.

[12] J. POLEC, J. Pavlovieova, T. Karlubikova, Application Of Shape-independent Orthogonal Transform For Image Inerpolation, *Radio-Engineering*, Vol. 11, No. 1, April 2002.

[13] E. G. Sánchez, Y.A. Dimitriadis, M. Sanchez-Reyes Mas, P. S. García, J.M. Cano Izquierdo, J. Lopez Coronado, On-Line Character Analysis and Recognition With Fuzzy Neural Networks, *Intelligent Automation and Soft Computing*, Vol. 7, No. 3,1998, pp. 161-162.

[14] R. Zenobi, R. Knochenmuss, Ion formation in MALDI Mass Spectrometry. *Mass Spectrom* 1998, **Rev**. **17, pp. 337**–**66**.