RAy: Increasing the Scope of Mass Spectrometry–Homology Based Database Search

Şule Yılmaz[1] and Jens Allmer[2]

[1] Biotechnology Department, Izmir Institute of Technology, Urla – İzmir, Turkey
[2] Molecular Biology and Genetic Department, Izmir Institute of Technology, Urla – İzmir, Turkey

Abstract

Mass spectrometry is an analytical technique used in protein sequencing and identification. There are two approaches to analyze mass spectrometric data: database search and de novo sequencing. The most commonly used approach in proteomics is database search but it requires a sequence database; otherwise de novo sequencing can be used.. Error tolerant matching can be achieved by peptide sequence tags (PST) and MS-BLAST but is somewhat limited. A PST is a short amino acid sequence with mass information (Figure 5). They are specific to given MS/MS spectra and they are used to filter database. MS-BLAST enables homology searches, and as a result of this it is possible to retrieve some matches using related species.

We propose a new algorithm based on profiles so as to enable comprehensive sequence homology-based database search. Our profile contains frequencies of all amino acids supported in an MS/MS spectrum. The algorithm has two steps – profile constructions and profile alignment (Figure 1). It has some features that overcome limitations of existing methods. First of all, they are more specific comparing to a PST and post translational modification causes less problems. Secondly, it enables to do purely sequence-based homology search information without restrictions as found in other algorithms.

Flowchart

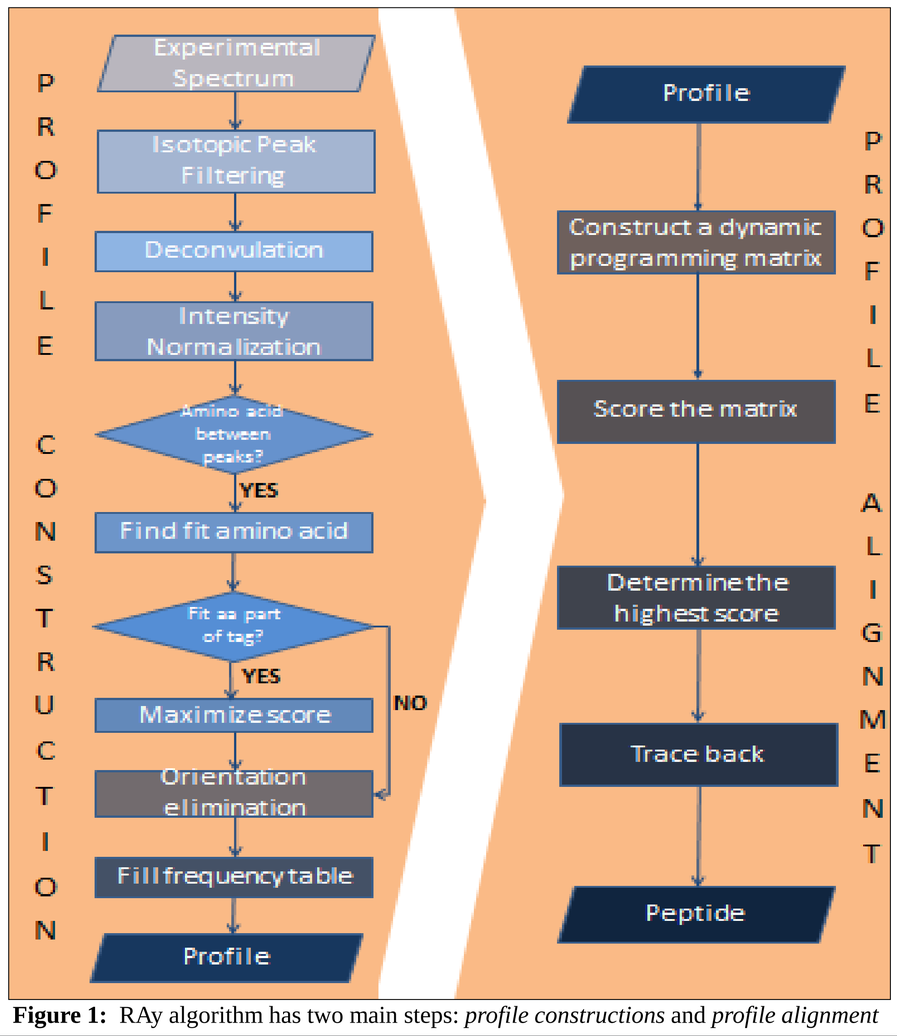

Figure 1: RAy algorithm has two main steps: *profile constructions* and *profile alignment*

Importance

Due the homology search functionality of RAy, proteins from related species can be found in database. Hence number of identifiable spectra and also sequence coverage of identified protein increases (Figure 2). Moreover, we hope to assign post translational modifications (PTM) to spectra (Figure 3) because we can still match the sequence and by comparing the calculated mass to the observed mass offer possible PTMs for unsure regions in the assignment. . This feature is important for analyzing disease conditions, as well (Figure 4)

RAy has features that overcome some limitations in existing approaches. Firstly, PSTs are used in filtering but orientations of PSTs are not certain (Figure 5), but in RAy this problem is reduced as much as possible by using a progressive greedy orientation algorithm. Secondly, database search algorithms usually work using enzymatic information and partial of full mass restriction (red circle in Figure 5) . Neither is required in RAy.

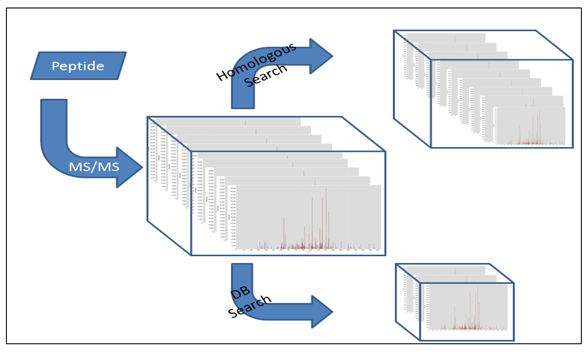

Figure 2: Increase in identifiable spectra and sequence coverage

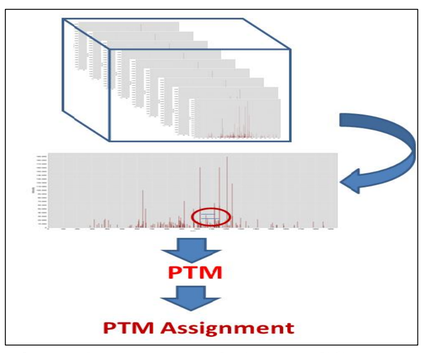

Figure 3: Post translation modification assignment

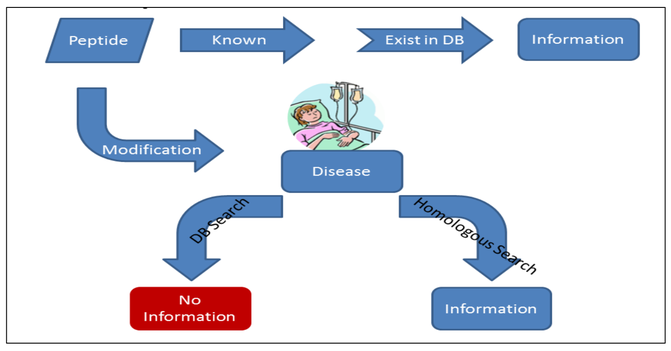

Figure 4: Peptide information under disease condition

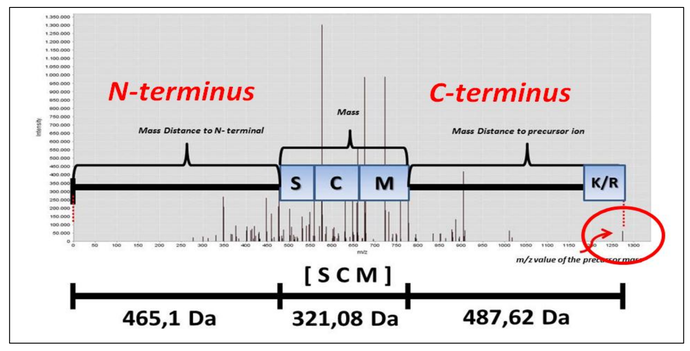

Figure 5: Peptide sequence tag.

Results

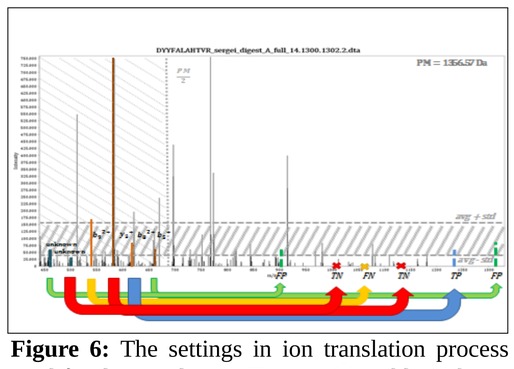

Figure 6: The settings in ion translation process used for deconvolution. True positive (blue) shows a translated 2+ peak, true negative (red) indicates a non-translated unknown or1+ charged peak. False positives (green) are any translated unknown or 1+charged peaks whereas false negatives (yellow) are non-translated 2+ charged peaks

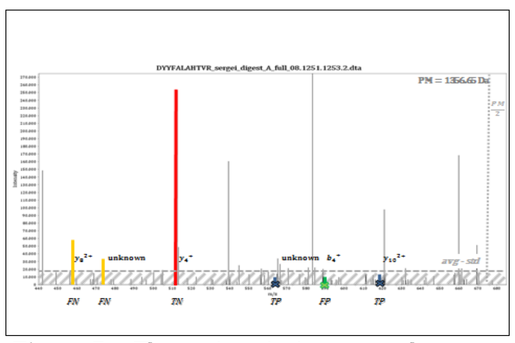

Figure 7 : The settings in ion removal process used for deconvolution. True positive (blue) shows any removed any unknown or 2+ charge peaks, true negative (red) indicates a non-removed 1+ charged peak. A false positive (green) is a removed 1+ charged peak whereas false negatives (yellow) are any non-removed unknown or 2+ charged peaks.

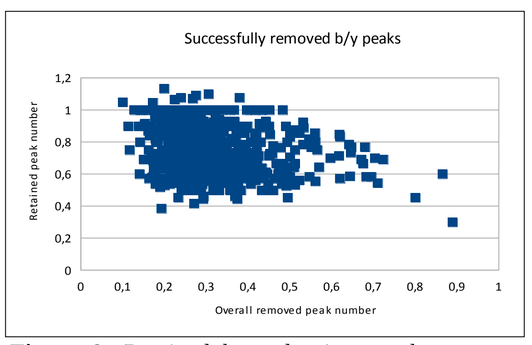

Figure 8: Retained b- and y-ion number versus overall peak number after filtration and deconvolution. While removing rate increases, a reasonable amount of b- and y-ions still remains new ions being deduced and added to the spectrum.

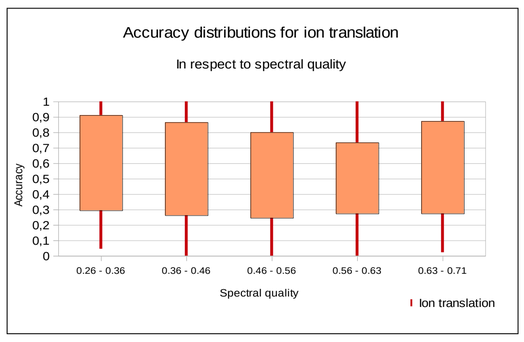

Figure 9: The accuracy distributions for ion translation regarding to spectral quality groups

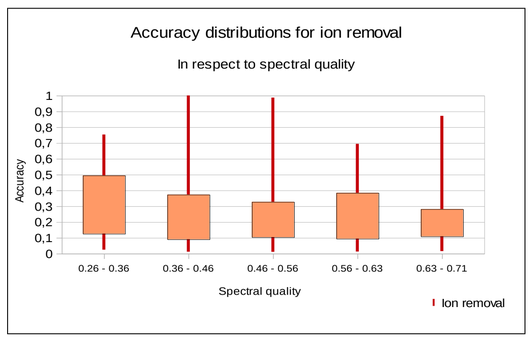

Figure 10 : The accuracy distributions for ion removal regarding to spectral quality groups

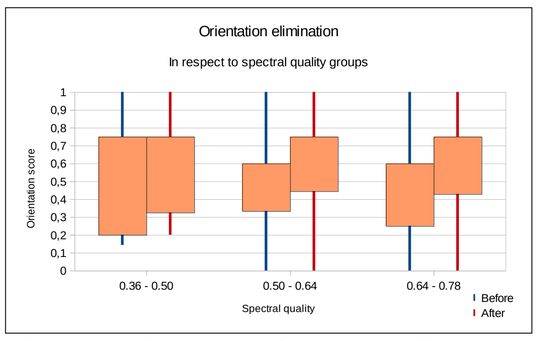

Figure 11: The success of orientation elimination process. While spectral quality is increases, percentage of amino acids that show the same direction increases.

Further Studies

Improvement of the orientation algorithm including tag significance and peak abundance
Implementation of Smith-Waterman profile sequence alignment analysis on CUDA architecture
Synthetic peptide measurements and result analysis

The profile of the spectrum named AAQGDITAPGGAR_sergei_digest_B_full_04.0537.0537.2.dta is below:

G:	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,1	0,1	0,0	0,0	0,0	0,0	0,0
A:	0,0	0,0	0,0	0,0	0,1	0,1	0,0	0,0	0,2	0,1	0,0	0,1	0,2	0,0	0,0	0,1	0,1	0,0	0,0	0,0	0,0
S:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,1	0,0	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
P:	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,1	0,0	0,0	0,0	0,1	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0
V:	0,0	0,1	0,1	0,1	0,1	0,1	0,0	0,0	0,1	0,1	0,0	0,1	0,1	0,0	0,0	0,1	0,1	0,1	0,1	0,1	0,1
T:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,2	0,3	0,2	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
C:	0,6	0,4	0,2	0,1	0,1	0,1	0,1	0,1	0,0	0,0	0,0	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
J:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,1	0,0	0,0	0,0	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0
N:	0,0	0,1	0,1	0,3	0,1	0,1	0,1	0,2	0,1	0,1	0,0	0,1	0,1	0,2	0,1	0,1	0,1	0,2	0,1	0,1	0,0
D:	0,0	0,0	0,0	0,0	0,0	0,1	0,2	0,2	0,0	0,0	0,0	0,0	0,0	0,2	0,1	0,1	0,0	0,0	0,0	0,0	0,0
Q:	0,2	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,1	0,1	0,2
K:	0,2	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,1	0,1	0,2
E:	0,0	0,1	0,2	0,2	0,1	0,1	0,1	0,1	0,0	0,1	0,1	0,1	0,0	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,0
M:	0,0	0,0	0,0	0,1	0,1	0,1	0,1	0,1	0,0	0,0	0,1	0,0	0,0	0,1	0,1	0,1	0,1	0,1	0,1	0,0	0,0
F:	0,0	0,0	0,0	0,1	0,1	0,1	0,1	0,0	0,0	0,1	0,0	0,1	0,0	0,1	0,0	0,1	0,1	0,1	0,0	0,0	0,0
R:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
Y:	0,0	0,0	0,1	0,1	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,1	0,0	0,0	0,0
W:	0,0	0,0	0,0	0,0	0,1	0,1	0,2	0,1	0,1	0,0	0,0	0,0	0,1	0,1	0,1	0,1	0,1	0,0	0,0	0,0	0,0
X:	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

Figure 12 : A profile of an experimental spectrum. The rows show amino acids, except for isoleucine and leucine which are combined into one row (J). Each column is different from the next one by Glycine mass. Each box in the table shows frequency of that amino acid in a given location as deduced from the MS/MS spectrum.

Acknowledgement

This project is supported by TUBITAK with project number TUBITAK - 111E139