Exam Project for Analysis of Protein Expression

Nguyen Thi Ngoc Mai

December 2, 2018

1 Introduction

Mass Spectromatry play an important role in proteomics area and is applied widely in many other fields, especially a strong technique for sequencing peptides, discovering the structure and identifying protein.

This project mainly focuses on developing score algorithm to compare theoretical spectra with observed data obtained from Thermo Q Exactive Orbitrap instrument. Analyzing tandem mass (MS2) data to identify the peptide sequences, then based on identified peptide, define the protein analyzed.

Searching database includes the list of 50 target proteins (identified by unique accession number) and the list of 50 decoy proteins. Based on this information, a database for peptides and their mass can be created by digesting protein in database with Tryptic. Then, the peptide database is built for a reference source to define potential peptide candidates based on the experimental data from MS1.

5 spectrums of interest from MS1 were chosen for analyzing further more with MS2, fragmenting apeptide into b and y ions, data obtained measured by m/z of ions and their intensity values. The charge of theoretical ions is upto charge of precursor ions minus one.

Programming language used in this project is Python. Related to Mass Spectrometry, there is one package from third party code, **Pyteomics** [1], to get the m/z for fragment ions (daughter ions) from a give precursor mass from MS1. This package is also used to manipulate protein such as digest protein to get the list of peptides give protein sequence. Other third party packages applied for visualizing data and applying mathematics and statistics.

The method applied to match between observed data and theoretical data is straight forward with Uniform approach for getting intensity of theoretical ions.

With the mass of precursor ions, the tolerance of 100ppm is applied to search the available data to get the potential candidates for peptides tandem analysis. With scoring algorithm, non-probabilistic approach is applied, simply focusing on counting number of matches of peaks between observed and theoretical ions.

2 Working with Data

2.1 Data Exploring

Experimental data for MS2 are visualized in Figure 1. Based on the overview about m/z of observed fragment ions, m/z of theoretical fragment ions can be added with the tolerance of 0.1da to see how the peaks are matched.

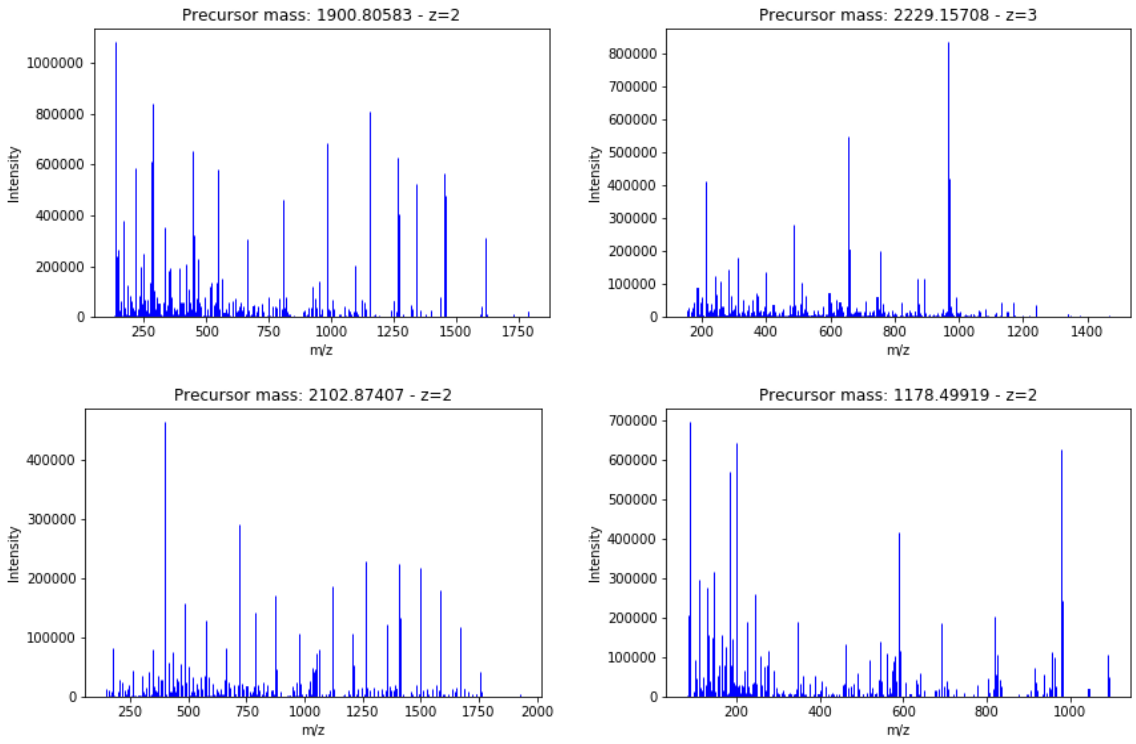

Figure 1: Intensity and m/z for MS2 data

2.2 Database building

2.2.1 Target and Decoy Protein Database

Firstly, fasta file for protein target and decoy are import to program to create database for searching, getting protein accession number based on identified peptides from MS2. There is 1 protein is duplicated in both fasta data file for targe and decoy proteins. Then, even in each data file contains 51 proteins, there are actually 50 proteins recorded by unique accession number. However, decoy data file uses the order of decoy protein instead of using accession number, protein 24 and 48 were duplicated. This was noted for further analysis.

The list of accession numbers as unique key after creating database for target protein is showed below. When searching by accession number, the whole protein sequence will be returned.

2.2.2 Peptides Database

Based on the protein databases, peptides database is also developed to find the potential peptide candidates based on the mass of peptides in this data with using tryptic for digesting and applying the tolerance 100ppm. The mass of each peptide is summed of all residues' mass in the peptide sequence.

2.2.3 Fragment Ions Database

After getting potential peptide candidates, fragment ions database is built up for matching the peaks from observed data. Each peptide is the unique key to access the NxM matrix of fragment ions with N+1 is the length of the peptide sequence and M the number of charge states defined. Charge of fragment ions is calculated as up to charge of precursor ions minus one.

3 Develop Function

The whole program is written by Python. The list of functions in project programming include:

- Reading fasta file and data sets of 5 spectrums
- Building databases (protein DB, peptide DB, fragment ions DB)
- Getting potential peptide candidates
- Getting fragment ions

- Getting accession number of protein from a given peptide sequence
- Matching theoretical fragment ions and experimental data
- Scoring the matched peaks
- Assigning the confidence on the findings
- FDR computation

4 Result

4.1 Potential Peptide Candidates

Using the tolerance at 100ppm to select potential peptide candidates, results are presented in tables 2, 3, 4, 5. In selected peptides, there are two spectrums have similar MH+ value, 2102.87407 and 2102.87345 with different charge, 3 and 2 respectively. These two spectrums have the same potential candidates with the tolerance at 100ppm. Thus, the result for these two spectrums are both represented in table 4.

4.2 Peak Matching

Theoretical fragment ions are generated to compare with experimental data. Peak matching, illustrated in Figure 2, is based on the tolerance at 0.1da between observed fragment ions and theoretical fragment ions.

Based on this, number of peak matching for each potential peptide candidate is counted to see for each peptide, how many ions are matched with observed data.

Peak intensities for theoretical fragment ions are done with the Uniform Level Model, a uniform intensity is assigned to each fragment. This level does not require any statistics approach.

4.3 Peak Scoring

Scoring can be obtained by comparing and an experimental and a theoretical spectrum by different scoring schemes. In this report, method approach is just focusing on searching for matching peaks, then getting the non-probabilistic score. This method has the disadvantage of finding corresponding peaks.

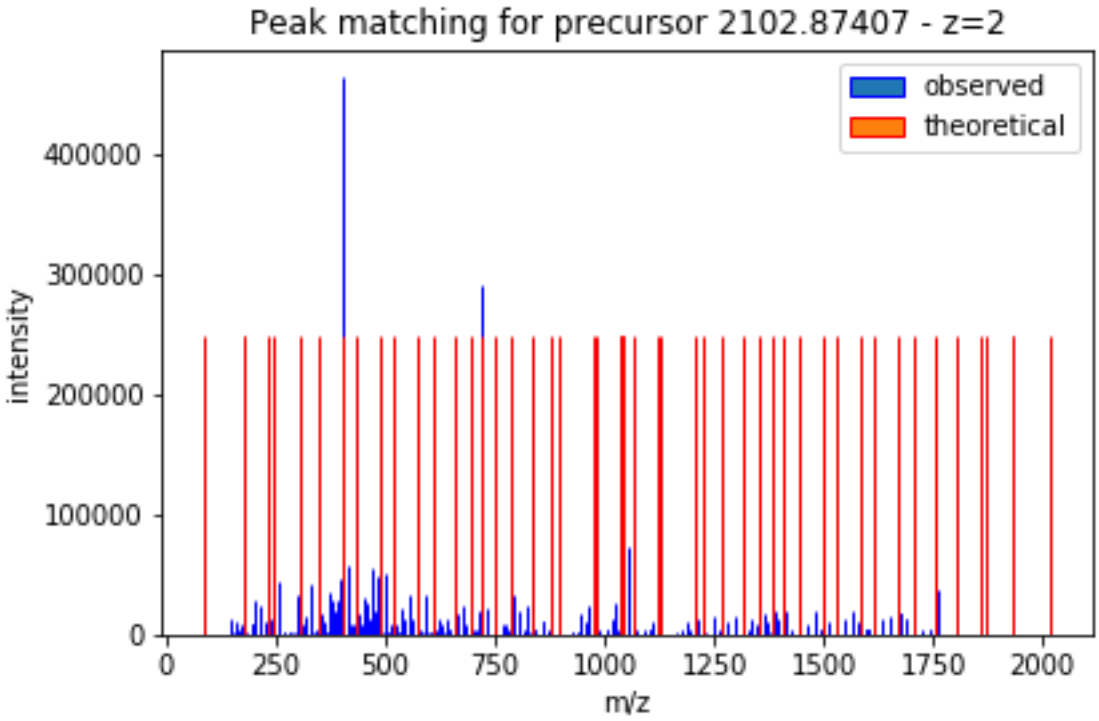

Figure 2: Peak Matching Illustration

4.3.1 Number of matching peaks

This is the simplest procedure for comparison without considering intensities into scoring function. For each potential candidate and its theoretical fragments, counting the number of matching peaks will be used to score the candidate. Higher number of matching peak higher score.

4.3.2 Number and intensities of matching peaks

This method is extended from number of matching peaks by taking intensities into account and interval division is used by following formula:

$$C = \sum_{j=1}^{n} I_j^R I_j^T \tag{1}$$

However, this approach assumes a linear increase in score as the number of matches increases. This is unreasonable due to random occurring. Alternative score proposed by Fenyo and Beavis (2003) taking number of matched b (y) ions:

$$C^{"} = Ce^{n_b + n_y} \tag{2}$$

Fenyo and Beavis (2003) showed that this alternative scoring function outperformed the simple score above.

The intensity is normalized instead of using the original intensity. The intensity is normalized by divided the intensity of each peak for the sum of intensities of all peaks recorded.

4.4 Assigning confidence on findings

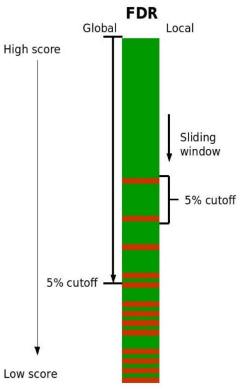

Figure 3: FDR computation

The confidence level of potential peptide candidates can be based on False Discovery Rate (FDR), local or global, to evaluate. The principle of computing local and global FDR is illustrated as in figure 3 [2].

Global FDR is calculated by the proportion of false discoveries in the list of ranked results above given threshold score. On the other hand, local FDR is the proportion of false discoveries in the slicing window along the list of results.

In this report, the score defined for threshold level in this report is 0 and this threshold is applied to calculate global FDR for peptide level only, not consider for protein level. FDR by each spectrum is presented in table 1 below.

As results shown in table 2, 3, 4, 5, after ranking based on score followed section **4.3.2**, considering intensities and number of peaks matching, there are different preferences for peptides from decoy database between spectrums, particularly, small size of database (50 target proteins and 50 decoy proteins), choosing slicing window to calculate local FDR could lead to too high false discovery rate. Here, only global FDR will be considered based on threshold level.

Precursor Mass (z)	FDR
1900.80583 (2)	0.20
2229.15708 (3)	0.3077
2102.87345 (2)	0.40
2102.87407 (3)	0.40
1178.49919 (2)	0.2727

Table 1: Table inside a floating element

4.5 Protein identification

Based on score obtained from two methods mentioned above and the assigned confidence, potential peptide candidates for each precursor mass are ranked from the highest to the

lowest score. Protein identification can be defined by integrating all evidences from spectra associated with a given protein. As result shown in table 2, 3, 4, and **??**, Protein with accession number Q12797 got four hits from five spectrum analyzed with the highest score and extremely low p-value.

In the data analyzed here, there are some shared peptides between proteins, each of potential peptide referred to a unique accession number. Integrating results from 5 spectrum, protein **Q12797** seemed to be the analyzed protein with the high coverage due to number of identified peptides, potential peptide candidates from this protein appear in 4 over 5 spectrums in Tandem Mass Analysis.

Protein **Q8IVF2** has two peptide candidates appear in 3 of 5 spectrums. With decoy database, decoy 22 also contains 2 peptides appear in 3 of 5 spectrum analyzed. And the rest of candidate peptides are come from different proteins.

As results presented, there are some proteins share the same peptide, such as in table 4, precursor peptide with mass 2102.8704 charged 2, searching from target database, the potential peptide candidate 'NGQGWVPSNYITPVNSLEK' appears in both protein P00519 an P42684; candidate 'AVDWWGLGVVMYEMMCGR' is shared in three proteins P31749, P31751, and PQ9Y243. When searching in decoy database, candidate 'ELSNVPTIYNSPVWGQGNK' can belong to protein decoy8 or decoy9, and 'GCMMEYMVVGLGWWDVAR' also shows up in decoy23, decoy25 and decoy26.

5 Conclusion

There is many other scoring schemes as well as method to define the confidence for findings. This report is just provide the simplest process to analyze data and get some intuitive overview about how to construct a theoretical framework to identify protein.

In this report, the issues related to post modification, isotope distribution, miscleaved peptides are not considered yet. Including decoy database here also give us the score-based decision, even decoy protein can get higher score than the real protein, this leads to the false discovery for the results achieved.

References

[1] Pyteomics - Goloborodko, A.A.; Levitsky, L.I.; Ivanov, M.V.; and Gorshkov, M.V. (2013) Pyteomics - a Python Framework for Exploratory Data Analysis and Rapid Software Prototyping in Proteomics, Journal of The American Society for Mass Spectrometry, 24(2), 301304. DOI: 10.1007/s13361-012-0516-6

[2] Sennels et. al Sennels, P., Bukowski-Wills, J.C., Rappsilber, J. (2009). Improved results in proteomics by use of local and peptide-class specific false discovery rates. *BMC Bioinformatics*, 2009, 10:179. DOI: 10.1186/1471-2105-10-179

Results

target peptide	s1	s2	rank	access.no	decoy peptide	s1	s2	rank	decoy.no
LGIYDADGDGDFDVDDAK	26	10.80	1	Q12797	ADDVDFDGDGDADYIGLK	7	3.75	2	decoy24(48)
WFGSDVLQQPLPSMPAK	11	2.58	3	Q9NRA8					
VPDVPSESGSPVYVNQVK	7	1.24	4	P28222					
VTFPNGFTLEGSFGSGAGR	8	1.09	5	Q60I27					
CFSLDAFCHHFSNMNK	6	1.00	6	Q2TB18	APMSPLPQQLVDSGFWK	14	0.66	7	decoy1
LHLLCLLCAEEEEEEK	5	0.53	8	Q2VPB7					
GPNPAFWWVNGQGDEVK	5	0.26	9	Q08AH1					
FYATGYISSAFLFGATGK	6	0.23	10	Q16853					
TSSSSCSAHSSFSSTGQPR	5	-0.16	11	Q15052	NMNSFHHCFADLSFCK	4	-0.17	12	decoy49
ANPMYNAVSNADLMDFK	4	-0.24	13	P01160					
NNFEFCEAHIPFYNR	5	-0.41	14	Q9H0R1					
Total Matching Count	11					3			

Table 2: Potential candidates for precursor ion mass 1900.80583 - charge 2

target peptide	s1	s2	rank	access.no	decoy peptide	s1	s2	rank	decoy.no
GAIETYQEVASLPDVPADLLK	35	9.37	1	Q12797					
FWAANVPPSEALQPSSSPSTR	18	5.00	2	Q96P48					
CPDIWLGCIMEDTGFYLPR	16	4.96	3	O75078					
TEVEAGASGYSVTGGGDQGIFVK	16	3.83	4	Q8IVF2					
LLLENVLPAHVAPQFIGQNR	13	3.31	5	Q8NFM4					
LAPDPSLVIYAIFPSGGVVADK	13	2.89	6	A8K2U0	VFIGQDGGGTVSYGSAGAEVETK	12	2.89	7	decoy22
AFNSHTWEQLGTLQLLSQR	11	2.82	8	Q9BUJ0	TSPSSSPQLAESPPVNAAWFR	11	2.03	9	decoy42
TALHLACANGHPEVVTLLVDR	10	1.84	11	Q9UPS8	NQGIFQPAVHAPLVNELLLR	11	1.86	10	decoy16
HENNLSVLAISNMEASSTLAK	9	0.84	13	Q68CP9	DVLLTVVEPHGNACALHLATR	9	1.71	12	decoy33
Total Matching Count	9					4			

Table 3: Potential candidates for precursor ion mass 2229.15708 - charge 3

target peptide	s1	s2	rank	access.no	decoy peptide	s1	s2	rank	access.no
SSGNSSSSGSGSGSTSAGSSSPGAR	44 (33)	17.90	1	Q12797	GCMMEYMVVGLGWWDVAR	6	0.57	2	decoy23
									decoy25
									decoy26
NFPGSSQSEIIQAIQNLTR	4 (2)	0.02	3	Q8N1W1					
GHFDAVQYLLEHGASCLSR	5 (5)	0.71	4	Q9BZ19	SFAYHTNLQTSADSGNFDK	4	0.001	5	decoy32
AVDWWGLGVVMYEMMCGR	1 (6)	-0.02	6	P31749	SLCSAGHELLYQVADFHGR	4	-0.05	7	decoy35
				P31751					
				Q9Y243					
DFNGSDASTQLNTHYAFSK	4 (8)	-0.63	8	Q6UB98					
NGQGWVPSNYITPVNSLEK	4 (2)	-0.65	9	P00519					
				P42684					
GDDEEGECSIDYVEMAVNK	3 (3)	-0.90	10	O75891	IIAGWYHSCLGSDPDTVLR	3	-0.95	11	decoy18
GDDEEGECSIDYVEMAVNK	3 (3)	-0.90	10	O75891					
MIQTNFIDMENMFDLLK	3 (1)	-1.01	13	Q9NP58	TLNQIAQIIESQSSGPFNR	3	-0.95	12	decoy43
LVTDPDSGLCSHYWGAIIR	3 (2)	-1.40	14	P55196	ELSNVPTIYNSPVWGQGNK	2	-1.59	15	decoy8
									decoy9
KPDAEVLTVESPEEEAMTK	1 (3)	-2.36	17	Q8IVF2	SQVGSQLFEDQQLTPLGQK	2	-1.78	16	decoy20
GHQEVLEGYPSSETELSLK	1 (1)	-2.36	17	Q9ULJ7					
QGLPTLQQDEFLQSGVQSK	5 (3)	-2.36	17	Q6ULP2					
Total Matching Count	12					7			

Table 4: Potential candidates for precursor ion mass 2102.87407 - charge 2

target peptide	s1	s2	rank	access.no	decoy peptide	s1	s2	rank	access.no
QEDSPFQCPK	10	2.86	1	Q96L96					
IMNMEAGTLAK	9	2.62	2	Q96CM8	EPELAIYCLK	9	2.01	3	decoy14
TLNSTSPFPSK	7	1.58	4	O95477	IMSQTSEVQR	8	1.34	5	decoy47
GLAAHYFFPR	7	1.27	6	O00203	QDCWVQMLR	9	1.21	7	decoy31
LLASCSADGTLK	10	1.11	9	O14727	LTGDASCSALLK	6	1.17	8	decoy41
NGGCNHMQCSK	6	1.10	10	O95376					
NLLSAYGEVGR	7	0.93	11	Q9ULW3	DALSMDEELR	10	0.89	12	decoy22
QVESTQSMIR	6	0.81	13	Q8N3C0					
DEFEIVGDVR	8	0.78	14	Q9BYV1					
DGYFWFMGR	8	0.77	15	Q6NUN0					
ENGSCQIITIT	7	0.46	16	Q9H7F0					
LMQVWCDQR	6	0.30	17	Q9UKU9	SCQMHNCGGNK	4	0.29	18	decoy45
AEQEALTGECK	6	0.25	19	P18848					
LCYIALEPEK	5	0.24	20	Q8TDG2					
ESVELHDPPR	7	0.09	21	P51816					
ASSPQGFDVDR	6	0.06	22	P27216					
GHEDDSYEAR	6	-0.07	23	Q9H2P0	GVEGYASLLNR	7	-0.10	24	decoy10
AQSYEYVYR	3	-0.41	25	Q86V21	VDGVIEFEDR	4	-0.57	26	decoy21
Total Matching Count	18					8			

Table 5: Potential candidates for precursor ion mass 1178.49919 - charge 2