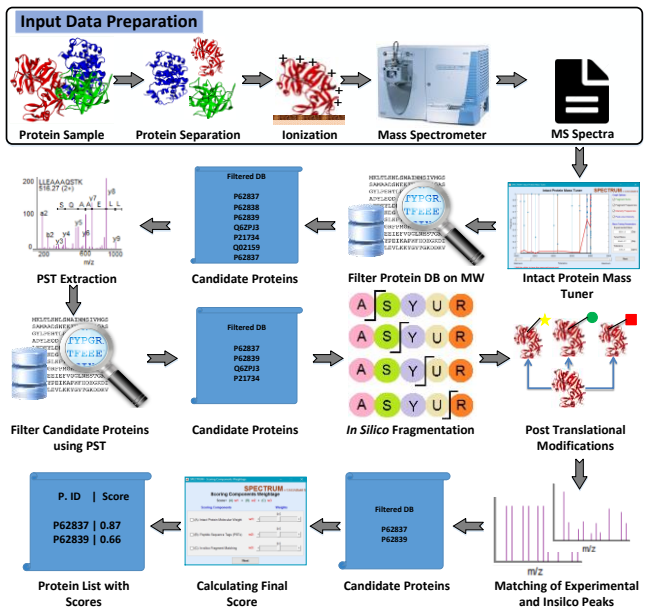
Basic Algorithms & Scoring Schemes for Searching Protein Spectra

Department of Life Sciences, SBASSE, LUMS

1

Out of Experiment and into Algorithms



2

Format of Tandem MS Data

8559 1 <-MS1 Data

635.39084
654.44981
763.490304
864.543503
866.611999
1135.705343
1217.828113
1248.798419
1263.796108
1304.861342
1433.914879
1477.96237
1535.004263
1562.021644

<- MS2 Data

3

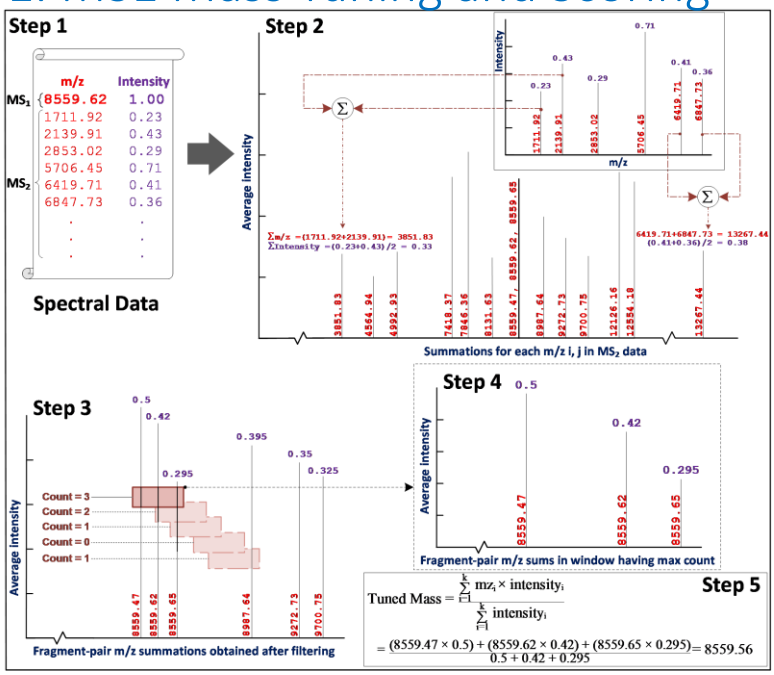
Things to watch out for in MS data

1. Intact Protein/Peptide Mass
2. Charge States
3. Relative Abundances
4. Technique-specific fragmentation patterns
5. Mass Shifts (PTMs, neutral losses)

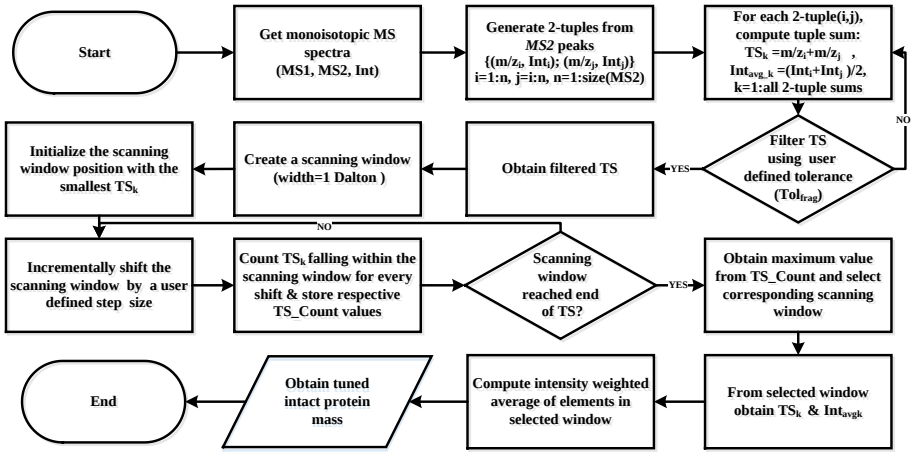
4

4

1. MS1 Mass Tuning and Scoring



Estimating Intact Mass



$$TunedMass = \frac{\sum_{i=1}^{m} TS_{k_i} \times Int_{k_i}^{avg}}{\sum_{i=1}^{m} Int_{k_i}^{avg}}$$

Intuitively Scoring Tuned Masses

- As a first step in protein search, protein database is filtered for proteins matching the MW reported in the experimental data
- What to do incase multiple proteins fall in the mass range?
- Scoring Philosophy: The closer the better!

$$M_{Score} = \frac{1}{\sqrt{(M_{Exp} - M_{Thr})^2}}$$

7

7

What we do in SPECTRUM?

$$Mass_{diff} = | Mass_{experimental} - Mass_{theoretical} |$$

$$Score_{WPMW} \begin{cases} 1, Mass_{diff} = 0 \\ \frac{1}{2^{Mass_{diff}}}, 0 < Mass_{diff} \leq Thr \\ 0, Mass_{diff} > Thr \end{cases}$$

8

8

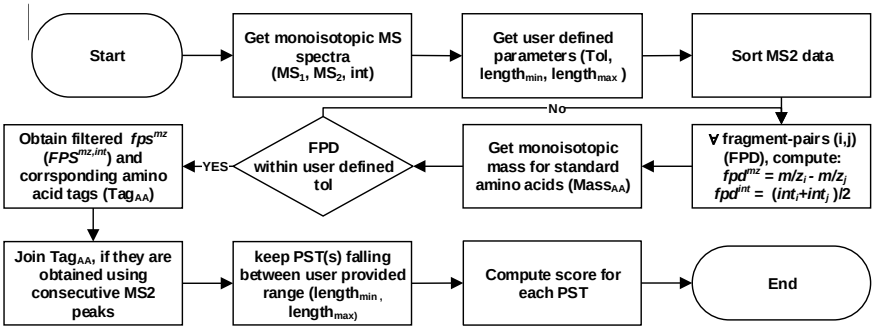
2. Peptide Sequence Tags

- Upon obtaining scores of all proteins in the protein database, we filter the database for “candidate proteins”
- Sequence tags are extracted from spectral data
- These sequence tags are then searched in the candidate proteins and a re-scoring is performed
- Upon obtaining the new scores, the “candidate proteins” are further shortlisted and sorted as per the newer scores.

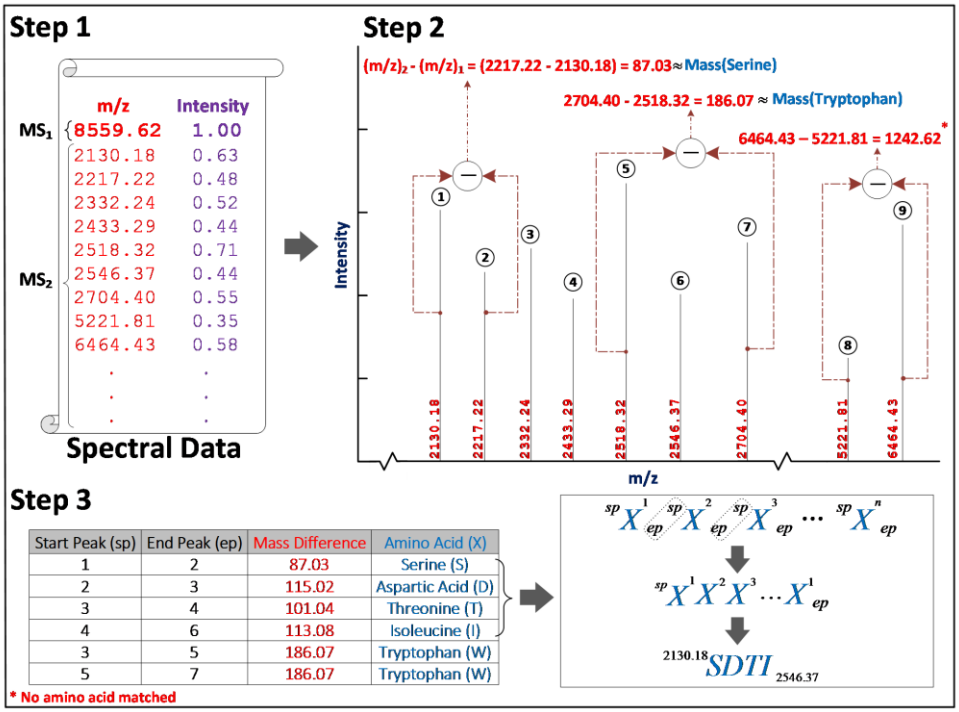
9

9

Extracting Peptide Sequence Tags



10



11

Scoring Sequence Tags - I

- Sequence Tag Examples: ‘M’, ‘MQ’, and ‘QV’ etc
- What can we consider to be “scorable” attributes of these tags?
  1. Length
  2. RMSE
  3. Abundance
- Scoring Philosophy:
  - The lengthier the tag, the better,
  - The smaller the RMSE, the better,
  - The more abundant the better!

12

Scoring Sequence Tags - II

- If a candidate protein matches 'n' PSTs, then its score can be given by:

$$PST_{Score} = \sum_{i=0}^{n} Length\,(PST_i)^2$$

Of a single PST

- Additionally, if we include RMSE to the scoring system, then it can highlight better PST matches.

Scoring Sequence Tags - III

- So, what is the RMSE for a specific sequence tag 'i' of length 'n'?

$$RMSE_i = \sum_{i=0}^{n} \frac{\sqrt{(M_{Hop} - M_{AA})^2}}{n}$$

of a single PST

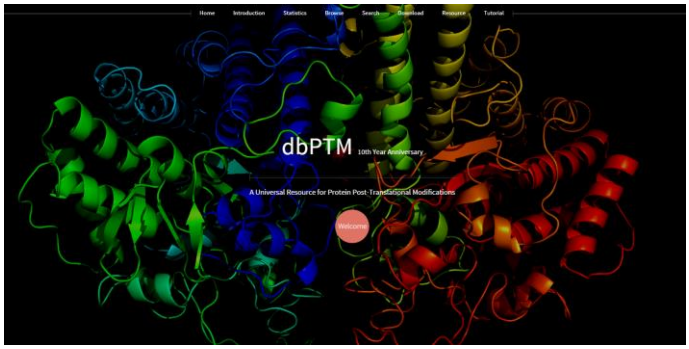
So, the updated relationship is:

$$PST_{Score} = \sum_{i=0}^{n} \left(\frac{Length(PSTi)}{RMSE_i}\right)^2$$

of 'n' PSTs in a protein

Cookie Point: How to cater for abundance? (0.25)

3. Post-translational Modifications



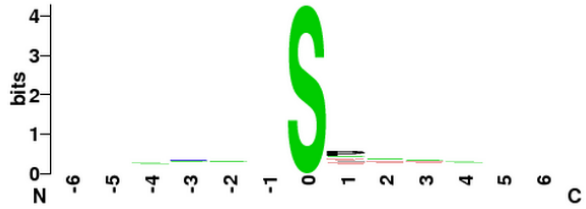
15

PTM Type	Ala(A)	Arg(R)	Asn(N)	Asp(D)	Cys(C)	Gly(G)	Glu(E)	Gln(Q)	His(H)	Ile(I)	Leu(L)	Lys(K)	Met(M)	Phe(F)	Pro(P)	Ser(S)	Thr(T)	Trp(W)	Tyr(Y)	Val(V)
Acetylation	1040	8	-	36	11	64	20	-	-	-	-	7732	616	-	15	676	132	-	6	29
ADP-ribosylation	-	126	1	-	18	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Allysine	-	-	-	-	-	-	-	-	-	-	-	37	-	-	-	-	-	-	-	-
Amidation	41	119	104	5	85	142	15	29	16	97	443	68	89	637	47	52	89	71	62	332
Biotin	-	-	-	-	-	-	-	-	-	-	-	10	-	-	-	-	-	-	-	-
Bromination	-	-	-	-	-	-	-	-	1	-	-	-	-	-	-	-	-	32	-	-
C-linked Glycosylation	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	156	-	-
Chromophore	-	-	-	-	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Citrullination	-	32	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Covalent protein-DNA linkage and Phosphorylation	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4	-	-	4	-
CTQ	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Cyclopeptide	3	8	1	1	8	106	-	-	2	1	2	2	-	-	2	7	-	-	-	-
D-amino acid	15	-	1	-	1	-	-	-	-	5	4	-	3	13	-	5	1	7	-	2
D-amino acid and Hydroxylation	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2
D-amino acid and Thioether bond	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	5	1	-	-	-
Deamidation	-	-	37	-	-	-	-	15	-	-	-	-	-	-	-	-	-	-	-	-
Decarboxylation	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	3	-	-	-
Dehydroxylation	-	-	-	-	2	-	-	-	-	-	-	-	-	-	-	28	41	-	6	-
Diphthamide	-	-	-	-	-	-	-	-	5	-	-	-	-	-	-	-	-	-	-	-
Disulfide bond	-	-	-	-	1137	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
FAD	-	-	-	-	6	-	-	-	12	-	-	-	-	-	-	-	-	-	1	-
FMN	-	-	-	-	9	-	-	-	1	-	-	-	-	-	-	-	2	-	-	-
Formylation	-	-	-	-	-	1	-	-	-	-	-	3	52	-	-	-	-	-	-	-
Gamma-carboxyglutamic acid	-	-	-	-	-	-	448	-	-	-	-	-	-	-	-	-	-	-	-	-

16

Predict Phosphorylation

Pos.	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	0.07	0.07	0.07	0.06	0.07	0.07	0	0.04	0.06	0.06	0.06	0.07	0.07
R	0.08	0.09	0.08	0.15	0.09	0.07	0	0.04	0.05	0.06	0.06	0.07	0.07
N	0.03	0.03	0.04	0.03	0.04	0.04	0	0.02	0.03	0.03	0.03	0.03	0.03
D	0.05	0.06	0.05	0.06	0.06	0.08	0	0.08	0.08	0.08	0.07	0.07	0.06
C	0.01	0.01	0.01	0.01	0.01	0.01	0	0.01	0.01	0.01	0.01	0.01	0.01
G	0.07	0.07	0.07	0.07	0.06	0.09	0	0.05	0.08	0.07	0.07	0.06	0.07
E	0.08	0.08	0.08	0.07	0.07	0.06	0	0.07	0.11	0.13	0.1	0.09	0.09
Q	0.04	0.04	0.05	0.04	0.04	0.04	0	0.05	0.04	0.04	0.04	0.04	0.04
H	0.02	0.02	0.02	0.02	0.02	0.02	0	0.01	0.02	0.02	0.02	0.02	0.02
I	0.03	0.03	0.03	0.03	0.03	0.03	0	0.03	0.03	0.02	0.03	0.03	0.03
L	0.07	0.09	0.07	0.07	0.07	0.09	0	0.08	0.06	0.06	0.08	0.07	0.07
K	0.07	0.07	0.07	0.07	0.05	0.05	0	0.03	0.05	0.06	0.05	0.06	0.07
M	0.02	0.02	0.01	0.01	0.01	0.01	0	0.01	0.01	0.01	0.02	0.01	0.02
F	0.02	0.02	0.02	0.02	0.02	0.02	0	0.03	0.02	0.02	0.02	0.02	0.02
P	0.08	0.08	0.08	0.07	0.09	0.08	0	0.27	0.09	0.08	0.08	0.09	0.08
S	0.12	0.12	0.14	0.13	0.16	0.12	0	0.1	0.15	0.14	0.14	0.12	0.12
T	0.06	0.05	0.06	0.05	0.06	0.05	0	0.04	0.06	0.05	0.06	0.05	0.06
W	0	0	0	0	0	0	0	0.01	0	0	0	0.01	0.01
Y	0.02	0.02	0.02	0.01	0.01	0.02	0	0.01	0.01	0.01	0.01	0.02	0.02
V	0.05	0.05	0.05	0.04	0.04	0.04	0	0.04	0.05	0.04	0.05	0.05	0.05



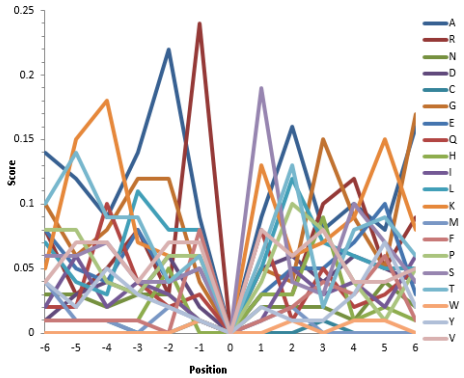
17

17

Methylation [Lysine]

*Modification will occur at K i.e. position 0

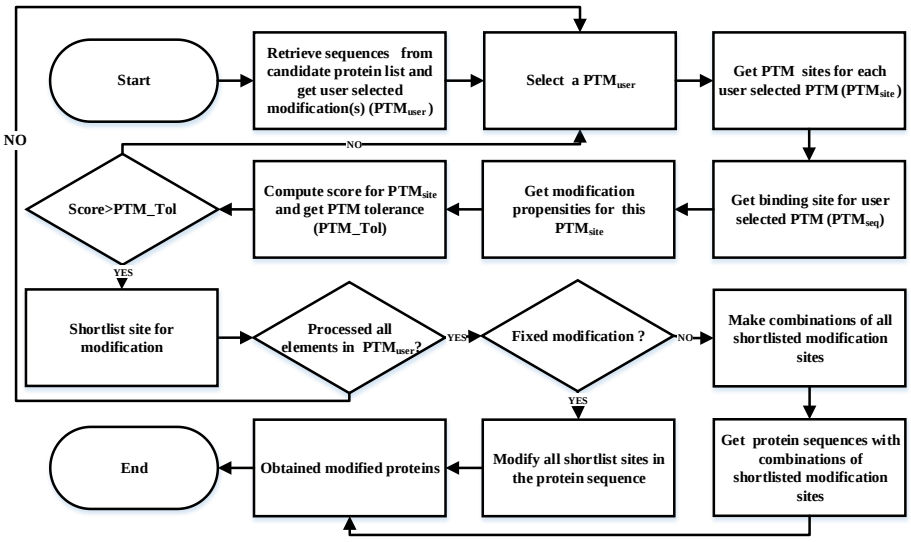
Pos.	-6	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
A	0.14	0.12	0.09	0.14	0.22	0.09	0	0.09	0.16	0.08	0.1	0.08	0.16
R	0.08	0.03	0.05	0.08	0.03	0.24	0	0.03	0.03	0.1	0.12	0.05	0.09
N	0.03		0.02	0.03	0.03	0.01	0	0.02	0.02	0.02	0.01	0.04	0.02
D	0.01		0.04	0.03	0.06	0.02	0	0.05	0.06	0.04	0.06	0.05	0.04
C	0.01	0.01	0.01	0	0	0.01	0	0	0	0.01	0	0	0
G	0.1	0.06	0.08	0.12	0.12	0.04	0	0.05	0.04	0.15	0.09	0.05	0.17
E	0.08	0.05	0.04	0.08	0.04	0.05	0	0.03	0.05	0.05	0.07	0.1	0.03
Q	0.02	0.02	0.1	0.04	0.02	0.03	0	0.08	0.01	0.05	0.02	0.03	0.05
H	0.01	0.01	0.01	0.01	0.05	0	0	0.03	0.03	0.09	0.01	0.02	0.01
I	0.02	0.06	0.02	0.04	0.03	0.01	0	0.01	0.05	0.03	0.04	0.02	0.06
L	0.07	0.04	0.03	0.11	0.08	0.08	0	0.05	0.12	0.07	0.06	0.05	0.05
K	0.05	0.15	0.18	0.07	0.06	0.06	0	0.13	0.06	0.07	0.09	0.15	0.08
M	0.04	0.01	0.01	0	0.02	0.01	0	0.01	0.02	0	0	0	0
F	0.01	0.01	0.01	0.01	0	0.08	0	0.01	0.02	0.04	0.03	0.06	0.01
P	0.08	0.08	0.04	0.03	0.06	0.06	0	0.04	0.1	0.08	0.04	0.01	0.05
S	0.06	0.06	0.07	0.04	0.04	0.05	0	0.19	0.04	0.03	0.1	0.07	0.04
T	0.1	0.14	0.09	0.09	0.04	0.06	0	0.06	0.13	0.02	0.08	0.09	0.06
W	0	0	0	0	0	0.01	0	0	0.01	0	0.01	0.01	0
Y	0.04	0.02	0.05	0.03	0.02	0.01	0	0.02	0.01	0.01	0.03	0.07	0.02
V	0.04	0.07	0.07	0.04	0.07	0.07	0	0.08	0.06	0.08	0.04	0.04	0.05



18

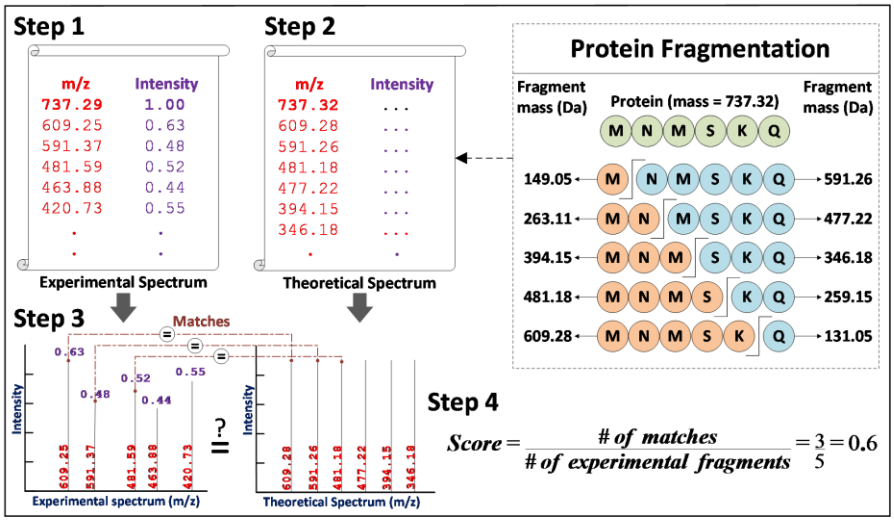
18

Post-translational Modifications



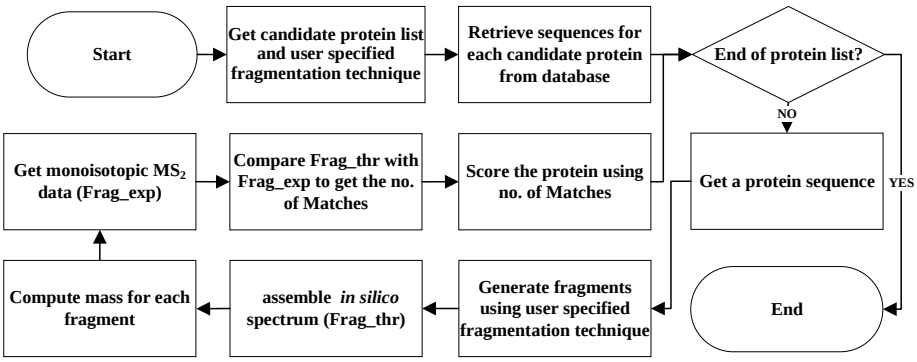
19

4. In silico Fragment Generation & Comparison



20

Spectral Comparison - Flowchart



Scoring Exp. & Thr. Peaks - I

- Upon computing the PST scores, the candidate list is further filtered for the highest scoring proteins.
- Finally, for each protein in this yet newer candidate list, we compute the theoretical fragments.
- Each proteins theoretical fragments is compared with the experimental fragments.
- Now, the question is, how to score?

Scoring Exp. & Thr. Peaks - II

1. Count the matches between thr. and exp. Peaks and give an equivalent score to the candidate protein

$$Score_{insilico} = \frac{Matches_{num}}{Frag_{exp}}$$

2. Weigh each of the aforementioned match by the mass error and abundance, and then accumulate the score

Computing Cumulative Scores - I

- So now we have obtained three individual scores
  1. Scores from MW Matches
  2. Scores from PST Matches
  3. Scores from Exp<>Thr Peak Matches
- It is necessary to compute an overall cumulative score (Why?)
- What are the options that we have? (Discussion!)

Scoring Scheme in SPECTRUM

$$\boldsymbol{Score\_MW} = \begin{cases} 1, & MWP_{Diff} = 0 \\ \frac{1}{2^{MWP_{Diff}}}, & 0 < \boldsymbol{ABS}(MWP_{Diff}) \leq Thr \\ 0, & MWP_{Diff} > Thr \end{cases}$$

$MWP_{Diff} = |Tuned\ mass - MWP|$

$$\boldsymbol{Score\_PST} = \sum_{i=0}^{M} PSTMatches_i \times (ErrorScore_i + FrequencyScore_i)$$

$FrequencyScore = Intensity \times LengthScore$

$LengthScore = N^2$

$Intensity = \frac{\sum_{i=1}^{N} Int\_AA_i}{N}$

$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Error_i)^2}{N}}$

$ErrorScore = e^{-RMSE \times 2}$

$$\boldsymbol{Insilico\ Score} = \frac{No\ of\ matches}{Number\ of\ Experimental\ Fragments}$$

$$\boldsymbol{Score\_Final} = \left.\begin{array}{l} (Score\_MW \times W1) + \\ (Score\_PST \times W2) + \\ (Score\_Insilico \times W3) \end{array}\right)$$

Computing Cumulative Scores - II

- Simply sum the scores up (a linear function)

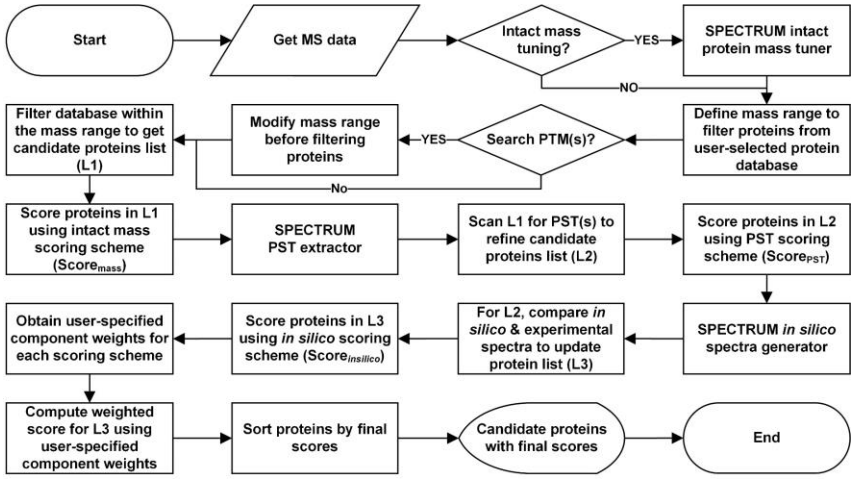
$$Score_{MW} + Score_{PST} + Score_{Exp <> Thr} = Score$$

- Weigh each scoring component up by respective RMSE before summing them up

$$Score_{final} = (Score_{mass} * W_1) + (Score_{PST} * W_2) + (Score_{insilico} * W_3)$$

- Develop a non-linear function to integrate the scoring components (e.g. Mascot etc)
  - Highly proprietary for commercial proteomics software

Overall Search Flowchart



Home Task

- Read

SPECTRUM – A MATLAB Toolbox for Proteoform Identification from Top-Down Proteomics Data

- https://www.nature.com/articles/s41598-019-47724-1