Scoring Intact Protein Mass

Scoring Scheme:

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

Example: A Protein Sequence from Database "MQLF"

MW: 537.67

http://web.expasy.org/compute_pi/

All experimental masses are compared with theoretical masses of database and mass is selected on the base of closeness.

128 Protein Fragmentation Techniques

We compare the experimental mass with theoretical data base mass of protein and on base of closeness we rank or score it.

If several proteins have same score than selection is done by using another technique protein fragmentation. We fragment the protein or peptide and ionize it, it helps us to measure the fragment masses as the same ways as their precursor.

There are different techniques for protein fragmentation.

- **Electron Capture Dissociation (ECD)**
- **Electron Transfer Dissociation (ETD)**
- **Collision Induced Dissociated (CID)**

Each fragmentation technique gives result of specific type of fragments.

ECD gives out 'C' and 'Z' ions. CID gives out 'B' and 'Y' ions, etc.

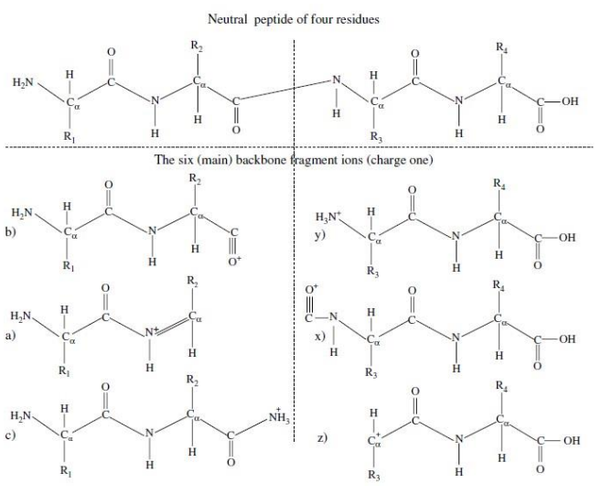
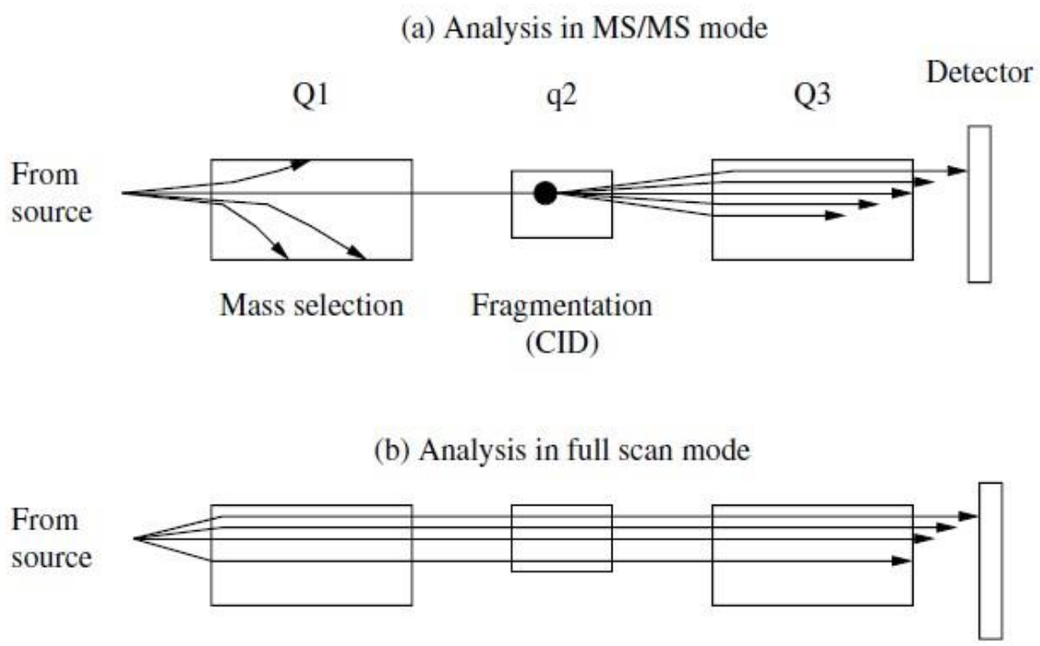


Figure 0.78 natural peptide of four residue

If we can measure the mass of fragments using MS, Calculate the theoretical mass of the fragments. Then, we can award score on the basis of the similarity of experimental and theoretical mass.

129 Tandem MS

Intact masses can measure the intact proteins or peptides. And this can be followed up by their fragmentation in MS chamber.



Tandem MS can be extended to the fragments of the intact fragment. All you need is the MS instrument capability to,

(i) select fragment's mass range.

(ii) Fragment the precursor fragment.

Tandem MS helps us to measure masses of fragments. By this scoring and protein identifications so easy.

130 Measuring Experimental Fragment's Mass

In MS1, the molecular weight of intact sample molecule is measure and then intact molecule is fragmented in two afterward, these two fragments are measured by MS or MS2

FRAGMENTATION TECHNIQUES AND MOLECULAR WEIGHT

Fragmentation techniques include ECD, CID etc. intact molecule fragmentation splits the molecule into two parts.

FRAGMENT MASS

Mass of fragment is produced by MS2 deepening upon the technique because each techniques splits the protein or peptide at different location.

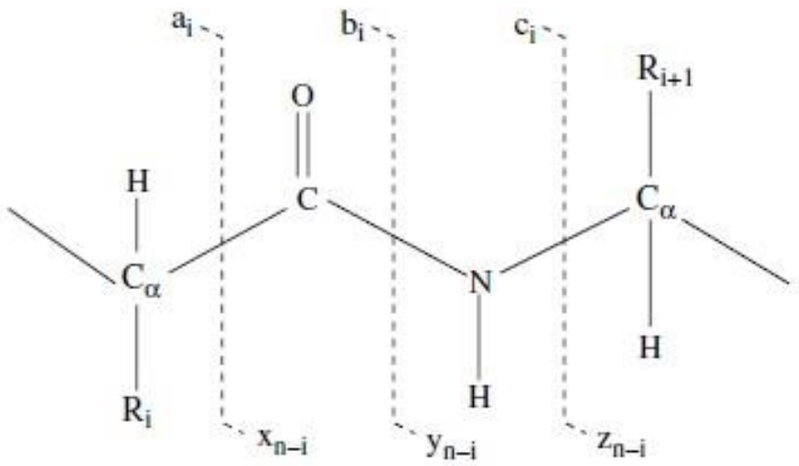


Figure 0.89 Masses after Fragmentation by ECD, CID & ETD

Experimental mass reported from MS2 is matched with theoretical peptides of candidate proteins (from DB). Score is awarded on the basis of the closeness between experimental and theoretical masses.

131 Calculating Theoretical Fragment's Mass

After measuring the mass of intact molecule from MS2 we compare that mass with theoretical mass of databased proteins.

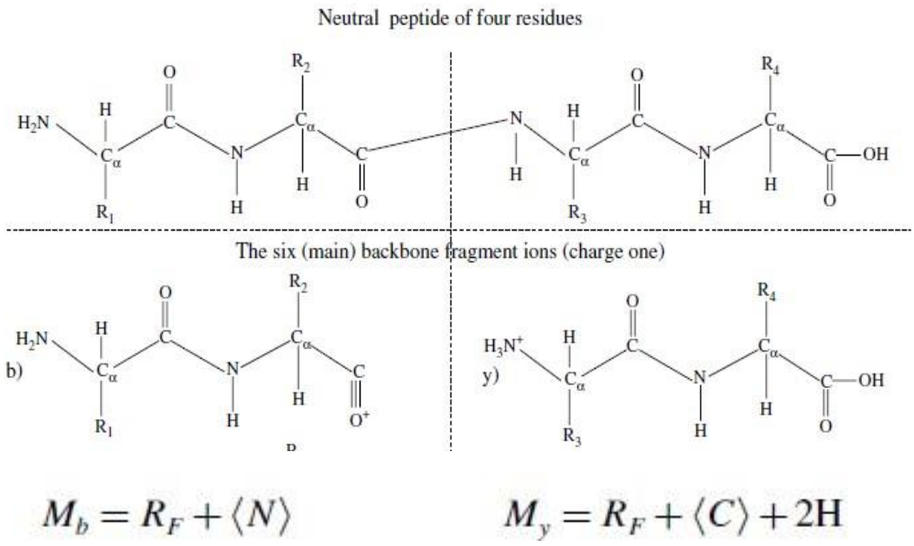


Figure 20 Masses after Fragmentation by CID

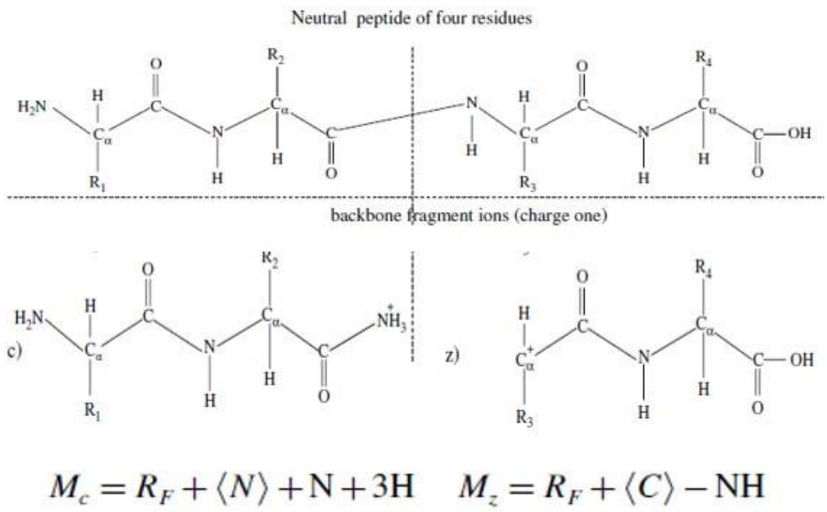


Figure 0.91 Masses after Fragmentation by ECD

$$M_b = R_F + \langle N \rangle \qquad M_y = R_F + \langle C \rangle + 2\mathrm{H}$$
$$M_a = R_F + \langle N \rangle - \mathrm{CO} \qquad M_x = R_F + \langle C \rangle + \mathrm{CO}$$
$$M_c = R_F + \langle N \rangle + \mathrm{N} + 3\mathrm{H} \qquad M_z = R_F + \langle C \rangle - \mathrm{NH}$$

$$M_b + M_y = R_P + \langle N \rangle + \langle C \rangle + 2\mathrm{H} \; (= M_P + 2\mathrm{H})$$
$$M_a + M_x = R_P + \langle N \rangle - \mathrm{CO} + \langle C \rangle + \mathrm{CO} \; (= M_P)$$
$$M_c + M_z = R_P + \langle \mathrm{N} \rangle + \mathrm{N} + 3\mathrm{H} + \langle C \rangle - \mathrm{N} - \mathrm{H} \; (= M_P + 2\mathrm{H})$$

Figure 22 Masses after Fragmentation by ECD

132 Peptide Sequence Tags

Peptide sequence tag are the sequence of peptide which are produced after MS2. We can obtain the sequence of peptide through variation in fragmentation site.

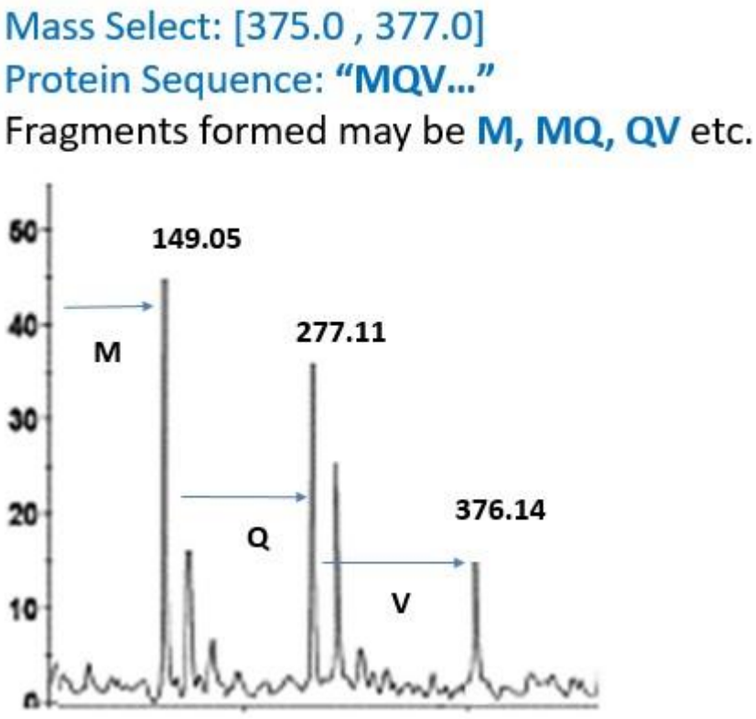


Figure 23 variation sites

Precursor proteins or peptides fragmentation leads to formation of multiple ions of the same fragment type. However, fragments have variation in their molecular weights due to variation in site of fragmentation

Fragmentation at consecutive sites leads to a mass difference equal to that of a single amino acid. Such consecutive peaks can reveal partial peptide sequence tags

133 Extracting Peptide Sequence Tags

PSTs are formed due to sequential cleavage of precursor protein/peptide's backbone.

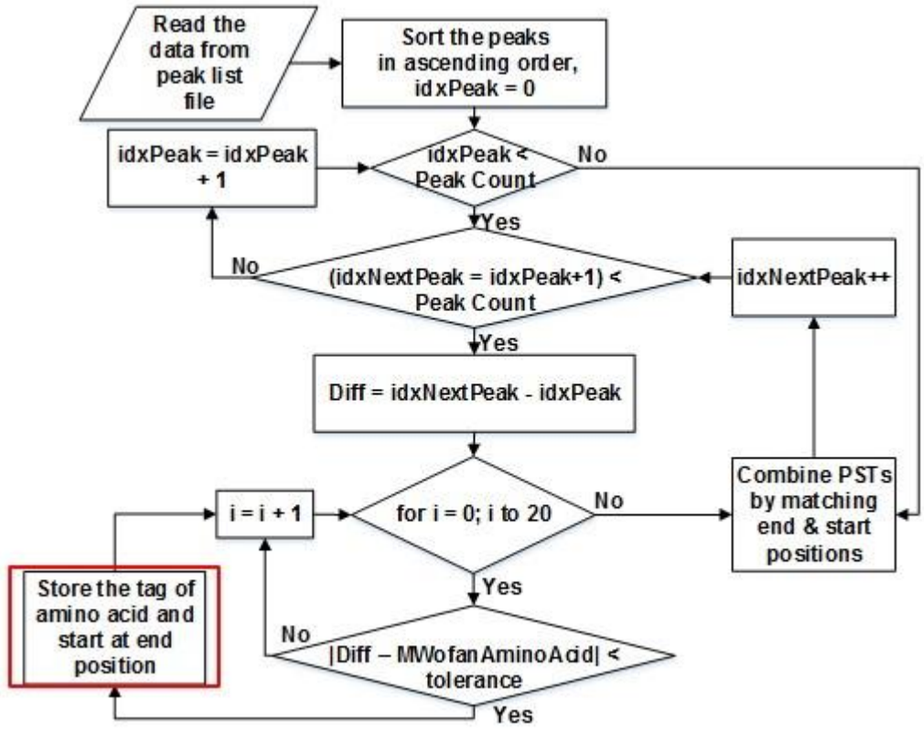


Figure 24 peptide sequencing tagging

Peptide sequence tags can be extracted from peak list iteratively. A high quality mass spectrum will produce large number of PSTs. The bigger the peptide sequence tags, the better!

134 Using Peptide Sequence Tags in Protein Search

PSTs provide clues of the precursor protein/peptides sequence. Consider that we extract the following PSTs: M, MQ, QV etc. Search protein sequence database (e.g. Uniprot, Swissprot)

Sample sequence in protein DB

>>sp|Q6GZ4X|0X1R_FRG3G Putative transcription factor 0X1R OS=Random virus 3 (isolate Goorha) GN=FV3-0X1R PE=4 SV=1

MAFSAEDVLKEYDRRRRMEALLLSLYYPNDRKLLDYKEWSPPRVQVECPKAPVEWNNPPSEKGLIVG
HFSGI
KYKGEKAQASEVDVNKMCCWVSKFKDAMRRYQGIQTCKIPGQVLSDLDAKIKAYNLTVEGVEGFVRYS
RVT
KQHVAAFLKELRHSKQYENVNLIHYILTDKRVDIQHLEKDLVKDFKALVESAHRMRQGHMQNVKYILYQ
LLK KHGHGPDGPDILTVKTGSMQLYDDSFRKIYTDLGWKFTPL

For all the proteins in the database, we find out which PSTs exist in which proteins. The protein reporting the most PSTs is more probable to be the precursor protein.

If many PSTs report the same number or protein report the longer PSTs than through scoring we find the greater number. After extracting the PSTs we search the entire database for protein who report it.

135 Scoring Peptide Sequence Tags

According to scoring scheme if a candidate protein matches ‘n’ PSTs, then its score can be given by:

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

Additionally, if we include RMSE to the scoring system, then it can highlight better PST matches. And RMSE is the root mean square error.

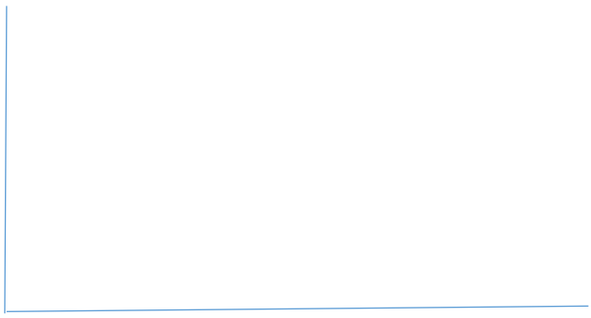


Figure 25 root mean square error

RMSE for a sequence tag 'i' of length 'n'?

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

So, the updated relationship is:

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

136 In silico Protein Fragmentation

MS1 reports the intact mass of molecules (proteins or peptides) in the sample. Intact mass can be compared with every protein's mass in database to identify the molecule in the sample.

Incase multiple candidate proteins are reported, MS2 can be performed. MS2 helps measure fragment peptide masses. MS2 data can be used to extract peptide sequence tags

If the protein identification is still not conformed than each experimentally reported MS2 fragment is compared with the in silico spectra of proteins from database.

Fragmentation techniques determine product ions e.g. ECD -> c/z and CID -> b/y ions etc. With known fragment types, we can compute the MW of all possible protein fragments

For obtaining all possible theoretical fragments in a protein, we need to compute the MWs of each fragment individually

Consider a random protein sequence from DB:

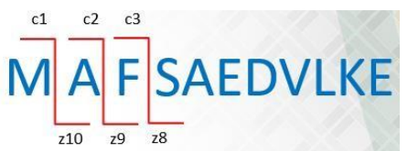


Figure 26 random protein sequence

Matching experimental fragments with in silico fragments is the final resort in protein search and identification.

137 In silico Fragment Comparison and Scoring

Experimental MS2 can be compared with the in silico spectra of protein from database.

- Count the matches between in silico and in-vitro peaks.
- Give an equivalent score to candidate protein.
- Weigh each of the aforementioned match by the mass error

- Accumulate the score

With “all possible” fragments in in silico spectrum, and “reported” fragments in experimental spectrum, we can match and rank.

Scoring scheme should also consider the errors in peak matching

138 Protein Sequence Database Search Algorithm

MS1 and MS2 provide us with a host of data towards enabling us in identifying unknown proteins. A step by step approach combining MW, PSTs and insilico spectral matching is required.

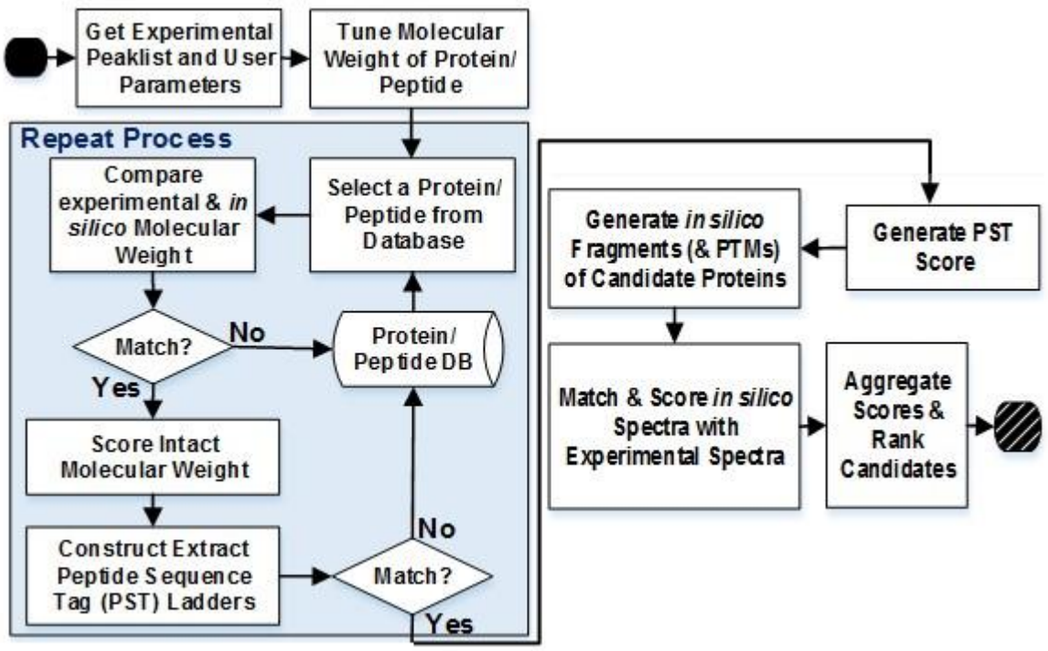


Figure 27 protein sequencing flowchart

Integrating MW, PST and insilico comparison algorithms in a workflow can help create a composite protein search engine. A composite scoring system is also required for this search engine.

139 Integrative Scoring Schemes

Three individual scores can be obtained:

- MW Match score
- PST Match score
- In vitro & In silico Match score

For overall cumulative score computation:

- We 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

$$Score = \frac{Score_{MW}}{E_{MW}} + \sum_{i=0}^{m} \frac{Score_{PST}}{RMSE_{PST}} + \sum_{i=0}^{n} \frac{Score_{Exp <> Thr}}{E_{EXP <> Thr}}$$

- Complex non-linear functions integrate the scoring components in Mascot etc.
- Highly proprietary for commercial proteomics software are used.

Composite scoring schemes are needed to combine scores coming in from multiple criteria. The ability of a scoring scheme to better isolate true positives from false positives is important.

140 Large Scale Proteomics

Peptide mixtures in bottom up proteomics are very complex. Tryptic peptides may reach up to an order of 300,000–400,000.In whole proteome samples, protein count may be over 10,000. Experiments have shown that it is difficult or even impossible to analyze all these peptides in a single analysis, as the mass spectrometer is essentially overwhelmed.

Over half a million peptides reported in a typical LSP experiment are redundant.

If we could find a unique peptide for a protein, that would make sequence coverage suffer and we have to strike a compromise between sequences coverage and sample coverage.

TECHNIQUE

One way forward would be to transfer peptides to the MS chamber in a step-by-step manner. However, this imposes a precondition that a peptide is not selected earlier as well (i.e. more than once)

STEP BY STEP TECHNIQUE

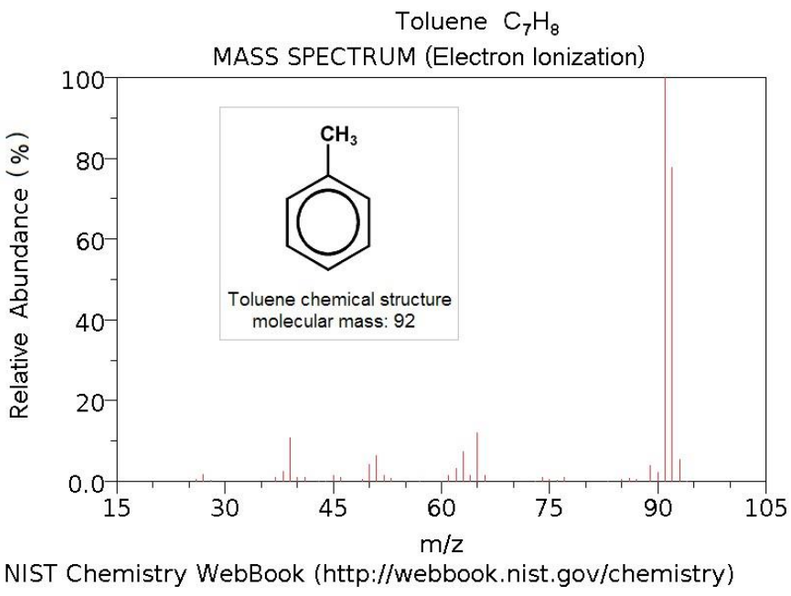
1. The instrument alternates between MS and MS/MS modes.
2. Three most intense peaks are chosen for MS/MS analysis.
3. After the initial MS scan, an MS/MS spectrum from peptide A is obtained by selectively fragmenting this mass only.
4. Next, a spectrum for peptide B is produced, followed by a recording of the MS/MS spectrum for peptide C.
5. After these three fragmentation spectra have been obtained, a new MS scan is started.

From this scan, three more peptides A B C are selected for fragmentation and the cycle starts over again.

The number of MS1/2 scans can be limited by carefully selecting the peptide peaks. Once the intense peptides are identified, next batch of peptides is chosen for MS2.

141 Proteomics Data File Formats

Mass spectrometer is used to measure mass/charge ratio of ionized proteins and peptides. Data output from the MS comprises of m/z ratios and intensities of each molecule that is measured.



Followings are the formats for proteomics data:

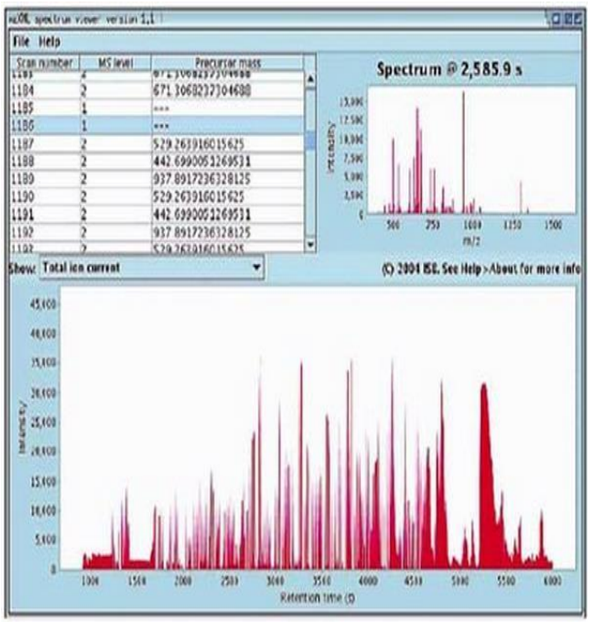
Agilent	.D/.YEP
Bruker	.BAF
ABI/Sciex	.WIFF
	.t2d
Thermo Xcalibur	.RAW
Micromass	
PerkinElmer	
Waters	
Shimadzu	.QGD

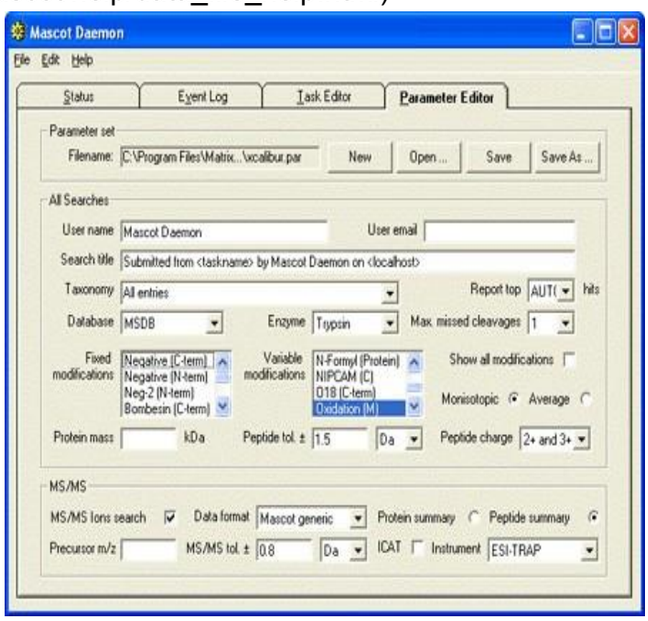
Figure 28 Formats used for proteomics data

OPEN FORMTAS:

mzXML (tools.proteomecenter.org/mzXMLViewer.php)

MGF (proteomicsresource.washington.edu/mascot/help/data_file_help.html)





Multiple MS data formats exist. Proprietary formats exist which come implemented as software with hardware. Also, open software standards exist for interoperability etc.

142 RAW File Format

Mass spectrometer outputs data with ionic mass/charge ratios & respective ion intensities. RAW file is a format in which an instrument outputs data in binary form.