

The dot plot method for RNA structure prediction is easy. Draw a square and partition by drawing gridlines. Put RNA sequence on top and left sides of the square. Put a "DOT" on complementary nucleotides

For example:

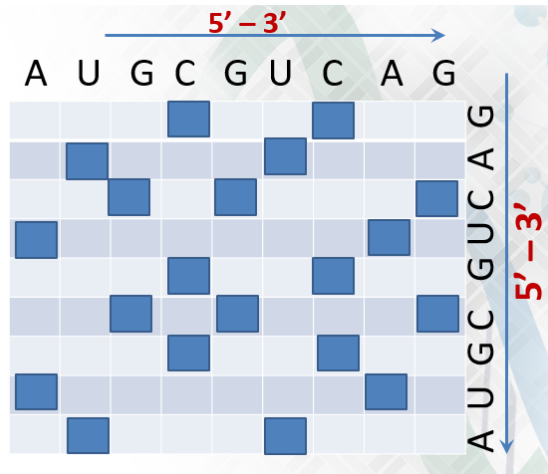


Figure 16 dot are placed at complimentary base pair place.

Connect regions of paired nucleotides to form 2' structures in following image.

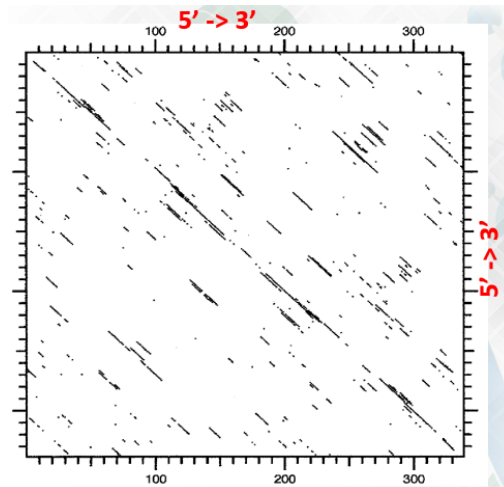


Figure 17 Potato Tuber Spindle Viroid

In longest RNA nucleotides the gaps between complementary nucleotides becomes bulges and loops of the structure.

Module 014: ENERGY BASES METHODS

Experimental prediction of RNA structure is slow and costly that's why a few 2' RNA structures are reported experimentally.

While prediction we get many possible 2' structures of RNA and for optimal structure selection we calculate their overall stability.

	Y:	A	C	G	U

X: A		.	.	.	-2.1
C		.	.	-3.3	.
G		.	-2.4	.	-1.4
U		-2.1	.	-2.1	.

Figure 18 energy table

- **STABILIZING ENERGY**

Energy table helps us to find the optimal prediction of structure because energy is released when complementary nucleotides make bonds.

- **DESTABILIZING ENERGY**

Remaining unpaired nucleotides destabilized the RNA structure in form of hairpin or bulge structure.

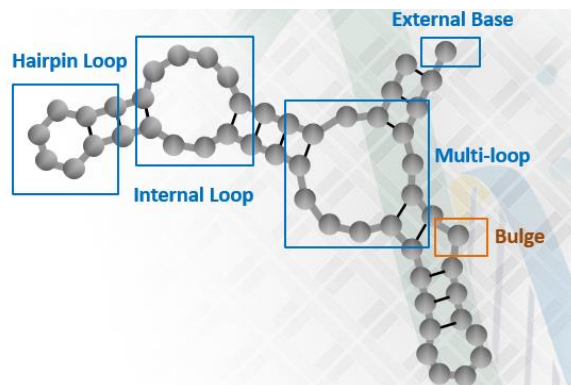


Figure 19 Hairpin+IntLoop+ExtLoop+Bulge+Hairpin

- **SUM OF ENERGIES**

Sum of stabilizing and destabilizing energies can help determine the quality of a 2' RNA structure. 2' structure with longest coupled sequences vs. one with lowest energy

Module 015: Zuker's Algorithm

Energy based methods involve evaluating the free energy structures. To compute the RNA sequence for 1' or 2' optimal structure prediction we use Zuker's Algorithm.

Zuker's Algorithm helps us to compute the stabilizing energies (-ve) and also destabilizing energies (+ve values). And also compute the sum of +ve and -ve energies.

	Y:	A	C	G	U	

X:	A		.	.	.	-2.1
	C		.	.	-3.3	.
	G		.	-2.4	.	-1.4
	U		-2.1	.	-2.1	.

Figure 20 stacking energies

SIZE	INTERNAL	BULGE	HAIRPIN
1	.	3.8	.
2	.	2.8	.
3	.	3.2	5.6
4	1.7	3.6	5.5
5	1.8	4.0	5.6
6	2.0	4.4	5.3
7	2.2	4.6	5.8
8	2.3	4.7	5.4
		...	
30	3.7	6.1	7.7

Figure 21 destabilizing energies

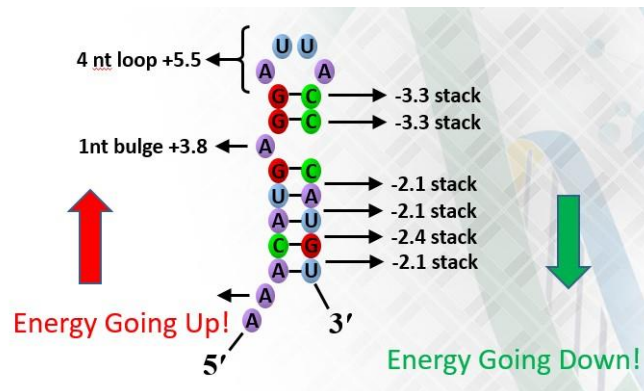


Figure 22 working principle of Zuker's Algorithm (2003)

It Compute energies of all possible 2' structures. Generate combinations of all computed 2' structures. Select the one with lowest energy.

Module 016: Zuker's Algorithm EXAMPLE

Zuker's Algorithm involves computing stabilizing and destabilizing energies of a 2' structure. All possible 2' structures are generated. The best 2' structure is selected!

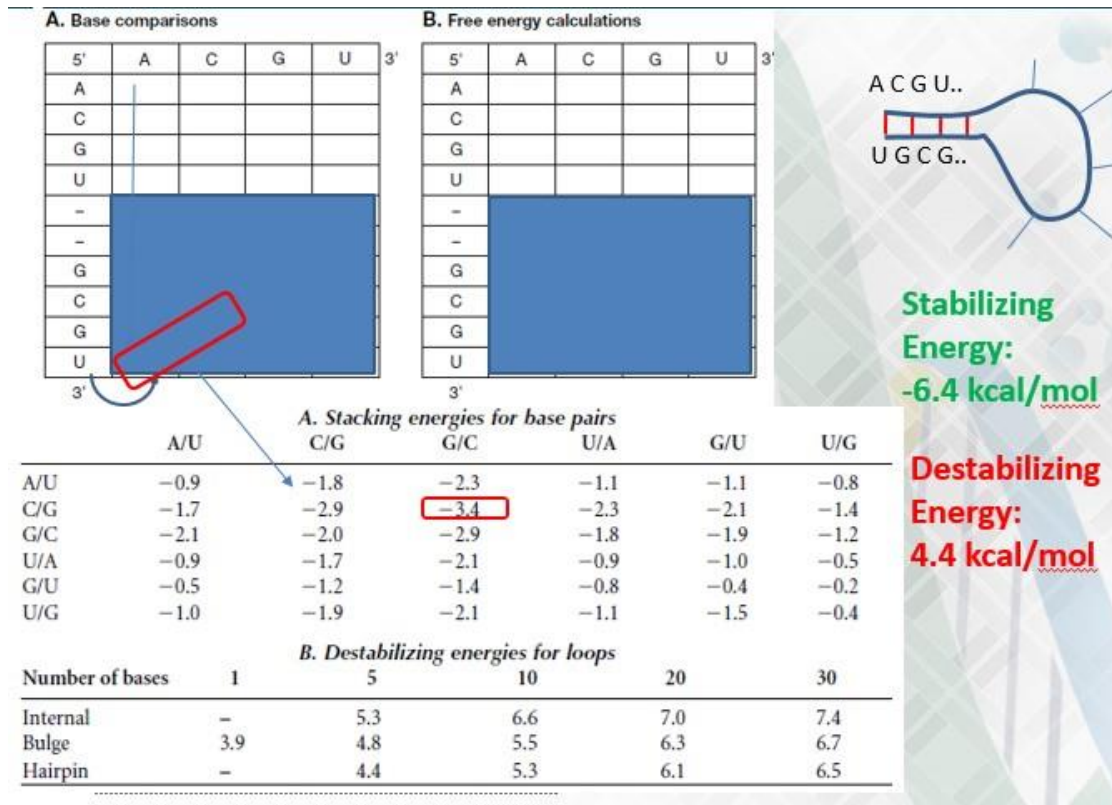


Figure 23 Calculation of all possible structure combinations

We need to construct all the possible combinations of nucleotides for selection of optimal 2' RNA structure.

Module 017: Zuker's Algorithm – A Flow Chart

Zuker's Algorithm involves computing stabilizing and destabilizing energies of 2' structure. And it also computes the overall energy by summing up the positive and negative energies.

A. Base comparisons

5'	A	C	G	U	3'
A					
C					
G					
U					
-					
-					
G		C/G		U/G	
C			G/C		
G		C/G		U/G	
U	A/U	C/U	G/U		
3'					

B. Free energy calculations

5'	A	C	G	U	3'
A					
C					
G					
U					
-					
-					
G				-6.4	
C			-5.2		
G		-1.8			
U					
3'					

The two diagonals ('D') given above include:

1. **A/U, C/G, G/C, U/G**
2. **G/U, U/G**

The flow chart for energies is:

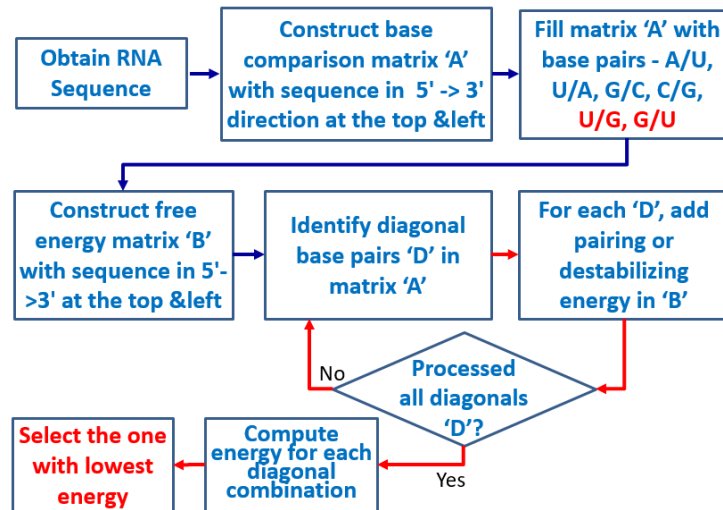


Figure 24 flow chart

The diagonal combination from all possible is selected with overall lowest energy.

Module 018: Martinez Algorithm

Zuker's Algorithm involves computing stabilizing and destabilizing energies of 2' structure. And it also computes the overall energy by summing up the positive and negative energies. Martinez Algorithm is improvement on it.

Making combination of all possible structures is time consuming, Martinez Algorithm favors those 2' structure which are energetically more feasible.

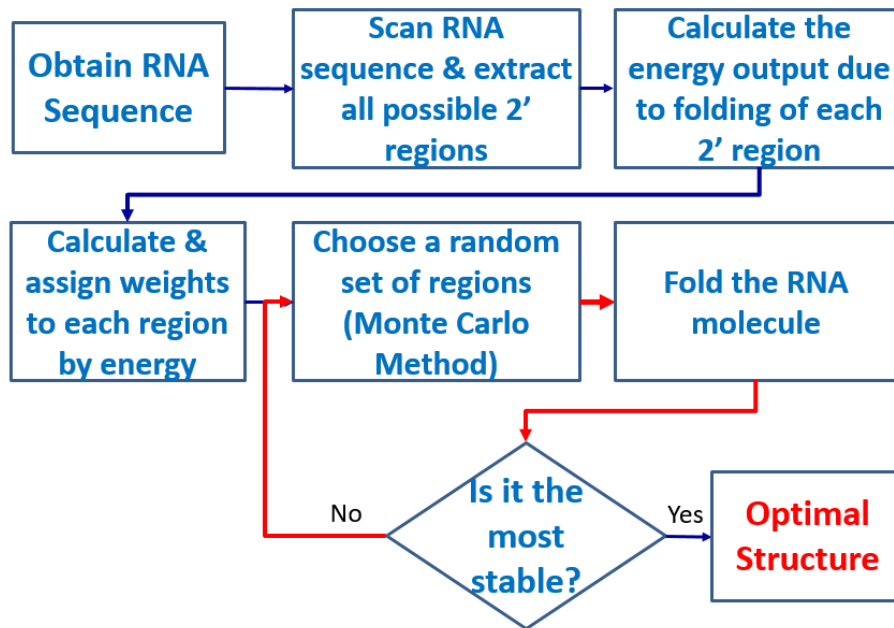


Figure 25 Martinez Algorithm flow chart

In Martinez algorithm all the 2' structures are weighed by its stability and optimal one is sorted out. Monte Carlo methods (or Monte Carlo experiments) are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. They are often used in physical and mathematical problems and are most useful when it is difficult or impossible to use other mathematical methods.

And Monto Carlo method do not provide a definitive solution.

Module 019: Dynamic Programming Approaches

RNA sequence contains 4 type of nucleotides G/C, G/U and A/U and it may contain hundreds of nucleotides it means there is possibility of many combinations.

In 2' RNA structure there may be large number of nucleotide sequence with large number of combinations hence it is hard to find the optimal one and for this prediction we use Dynamic Programming (DP) which breaks the larger problems into smaller one.

- **PRINCIPLE OF DYNAMIC PROGRAMMING**

For optimal structure combination selection we use the Dynamic Programming (DP) and we select the sequence of RNA nucleotides and list all the possible complementary positions for nucleotides in the given complete sequence.

For example:

AUGCUGUCAUCG
 AUGCUGUCAUCG
 AUGCUGUCAUCG
 AUGCUGUCAUCG

Figure 26 all possible complementary bases combinations.

Dynamic Programming then recombines such combinations in a process called “**Traceback**” to ensure that the highest coupled 2' structure is reported

Module 020: Nussinov-Jacobson Algorithm – An Overview

Nussinov-Jacobson (NJ) Algorithm is a Dynamic Programming (DP) strategy to predict optimal RNA 2' structures, Proposed in 1980. Computes 2' structures with most nucleotide coupling.

http://ultrastudio.org/en/Nussinov_algorithm

• HOW IT WORKS

- Create a matrix with RNA sequences on top and right
- Set diagonal & lower tri-diagonal to zero
- Start filling each empty position in matrix by choosing the maximum of 4 scores

		J								
		1	2	3	4	5	6	7	8	9
I		G	G	C	A	A	A	U	G	C
1	G	0								
2	G	0	0							
3	C		0	0						
4	A			0	0					
5	A				0	0				
6	A					0	0			
7	U						0	0		
8	G							0	0	
9	C								0	0

Figure 27 A. Note the I and J labels, B. Initialize tri-diagonal and lower tri-diagonals to zero.

The score $S(i, j)$ is the maximum of the following four possibilities

$$S(i, j) = \max \left\{ \begin{array}{lll} S(i+1, j) & \text{Lower Element} & r_i \text{ unpaired} \\ S(i, j-1) & \text{Left Element} & r_j \text{ unpaired} \\ S(i+1, j-1) + e(r_i, r_j) & \text{Energy of pairing} & i, j \text{ base pair} \\ \max_{i < k < j} \{S(i, k) + S(k+1, j)\} & & i, j \text{ paired, but not to each other} \end{array} \right.$$

Left Row Bottom Column

Module 021: Nussinov-Jacobson Algorithm – The Flowchart

NJ algorithm is actually a dynamic programming (DP) approach to predict the 2' RNA structure. A scoring matrix is initialized to record scores in NJ Algorithm. For filling scoring matrix, the maximum score from 4 matrix positions is chosen.

D(i,j)	G	G	G	A	A	A	U	C	C
G	0	0	0	0	0	0	0	1	1
G	0	0	0	0	0	0	0	1	1
G		0	0	0	0	0	0	1	1
A			0	0	0	0	1	0	0
A				0	0	0	1	0	0
A					0	0	1	0	0
U						0	0	0	0
C							0	0	0
C								0	0

Figure 0.3 for maximum score 4 positions are used in scoring

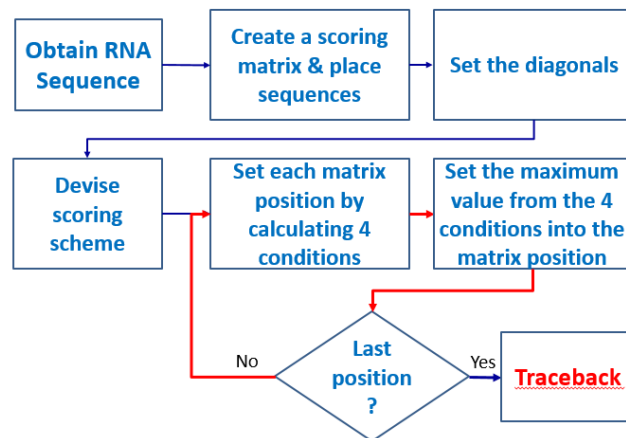


Figure 28 flow chart of NJ Algorithm.

Traceback is used to report the coupling of structures in sequences.

Module 022: Nussinov-Jacobson Algorithm – EXAMPLE

The main points to be focused in N-J Algorithm are:

- Scoring Matrix
- Matrix Initialization
- Scoring method
- The 4 different positions to be considered for calculating matrix

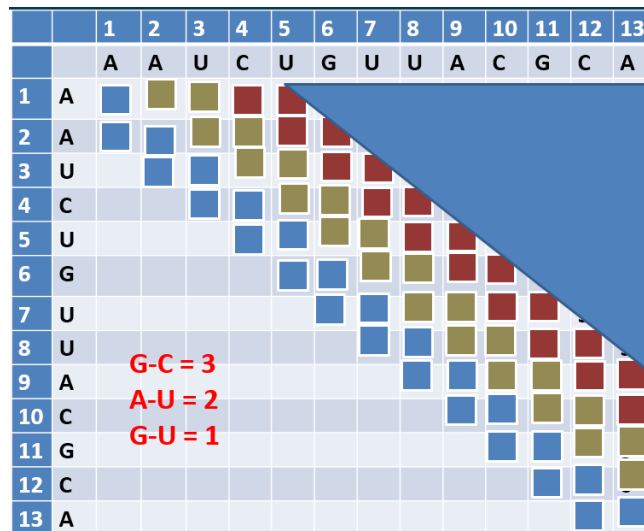


Figure 29 N-J Algorithm scoring

The matrix is filled by four different positions. **Left, Bottom, Diagonal, and Left/Bottom elements.** In this way all complementary nucleotides coupling is catered.

Module 023: Score Calculations & Traceback

From four positions the score is calculated and from each position we calculate the score contribution. And maximum score is sorted out.

		1	2	3	4	5	6	7	8	9	10	11	12	13
		A	A	U	C	U	G	U	U	A	C	G	C	A
1	A	0	0	2	2	4	5	7	7	9	9	12	12	12
2	A	0	0	2	2	2	5	5	5	7	7	10	10	12
3	U		0	0	0	0	3	3	3	5	5	8	8	10
4	C			0	0	0	3	3	3	5	5	8	8	10
5	U				0	0	1	1	1	3	5	6	8	10
6	G					0	0	1	1	3	5	6	8	8
7	U						0	0	0	2	2	5	5	7
8	U							0	0	2	2	5	5	5
9	A								0	0	0	3	3	3
10	C									0	0	3	3	3
11	G										0	0	3	3
12	C											0	0	0
13	A												0	0

$G-C = 3$
 $A-U = 2$
 $G-U = 1$
 4th condition of $\max\{S(i,k)+S(k+1,j)\}$.

Figure 30 scoring and traceback in N-J Algorithm.

There can be many traceback. Each traceback is used to make the RNA secondary structure. And traceback with highest number of nucleotide coupling is selected.

Module 024: Comparison of Algorithms

RNA has three different structures 1', 2' and 3'. For these structures predictions there are many algorithms. But in all algorithm there are two main strategies:

1. Nucleotides stacking
2. Energy minimization

- **ENERGY BASED ALGORITHM.**

Zuker's Algorithm involves energy minimization. It is updated version and incorporate the phylogenetic information. It is improved. Overcomes the pseudoknots assumes them and accommodate them. And this algorithm helps to predict the structures of RNA based on nucleotides.

- **NUCLEOTIDES STACKING ALGORITHM.**

NJ's Algorithm comes under this category. It involves the maximizing the nucleotides pairing. Traceback helps to find best 2' structure.

It predict the 75% accurate 2' structure. Because there may be more than two equal scores as it is calculated from four different positions. To get best results we need to combine the stacking and energy minimization methods together.

For further improvements in results we take help from:

- Sequences
- Comparison
- Nucleotide
- Covariance analysis

Module 025: WEB RESOURCES-I

For prediction of 1' and 2' structure of RNA we use different algorithm like Zuker's, Martinez and N-J. Online tools also.

The mfold web server is one of the oldest web servers in computational molecular biology. Mfold is upgraded version of Zuker's algorithm.

MFOLD is computationally expensive and can give results for 1' and 2' structures that have sequences less than 8000 nucleotides.

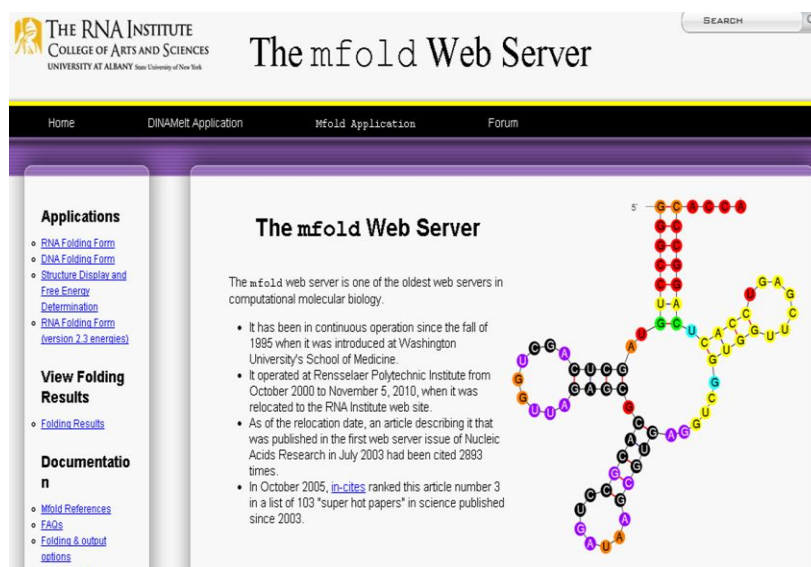


Figure 31 <http://unafold.rna.albany.edu/?q=mfold>

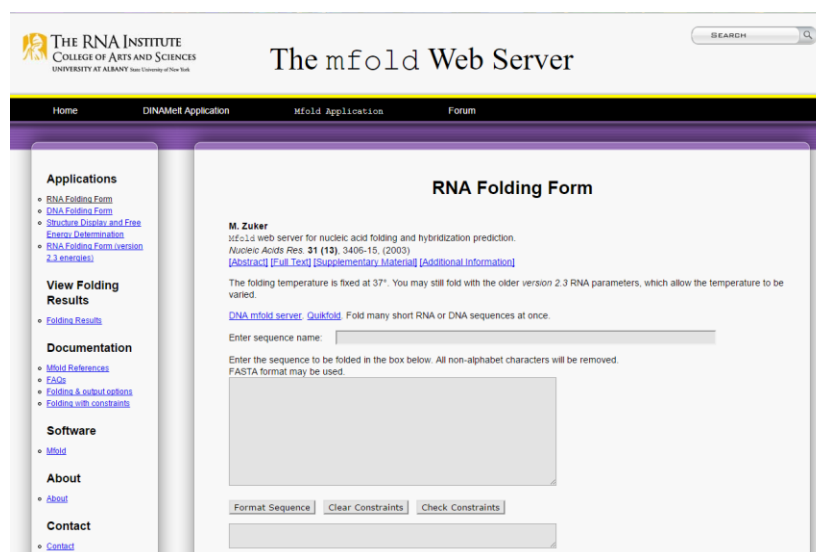


Figure 32 <http://unafold.rna.albany.edu/?q=mfold/RNA-Folding-Form>

The screenshot shows the mfold Web Server interface. The header includes the logo of The RNA Institute, College of Arts and Sciences, University at Albany, and the title 'The mfold Web Server'. A search bar is located in the top right. The navigation bar has links for Home, DINAMelt Application, mfold Application, and Forum. The left sidebar contains links for Applications (RNA Folding Form, DNA Folding Form, Structure Display and Free Energy Determination, RNA Folding Form version 2.3 energies), View Folding Results (Folding Results), Documentation (mfold References, E2DE, Folding & output options, Folding with constraints), Software (Mfold), About (About), and Contact (Contact). The main content area is titled 'Structure Display and Free Energy Determination' and includes a sub-header 'efn server - 1996-2016, Michael Zuker, Rensselaer Polytechnic Institute'. Below this is a 'View sample input and output.' link. The form contains several input fields: 'Enter a name for your structure:' with a text box; 'Cl file to upload:' with a 'Choose File' button and 'No file chosen' text; 'Vienna format file to upload:' with a 'Choose File' button and 'No file chosen' text; 'Is the sequence linear or circular?' with a dropdown menu set to 'linear'; 'Temperature (between 0° and 100° C):' with a text box set to '37'; 'What style of energies should be used?' with a dropdown menu set to 'mfold'; 'RNA free energy parameters:' with radio buttons for '3' (selected) and '2.3', and a note 'Version 3 parameters are for 37 deg only.'; 'Select nucleic acid type:' with a dropdown menu set to 'RNA'; 'Ionic conditions: [Na+]' with a text box set to '1.0'; '[Mg++]' with a text box set to '0.0'; 'Units: M' (selected) and 'mM'; and a 'Send data for processing.' button with a green smiley face icon. A 'Reset form' button is also present.

Figure 33 <http://unafold.rna.albany.edu/?q=mfold/Structure-display-and-free-energy-determination>

MFOLD helps fold an RNA nucleotide sequence into its possible 2' structures. MFOLD gives out several structures along with their energetic stability!

Module 026: WEB RESOURCES-II

RNA nucleotides folds to form 2' structure from simple portion of 1' nucleotides. For example CUUCGG occurs a wide variety in RNA and it mostly forms the stable hairpin loop. So we can make the list of all likely 2' structures arising from 1'.

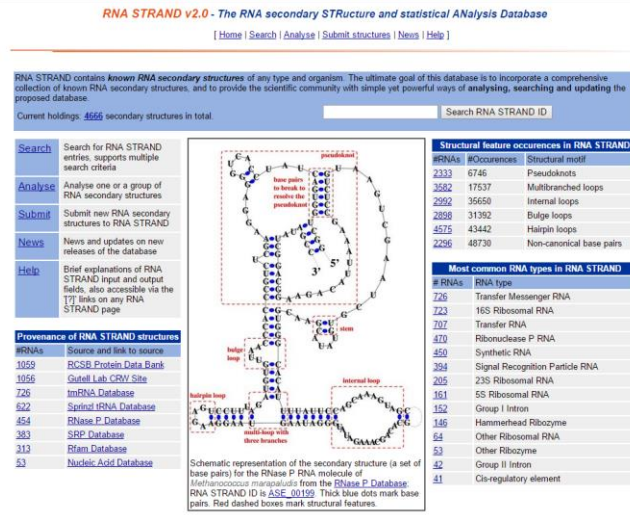


Figure 34 <http://www.rnasoft.ca/strand/>

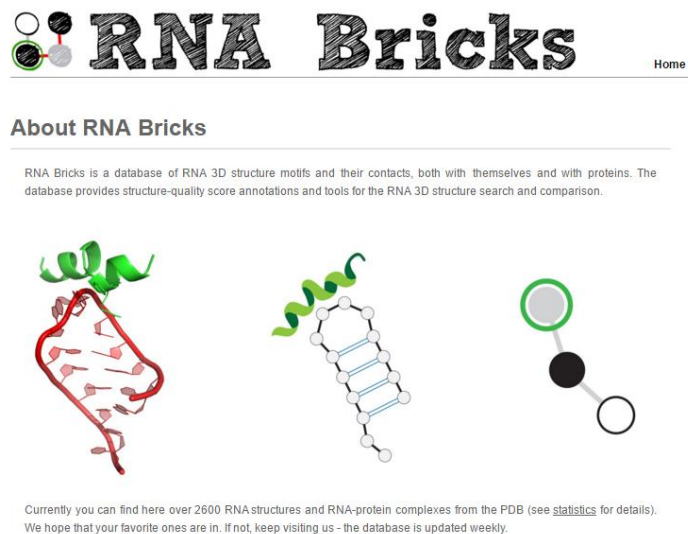


Figure 35 <http://iimcb.genesilico.pl/rnabricks>

RNA 1' folds and makes RNA 2' structure and this online database is established for 2' RNA structure and it act as dictionary for 2' RNA structure.

Chapter 5 - Protein Sequences

Module 001: FROM DNA/RNA SEQUENCES TO PROTEIN

We are aware that DNA has four nucleotides bases (**A, C, T & G**). RNA contains (**A, C, U & G**). And protein contain 20 different amino acids. **DNA to RNA then Protein** is called as **central dogma**. Which includes translation, transcription and protein modifications.

A set of three nucleotides called codon, codes the information for specific amino acids in protein synthesis.

Amino Acid	3-Letters	1-Letter			
Methionine			Methionine	Met	M
Alanine	Ala	A	Phenylalanine	Phe	F
Arginine	Arg	R	Proline	Pro	P
Asparagine	Asn	N	Serine	Ser	S
Aspartic Acid	Asp	D	Threonine	Thr	T
Cysteine	Cys	C	Tryptophan	Trp	W
Glutamic Acid	Glu	E	Tyrosine	Tyr	Y
Glutamine	Gln	Q	Valine	Val	V
Glycine	Gly	G			
Histidine	His	H			
Isoleucine	Ile	I			
Leucine	Leu	L			
Lysine	Lys	K			

Figure 0.1 amino acids letters information according to codons

		Second letter				
		U	C	A	G	
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } Ser UCC } UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	U C A G
	C	CUU } CUC } Leu CUA } CUG }	CCU } Pro CCC } CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } Arg CGC } CGA } CGG }	U C A G
	A	AUU } AUC } Ile AUA } Met AUG }	ACU } Thr ACC } ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G
	G	GUU } Val GUC } GUA } GUG }	GCU } Ala GCC } GCA } GCG }	GAU } Asp GAC } Glu GAA } GAG }	GGU } Gly GGC } GGA } GGG }	U C A G

Figure 0.2 codon (set of three nucleotide) codes for specific amino acid.

Codons select the amino acids and ribosomes make the protein by polymerization process and these nucleotides coil together to form 3D structure.

Module 002: CODING OF AMINO ACIDS

Nucleotides (A, G, C, and T) make set of three called codons for amino acid selection in protein synthesis. More than one codon can code for same amino acids as there are 20 amino acids involved in protein synthesis.

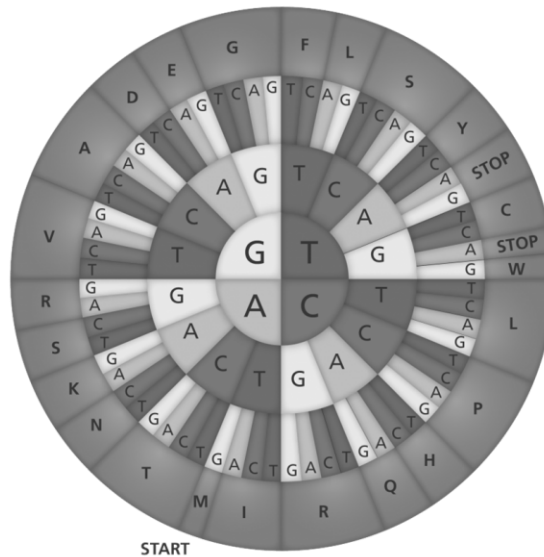


Figure 3 coding of amino acids

```
>gi|14456711|Homo sapiens hemoglobin, alpha 1 (HBA1) mRNA
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCACCATGGTGCTGTCTCCTGCCGACAAGAC
CAACGTC AAGCCGCCTGGGGTAAGGTCGCGCGCACGCTGGCGAGTATGGTGCAGAGGCCCTGGAG
AGGATGTTCTCTGTCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTG
CCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGA
CATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAAGCTTCGGGTGGACCCGGTCAAC
TTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCACCTCCCGCCGAGTTCACCCCTG
CGGTGCACGCCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCG
TAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCCT
↓
>gi|4504347|ref|NP_000549.1| alpha 1 globin [Homo sapiens]
MVLSPADKTNVKAANGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVAD
ALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLAS
VSTVLTSKYR
```

Figure 4 Start Codon ATG and Stop Codon TAG, TGA or TAA

Module 003: OPEN READING FRAMES

Codons codes information for amino acid and there are three stop codons and one start codon. For the valid open reading frame it must have longest sequence.

In molecular genetics, an **open reading frame (ORF)** is the part of a **reading frame** that has the potential to code for a protein or peptide. An **ORF** is a continuous stretch of codons that do not contain a stop codon (usually UAA, UAG or UGA).

https://en.wikipedia.org/wiki/Open_reading_frame

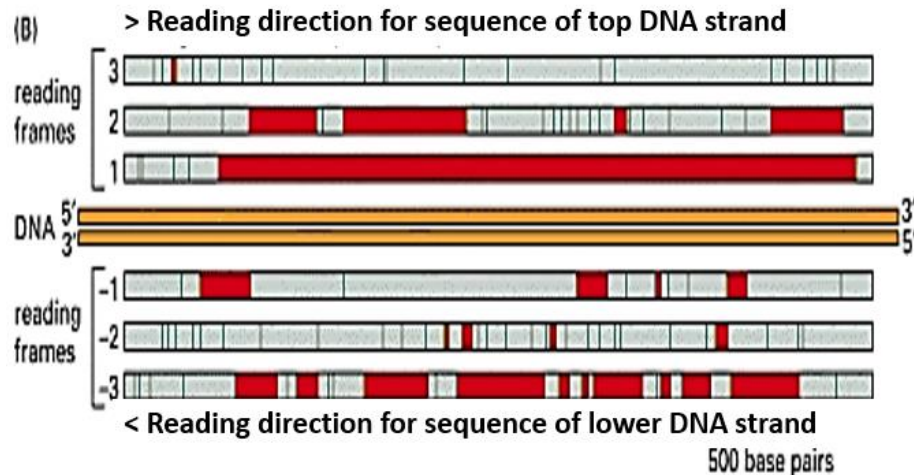


Figure 5 ORF 1 is valid, as it is the longest

There is online tool from which we can find ORF in any sequence.

NCBI

ORF Finder (Open Reading Frame Finder)

PubMed Entrez BLAST OMIM Taxonomy Structure

NCBI

Tools for data mining

GenBank sequence submission support and software

FTP site download data and software

The ORF Finder (Open Reading Frame Finder) is a graphical analysis tool which finds all open reading frames of a selectable minimum size in a user's sequence or in a sequence already in the database. This tool identifies all open reading frames using the standard or alternative genetic codes. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the WWW BLAST server. The ORF Finder should be helpful in preparing complete and accurate sequence submissions. It is also packaged with the Sequin sequence submission software.

Enter GI or ACCESSION OrFind Clear

or sequence in FASTA format

FROM: TO:

Figure 6 NCBI, ORF Finder

Six ORF exist in any DNA sequence and longest one is marked and first stop codon will be marked end of the protein.

Module 004: ORF Extraction – A Flowchart

Codons of 3 nucleotides code for each Amino Acid. There are 1 start and 3 stop codons. Selection of ORF is based on its length if it the longest one from others than it would be suitable for protein synthesis reaction.

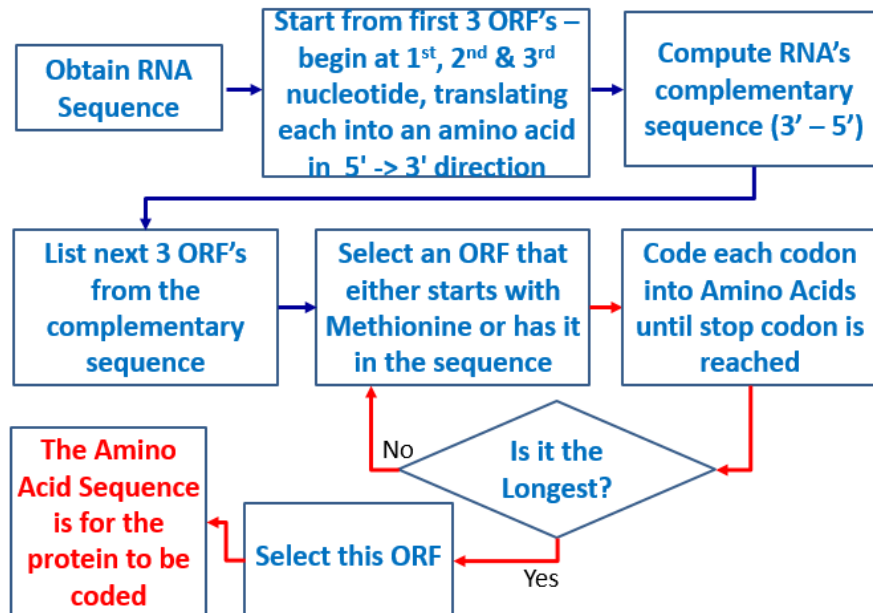


Figure 7 ORF extraction flowchart

Both reverse and forward RNA sequences are considered which may have many ORF and selection is based upon longest protein sequences having.

Module 005: SEQUENCING PROTEINS

Given the DNA/RNA sequence, ORFs can be extracted and protein sequence can be determined. But there are chances that protein may be unknown, that's why we use Edman degradation method in protein sequencing.

Edman degradation, developed by Pehr Edman, is a method of sequencing amino acids in a peptide.

In this method, the amino-terminal residue is labeled and cleaved from the peptide without disrupting the peptide bonds between other amino acid residues. It starts from the N-terminal and removes one amino acid at a time.

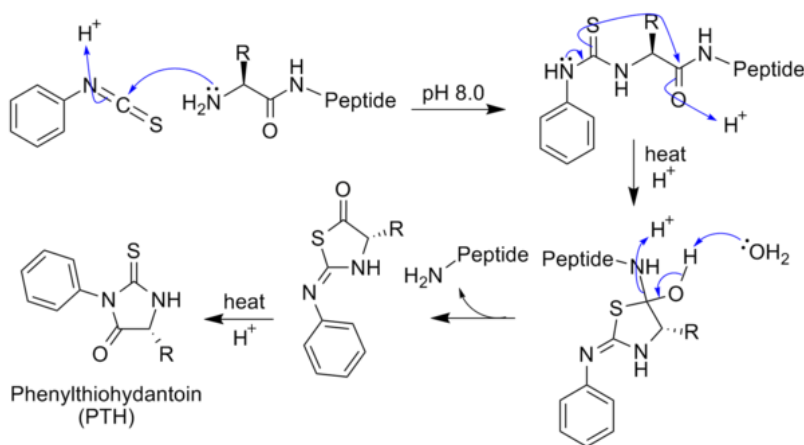
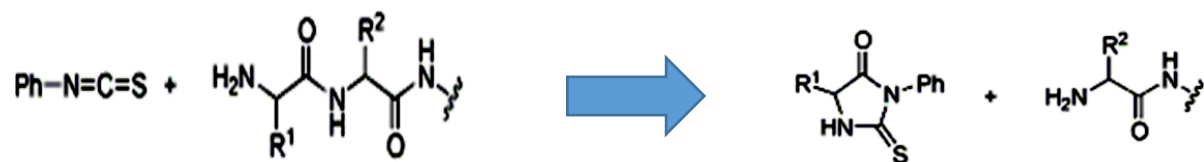


Figure 8 Mechanism

https://en.wikipedia.org/wiki/Edman_degradation

Cyclic degradation of peptides by Phenyl-iso-thio-cyanate (PhNCS). PhNCS attaches to the free amino group at N-terminal residue. 1 amino acid is removed as a PhNCS derivative.



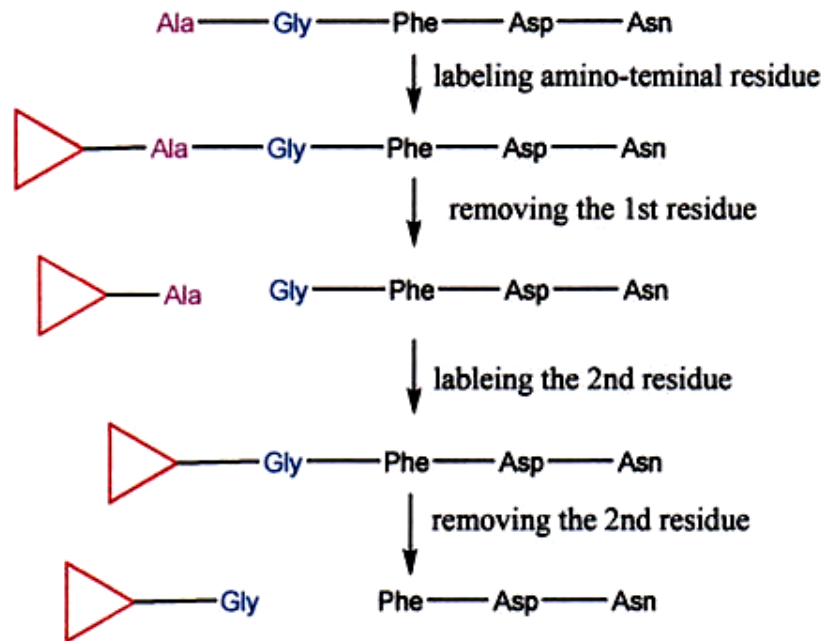


Figure 9 working of Edman degradation

- **DRAWBACKS**

- It is restricted to chain of 60 residues.
- It is very time consuming process 40-50 amino acids per day.

Modern techniques for this is Tandem mass spectrometry.

Module 006: Application of Mass Spectrometry in Protein Sequencing

Edman Degradation methods helps us to sequence the protein which is unknown. But it is restricted to 60 amino acids only.

Protein can be charged with electrons or protons and if moving charges are placed in between the magnetic field they get deflected. And their deflection is proportional to their momentum.

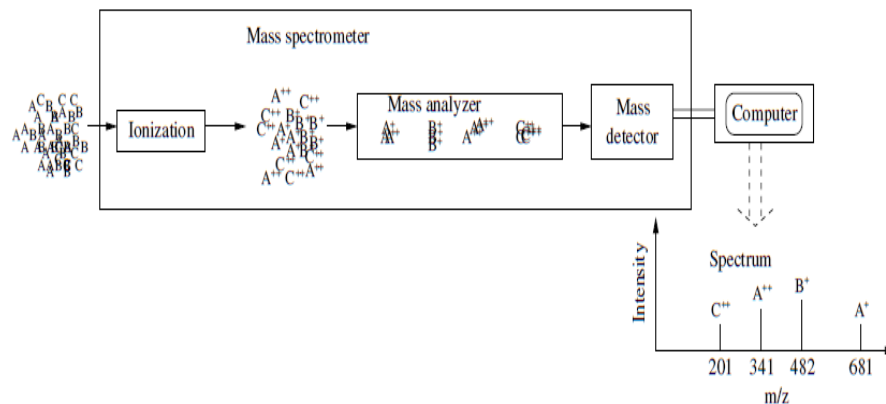
$$\mathbf{F} = Q(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \text{ (Lorentz force law)}$$

$$\mathbf{F} = m\mathbf{a} = m \frac{d\mathbf{v}}{dt} \quad \text{(Newton's second law of motion)}$$

Figure 10 Moving charged particles in a magnetic field

Where:

- **F** is the force applied to the ion
- **m** is the mass of the particle,
- **a** is the acceleration
- **Q** is the electric charge,
- **E** is the electric field
- **$\mathbf{v} \times \mathbf{B}$** is the cross product of the ion's velocity and the magnetic flux density.



$$\left(\frac{m}{Q}\right) \mathbf{a} = \mathbf{E} + \mathbf{v} \times \mathbf{B}.$$

Figure 11 equation for MS application in protein sequencing

- **COMPONENTS**

- Sample Injection
- Ionization Source
- Mass Analyzer
- Ion Detector
- Spectra search using computational tools

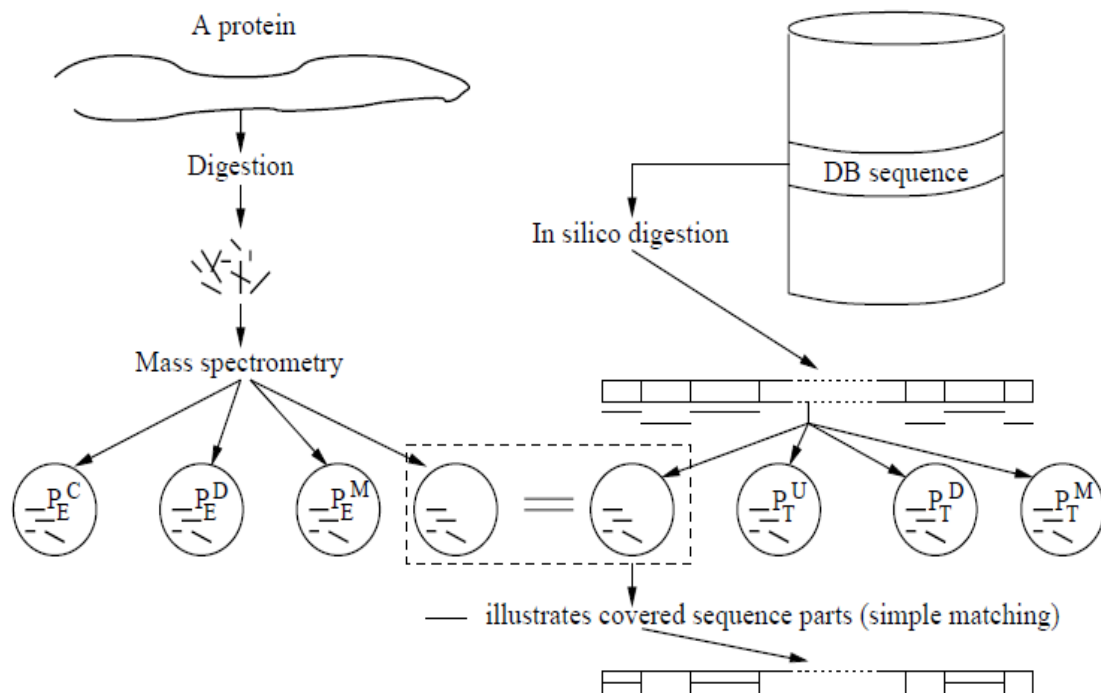
- **CONCLUSION**

Charged proteins can be set into motion within a magnetic field. Their deflections accurately correspond to their molecular mass. Deflections can be measured (hence protein's mass)

Module 007: Techniques for MS Proteomics

MS proteomics works on the principle of protein ionization which are placed in very high magnetic field. Each protein deflect to its proportion which is equal to its molecular weight in this way molecular mass is measured. The protein mass of unknown protein is compared with the masses of proteins in database and matching one is selected.

Example for protein sequences database is uniProt, swissprot etc.



Proteins are measured and sequenced if are unknown than matched with existing database if matched than are shortlisted.

Module 008: Types of MS-based proteomics

Proteins can be sequenced by Edman's degradation and Mass spectrometry. MS based proteomics helps us to sequence the larger and bigger proteins more quickly.

Following steps are involved in MS:

- Separation
- Ionization
- Mass analysis
- Detection

Two methodologies are involved

1. Bottom up proteomics
2. Top down proteomics

Bottom up proteomics measures the peptide masses produced after protein enzymatic digestion. And Top down proteomics measures the intact proteins followed by peptides after fragmentation.

- **BOTTOM UP PROTEOMICS**

In this methodology the protein complex is treated with site specific enzymes which cleaves them into amino acid residue and resultant peptides are measured for their masses. One peptide is selected at one time for processing and when all are processed then protein search engine is used for matches.

- **TOP DOWN PROTEOMICS**

In this methodology proteins are ionized and measured for their masses and one protein is mass selected at a time for fragmentation. And resultant peptide fragments are measured for mass.

We can say that bottom up proteomics deals with peptides while top down proteomics can handle the whole protein.

Module 009: BOTTOM UP PROTEOMICS

There are two types of proteomics protocol that are usually employed.

1. Bottom up proteomics
2. Top down proteomics

PROTOCOL

1. Sample containing the mixture of protein from cells and tissues is obtained.
2. Enzymes such as trypsin is use to cleave the proteins.
3. Enzyme cleaves the amino acids at specific sites of amino acid.
4. Several peptides are formed when protein is cleaved.
5. Number of peptide depends upon the number of sites where enzymes cleaved the protein. For example trypsin cleaves the protein at lysine (k)
6. Mass of each peptide is measured.
7. One peptide is selected at a time.
8. Different enzyme is use to cleave the protein at different site.
9. This process keep going until the possible number of peptides are formed or searched.
10. Peptides are searched in data base and matched.

Module 010: Two Approaches for Bottom up Proteomics (BUP)

There are two approaches for bottom up proteomics.

1. Peptide Mass Fingerprinting.
2. Shotgun Proteomics

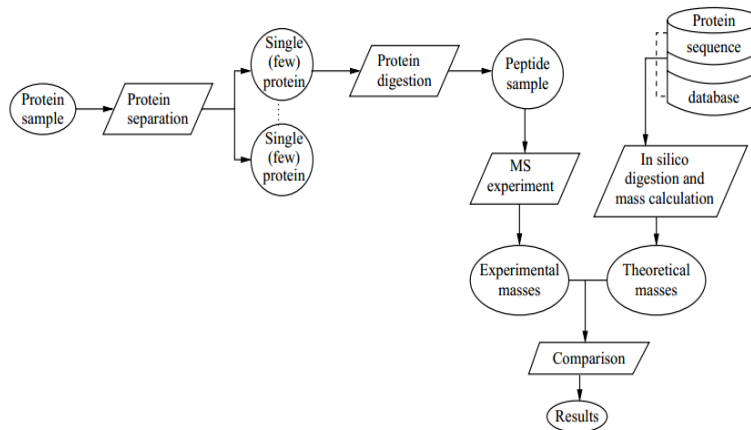


Figure 0.33 Peptide mass fingerprinting

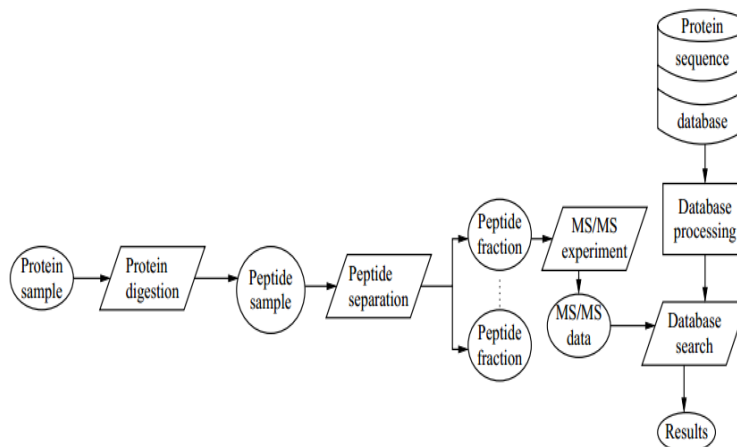


Figure 14 Shotgun Proteomics

Shotgun Proteomics digest the whole protein and mix first and compared with database. And peptide mass finger printing involves in protein separation followed by single protein's peptide analysis.

Module 011: TOP DOWN PROTEOMICS

Bottom up proteomics identifies the proteins by cleaving them into segments at specific sites and was not suitable to measure the direct protein mass.

- **PROTOCOL**

1. Sample containing the protein mixture from cells and tissues is obtained
2. The entire protein is mixed and analyzed for masses.
3. The list of masses is obtained.
4. TDP Measures all post translational masses of protein.
5. After MS1 one protein is selected at a time and fragmented to obtain its peptides.
6. The process is repeated many times.

Comparison is done from protein database uniProt and swissProt.

TDP also measure the masses of intact proteins and masses of post transcriptional changes.

Module 012: PROTEIN SEQUENCE IDENTIFICATION

Mass spectrometry helps us to measure the molecular weight of proteins and peptides, but several proteins can have same masses to identify them we follow the flow chart of following techniques.

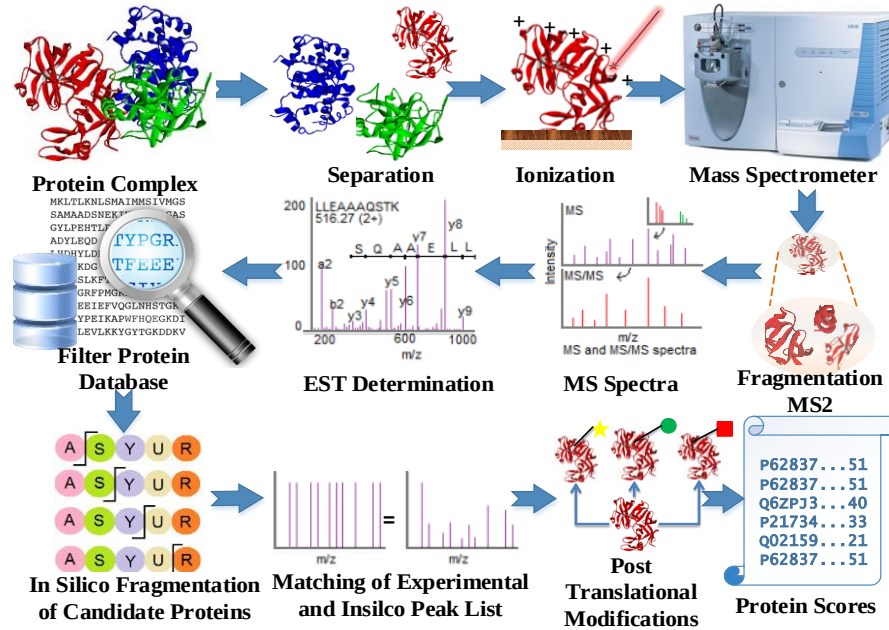
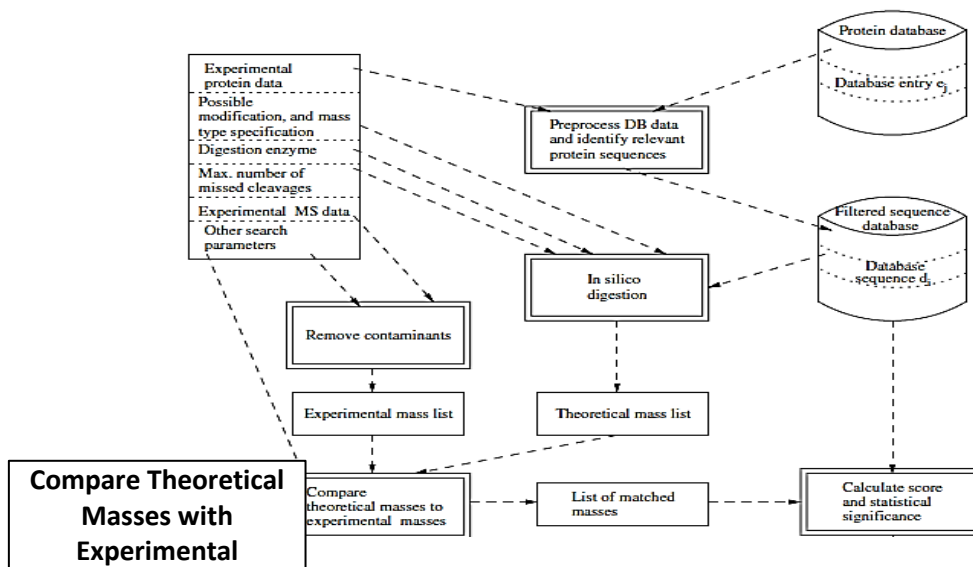


Figure 0.45 protein sequence identification flowchart



The flowcharts discussed above can help us arrive at the sequence of the protein in question. Scoring schemes are required to quantitatively represent the quality of results

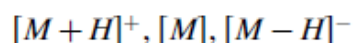
Module 013: PROTEIN IONIZATION TECHNIQUES

Protein ionization is used in Mass spectrometry based on proteomics protocols. Ionization involves loading of proton in protein or removal of proton. Ionizations can increase or decrease the mass of protein or peptide.

- **SALIENT IONIZATION**

Is the technique which include Matrix Assisted Laser Desorption Ionization (MALDI) & Electro Spray Ionization (ESI)

For example:



monoisotopic or average masses

- **MALDI**

In this technique one proton is added to protein or peptide and the molecular weight is increases by one and Mass spectrometry reports the molecule at +1.

- **ESI**

ESI adds many protons to protein or peptides and molecular weight is increased by the number of protons added. But it is difficult in ESI to find the molecule with +1.

- **EXAMPLE**

Example Suppose we have a peptide with mass 2000.0 Da, and that the ionization yields peptide ions of charge +1, +2, and +3, by the attraction of one, two, or three protons, respectively. The peptide ions will then be detected at

$$\text{ions with charge } +1: m/z = (2000 + 1)/1 = 2001$$

$$\text{ions with charge } +2: m/z = (2000 + 2)/2 = 1001$$

$$\text{ions with charge } +3: m/z = (2000 + 3)/3 = 666.7$$

Figure 0.56 resolving multiple charges

MS data from MALDI ionization is easier to handle as the product ions masses are mostly at “1+mass”. ESI is difficult to use as it does not easily give away the +1 charged ion

Module 014: MS1 & INTACT PROTEIN MASS

When we ionize the protein, it can be deflected by a magnetic field in proportion to its mass and the mass of protein can be measured by spectrometry.

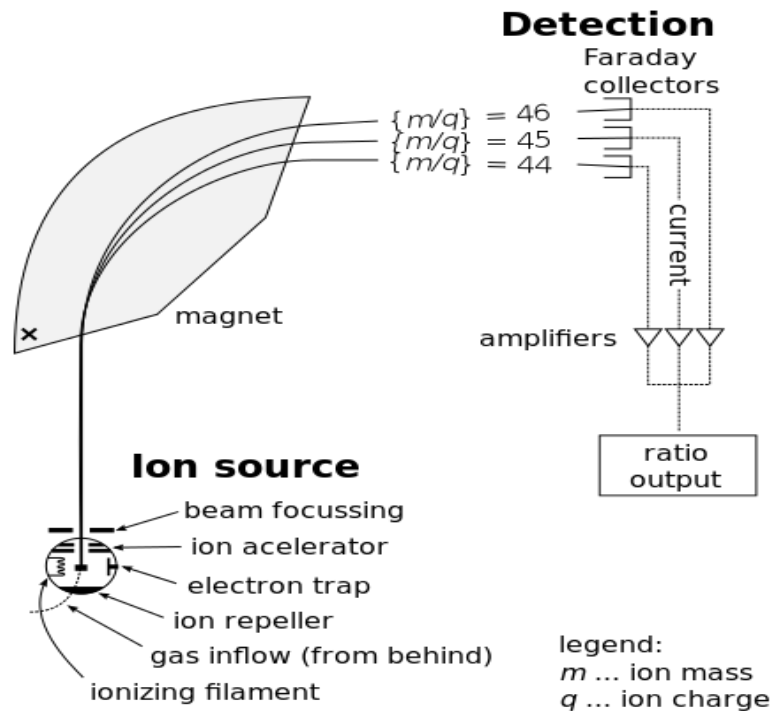


Figure 0.67 MS1 Schematic (Image courtesy Wikipedia)

Mass/charge helps us to calculate the mass of protein, "Mass Select" can help to select specific MS1 for further analysis.

MS1 results the intact masses of the peptides.

Module 015: SCORING INTACT PROTEIN MASS

MS1 helps us to obtain the intact masses of precursor molecules which depend upon the proteomics and protocol applied. Protein masses reported by MS1 are matched with protein database, but before match the masses are converted into +1 of all molecule.

- **SCORING**

We can score each protein in the way that it get maximum score and low quality matches should get low scores.

After filtering the multiple charges we get the only the peaks having charge 1. And after this filter we compare it with protein data base.

- **SCORING SCHEME**

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

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.

Module 016: 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.

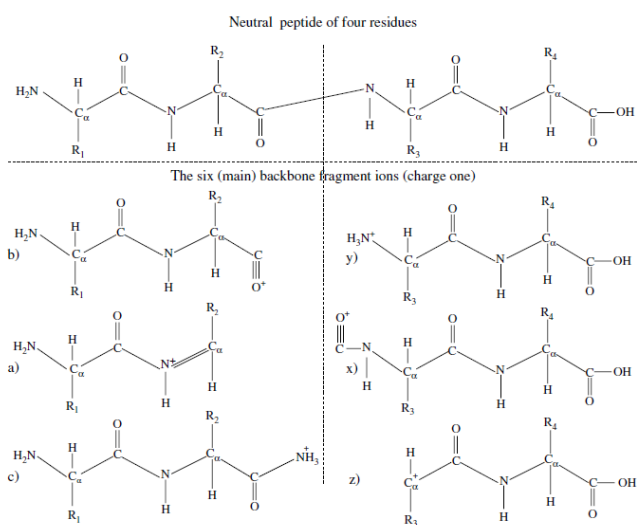
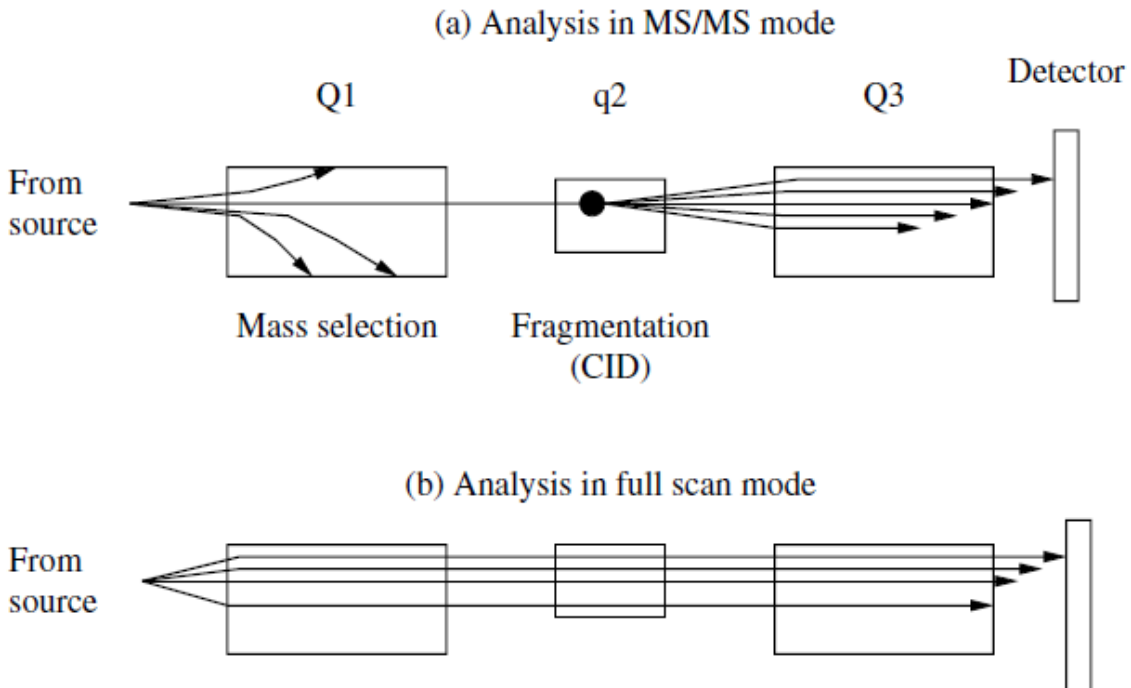


Figure 0.78 natural peptide of four residue

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.

Module 017: 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.

Module 018: MEASURING EXPERIMENTAL FRAGMENT'S MASS

In MS1, the molecular weight of intact sample molecule is measured and then the intact molecule is fragmented into 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 depending upon the technique because each technique splits the protein or peptide at different locations.

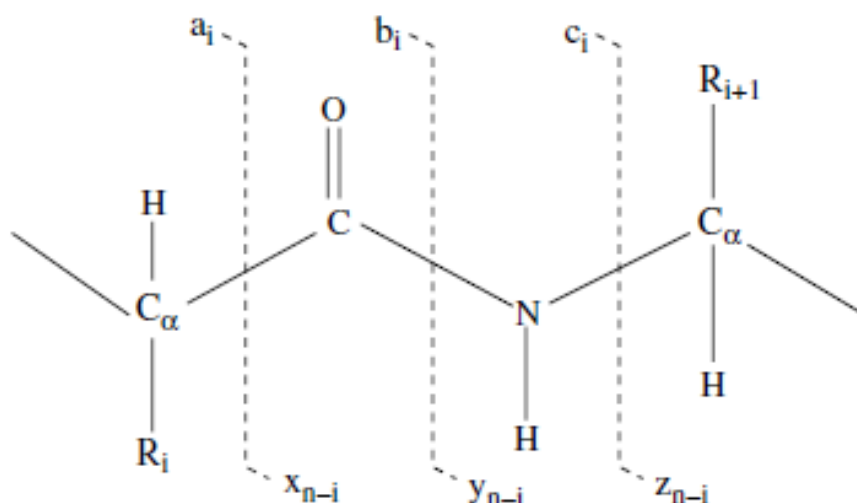


Figure 0.89 Masses after Fragmentation by ECD, CID & ETD

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.

Module 019: Calculating Theoretical Fragment's Mass

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

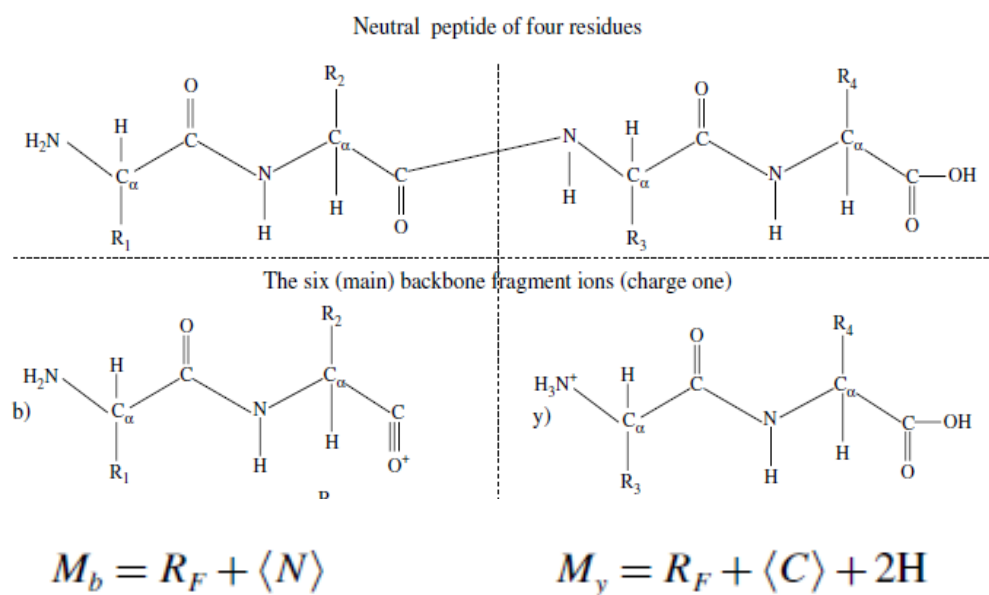


Figure 20 Masses after Fragmentation by CID

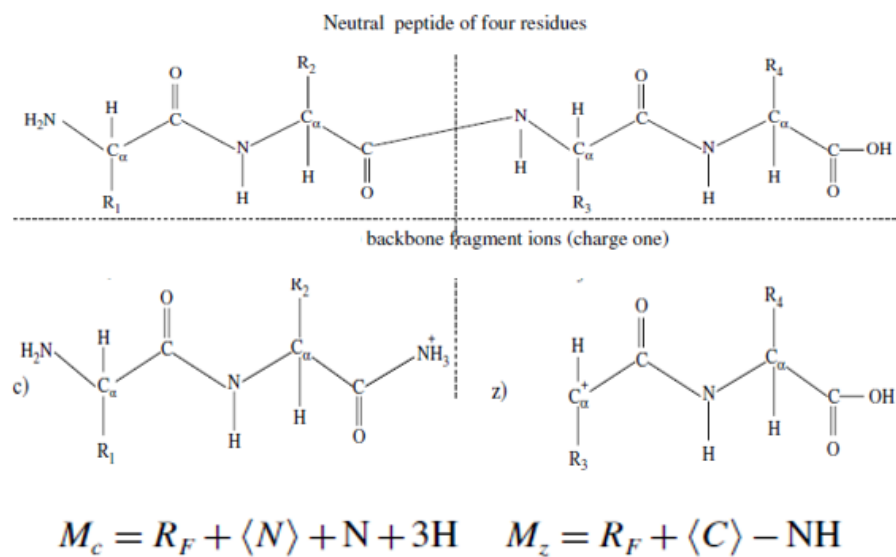


Figure 0.91 Masses after Fragmentation by ECD

$$\begin{array}{ll}
 M_b = R_F + \langle N \rangle & M_y = R_F + \langle C \rangle + 2H \\
 M_a = R_F + \langle N \rangle - CO & M_x = R_F + \langle C \rangle + CO \\
 M_c = R_F + \langle N \rangle + N + 3H & M_z = R_F + \langle C \rangle - NH
 \end{array}$$

$$M_b + M_y = R_P + \langle N \rangle + \langle C \rangle + 2H (= M_P + 2H)$$

$$M_a + M_x = R_P + \langle N \rangle - CO + \langle C \rangle + CO (= M_P)$$

$$M_c + M_z = R_P + \langle N \rangle + N + 3H + \langle C \rangle - N - H (= M_P + 2H)$$

Figure 22 Masses after Fragmentation by ECD

Module 020: PEPTIDE SEQUENCE TAG

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.

Mass Select: [375.0 , 377.0]

Protein Sequence: "MQV..."

Fragments formed may be **M**, **MQ**, **QV** etc.

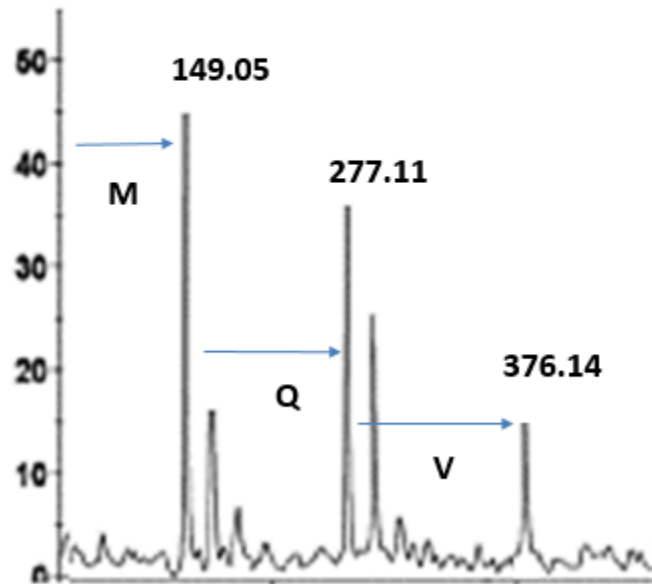


Figure 23 variation sites

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

Module 021: Extracting Peptide Sequence Tags

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

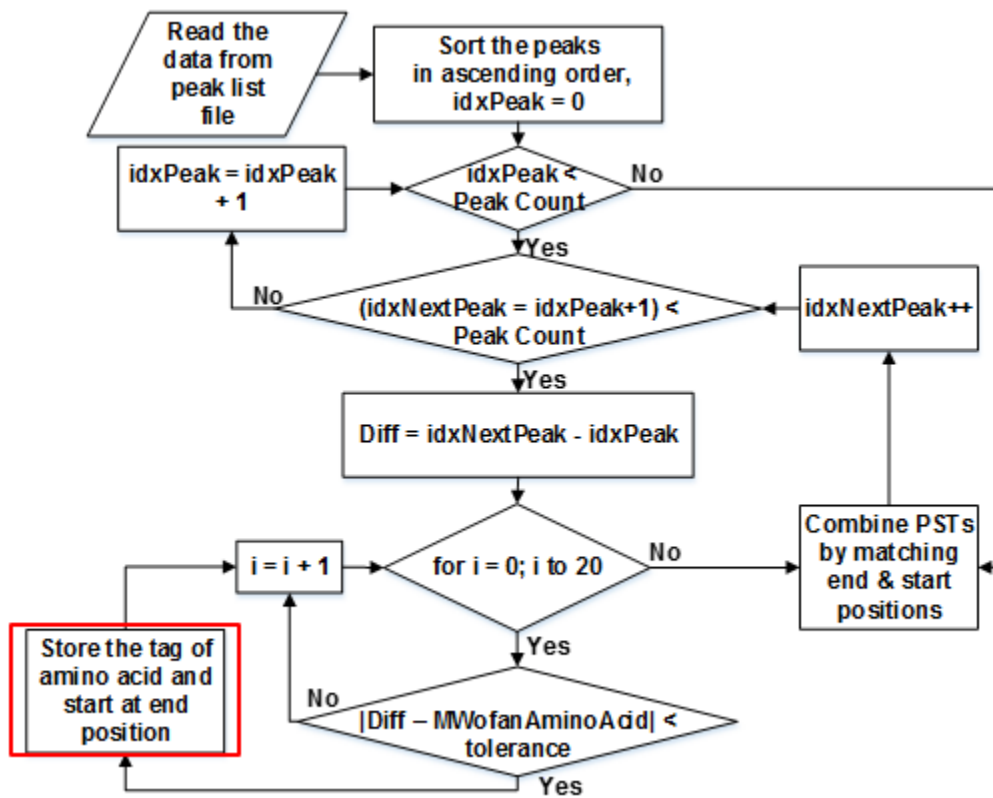


Figure 24 peptide sequencing tagging

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!

Module 022: 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|OX1R_FRG3G Putative transcription factor OX1R OS=Random virus 3 (isolate
Goorha) GN=FV3-OX1R PE=4 SV=1
```

```
MAFSAEDVLKEYDRRRRMEALLSLYYPNDRKLLDYKEWSPPRVQVECPKAPVEWNNPPSEKGLIVGHFSGI
KYKGEKAQASEVDVNKMCCWVSKFKDAMRRYQGIQTCKIPGQVLSDLDAKIKAYNLTVEGVEGFVRYSRVT
KQHVA AFLKELRHSKQYENVNLIHYILTDKRVDIQHLEKDLVKDFKALVESAHMRMQGHMQNVKYILYQLLK
KHGHGPDGPDILTVKTGSMQLYDDSFRIYTDLGWKFTPL
```

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.

Module 023: 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.

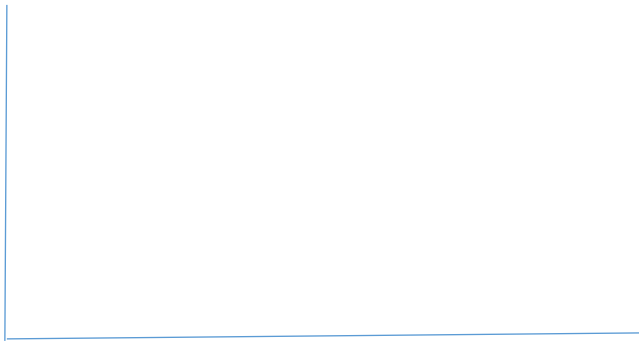


Figure 25 root mean square error

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

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

So, the updated relationship is:

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

Module 024: In silico Fragment Comparison

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.

In case 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 confirmed then 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.

Module 025: 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

Module 026: 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.

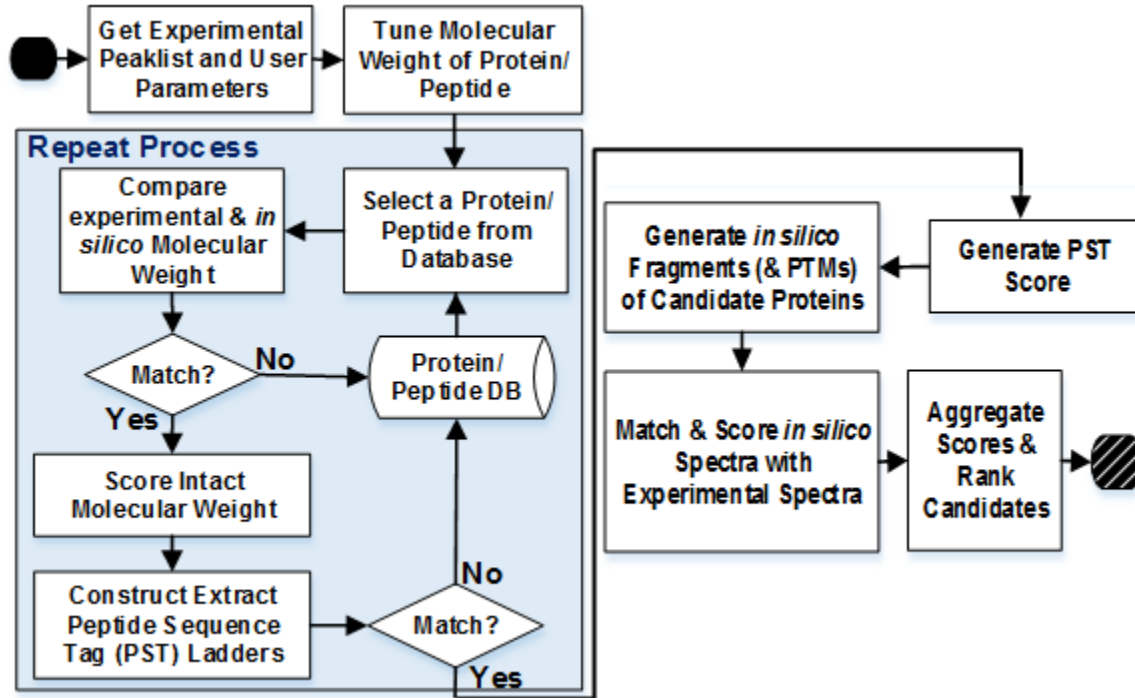


Figure 27 protein sequencing flowchart

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.

Module 027: 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).

$$Score_{MW} + Score_{PST} + Score_{Exp \leftrightarrow Thr} = Score$$

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

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

- 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.

Module 028: 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**

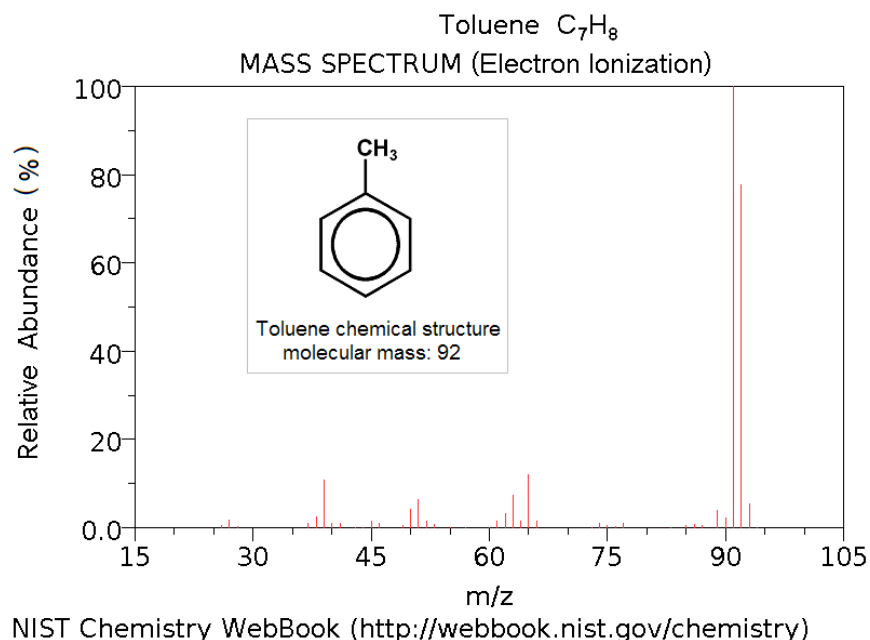
1. The instrument alternates between MS and MS/MS modes.
2. Three most intense peaks are chosen for MS/MS analysis.
3. After the initial MS scan, an MS/MS spectrum from peptide A is obtained by selectively fragmenting this mass only.
4. Next, a spectrum for peptide B is produced, followed by a recording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

Module 029: 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:

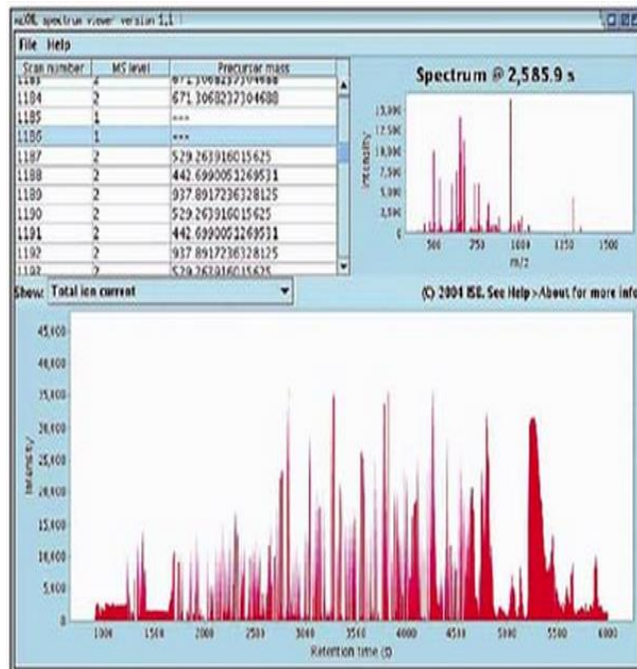
Agilent	.D/.YEP
Bruker	.BAF
ABI/Sciex	.WIFF
	.t2d
Thermo Xcalibur	.RAW
Micromass	
PerkinElmer	
Waters	
Shimadzu	
	.QGD

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)



The screenshot shows the Mascot Daemon software interface. It has a menu bar with 'File', 'Edit', and 'Help'. Below the menu bar are tabs for 'Status', 'Event Log', 'Task Editor', and 'Parameter Editor'. The 'Parameter Editor' tab is active, showing a 'Parameter set' section with fields for 'Filename', 'New', 'Open...', 'Save', and 'Save As...'. Below this is the 'All Searches' section with various search parameters: 'User name' (Mascot Daemon), 'User email', 'Search title' (Submitted from <taskname> by Mascot Daemon on <localhost>), 'Taxonomy' (All entries), 'Database' (MSDB), 'Enzyme' (Trypsin), 'Max. missed cleavages' (1), 'Fixed modifications' (Negative [C-term], Negative [N-term], Neg-2 [N-term], Bombesin [C-term]), 'Variable modifications' (N-Formyl (Protein), NIPCAM [C], O18 [C-term], Oxidation [M]), 'Protein mass' (kDa), 'Peptide tol.' (1.5 Da), 'Peptide charge' (2+ and 3+), 'MS/MS' section with 'MS/MS ions search' (checked), 'Data format' (Mascot generic), 'Protein summary' (unchecked), 'Peptide summary' (checked), 'Precursor m/z', 'MS/MS tol.' (0.8 Da), 'ICAT' (unchecked), and 'Instrument' (ESI-TRAP).

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

Module 030: 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.

Vendor	Instrument	Data Type	Converter & Set Up
Thermo	Ion Trap / Orbitrap	raw	MSFileReader
AB SCIEX	Triple TOF	wiff	AB SCIEX Data Converter
	QSTAR / QTRAP	wiff	Analyst + mzWiff
	TOF TOF	wiff	AB Extractor
Agilent	Ion Trap	yep	CompassXtract
	Q-TOF	d	MassHunter
Bruker	QTOF / Ion Trap / TOF TOF	baf / yep / lift	CompassXtract
Shimadzu	AXIMA-CFR	run	Kratos Converter
Varian	FTICR / Ion Trap	sms/xms	VarianMS
Waters	Q-TOF	raw	MassLynx

Figure 29 Raw file formats

Free Software Name	Input RAW Data Type	Output Data Types
ProteoWizard	Thermo, Agilent, Bruker, Waters, AB Sciex	mzML, TraML, mzIdentML, mzXML
OpenMS	All vendor formats	mzML, TraML, mzIdentML, mzData, mzQuantML
Trans Proteomic Pipeline	Thermo, ABI Sciex, Agilent, Waters. ^[14]	mzML, mzXML, pepXML, protXML (Proteowizard)
PEAKS	Thermo, ABI Sciex, Agilent, Waters, Bruker, Shimadzu, Varian	MSF, tandem, <u>mzML</u> , <u>omx</u> , <u>dat</u> , FASTA

Figure 30 list of tools for raw data processing

Multiple RAW file formats are prevailing in the industry. Each vendor has its own unique RAW file format. You can convert proprietary formats into open formats

Module 031: MGF FILE FORMAT

MGF – Mascot Generic Format. MGF is a simple human-readable format for MS/MS data developed by Matrix Science. Mascot Search Engine available at this link online.

http://www.matrixscience.com/search_form_select.html

Name	Description	Header	Local	Choices/Range	Notes
ACCESSION	Database entries to be searched	✓		List of double quoted, comma separated values	
CHARGE	Peptide charge	✓	✓	1-	M-H- on PMF form
				Mr	
				1+	MH+ on PMF form
				N- to N+ where N is an integer and combinations	Not PMF
CLE	Enzyme	✓		Trypsin etc., as defined in enzymes file	No default, so must be specified
COM	Search title	✓			Applies to the whole search

http://www.matrixscience.com/help/data_file_help.html

CUTOUT	Precursor removal	✓		Pair of comma separated integers	MIS only
COMP	Amino acid composition		✓		
DB	Database	✓		As defined in mascot.dat	
DECOY	Perform decoy search	✓		0 (false)	Default
				1 (true)	
ERRORTOLERANT	Error tolerant	✓		0 (false)	Default
				1 (true)	Not PMF
ETAG	Error tolerant sequence tag		✓		A single query can have multiple ETAGs

<u>FORMAT</u>	MS/MS data file	✓		Mascot generic	Default
				Sequest (.DTA)	
				Finnigan (.ASC)	
				Micromass (.PKL)	
				PerSeptive (.PKS)	
				Sciex API III	
				Bruker (.XML)	
				mzData (.XML)	
				mzML (.mzML)	
<u>FRAMES</u>	NA translation	✓		Comma separated list of frames	Default is 1,2,3,4,5,6
<u>INSTRUMENT</u>	MS/MS ion series	✓	✓	Default	Default
				ESI-QUAD-TOF etc., as defined in fragmentation_rules	
<u>IT_MODS</u>	Variable Mods	✓	✓	As defined in unimod.xml	
<u>ITOL</u>	Fragment ion tol.	✓		Unit dependent	

<u>ITOLU</u>	Units for ITOL	✓		ppm	
				Da	
				mmu	
<u>LOCUS</u>	Hierarchical scan range identifier		✓	string	MIS only
<u>MASS</u>	Mono. or average	✓		Monoisotopic	
				Average	
<u>MODS</u>	Fixed Mods	✓		As defined in unimod.xml	
<u>MULTI_SITE_MODS</u>	Allow two modifications at a single site	✓		0 (false) or 1 (true)	default 0
<u>PEP_ISOTOPE_ERROR</u>	Misassigned ¹³ C	✓		0 to 2	MIS only
<u>PEPMASS</u>	Peptide mass		✓	>100	optionally followed by intensity and charge

<u>PRECURSOR</u>	Precursor m/z	✓		>100	
<u>QUANTITATION</u>	Quantitation method	✓		as defined in quantitation.xml	MIS only
<u>RAWFILE</u>	Raw file identifier		✓	string	MIS only

Module 032: OPEN MS DATA FORMAT

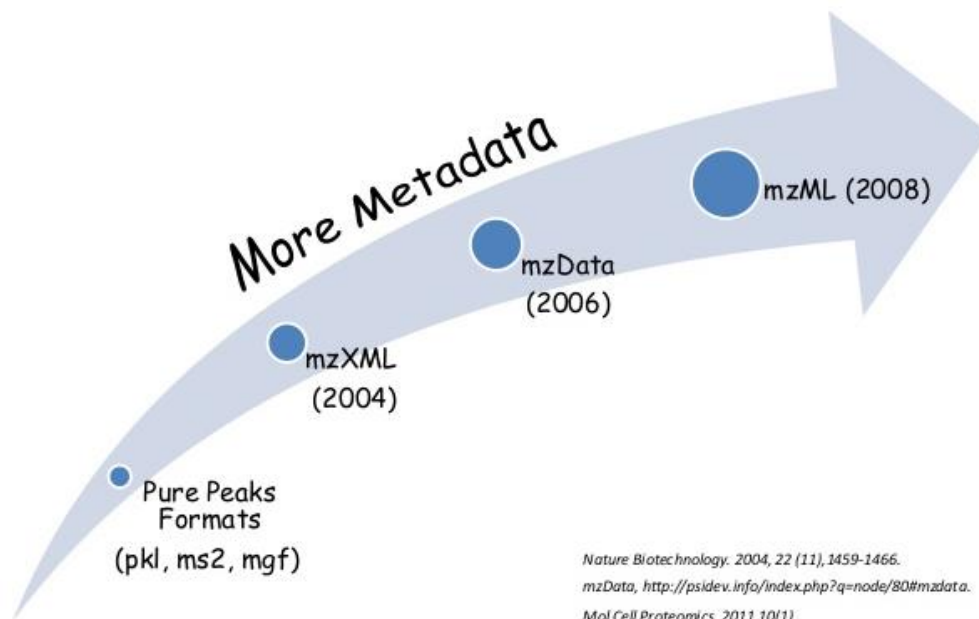
Mass spectrometer outputs data with mass/charge ratios & respective ion intensities. RAW file formats are specific to each instrument and each vendor has its own unique file format. Once an instrument is upgraded, data output from the instrument is also changed. Hence the underlying RAW file format needs to be upgraded as well.

NEED

Proprietary RAW formats are binary formats which are difficult to read and parse. If you have the software from the maker of the MS then you can read the RAW data file as well.

SOLUTION

- mzData was developed by HUPO-PSI
- mzXML was developed at the Institute for Systems Biology
- To combine them, a joint venture produced mzML



Tool	Formats
ProteoWizard	mzML, TraML, mzIdentML, mzXML, vendor formats
OpenMS	mzML, TraML, mzIdentML, mzData, mzQuantML, et al.
Trans-Proteomic Pipeline (TPP)	mzML, mzXML, pepXML, protXML (ProteoWizard)
compomics-utilities	MSF, tandem, mzML, omx, dat, FASTA
jmzReader	mzML, mzXML, mzData, PRIDE XML, dta, MGF, ms2, pkl
jTraML	TraML
multiplierz	Vendor formats
PEFF Viewer	PEFF
PRIDE Converter 2	mzTab, PRIDE XML (jmzReader)
Mascot & Distiller	MGF, mzML, mzXML, mzIdentML, vendor formats
SpectraST	msp, splib, blib, ASF, mzML, mzXML, pepXML, etc.
ProHits	PSI-MI (TPP formats)

Figure 31 Formats used for open use

Anubis	TraML, mzML, mzXML
Proteios	TraML, mzML, mzXML
Skyline	.sky, .skyd, mzML, mzXML, vendor formats
ATAQS	TraML, mzML, mzXML
Corra	APML, mzXML
Java MIAPE API	PRIDE XML, mzML, mzIdentML, GelML
Tool	Formats

Several software exist for converting RAW file formats into open software formats. Each open format has its own unique advantages. mzXML and MGF formats are most frequently used

Module 033: Online Proteomics Tools - MASCOT

Matrix Science developed an online Bottom up Proteomics Search Engine. "MASCOT". Mascot can search peptide mass fingerprinting and shotgun proteomics dataset

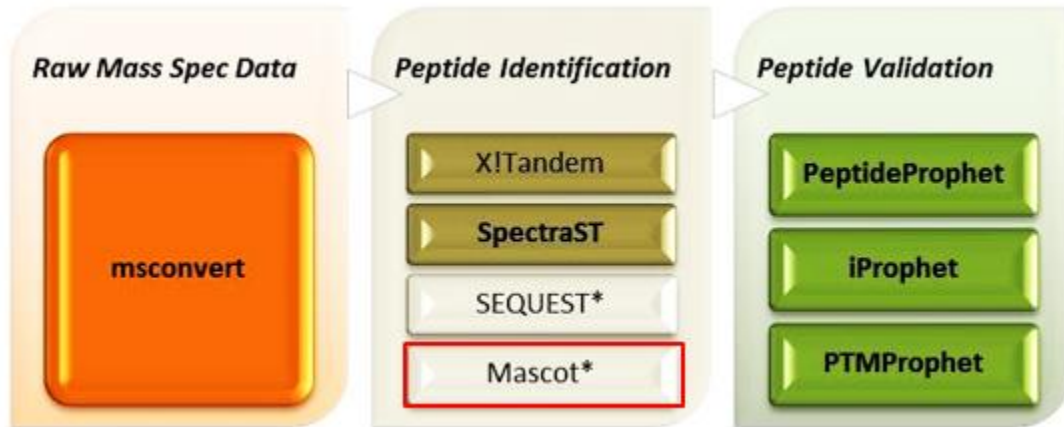


Figure 32 mzML <http://tools.proteomecenter.org/software.php>

MATRIX SCIENCE

Search this site

Home Mascot database search **Products** Technical support Training News Blog Newsletter Contact

[Access Mascot Server](#) | [Database search help](#)

Mascot database search > Access Mascot Server

Access Mascot Server

You are welcome to submit searches to this free Mascot Server. Searches of MS/MS data are limited to 1200 spectra and some functions, such as no enzyme searches, are unavailable. Automated searching of batches of files is not permitted. If you want to automate search submission, perform large searches, search additional sequence databases, or customise the modifications, quantitation methods, etc., you'll need to [license your own](#), in-house copy of Mascot Server.

Peptide Mass Fingerprint

The experimental data are a list of peptide mass values from the digestion of a protein by a specific enzyme such as trypsin.

[Perform search](#) | [Example of results report](#) | [Tutorial](#)

More info

- > [Mascot overview](#)
- > [Search parameter reference](#)
- > [Data file format](#)
- > [Results report overview](#)

Figure 33 http://www.matrixscience.com/search_form_select.html

Your name Email
 Search title
 Database(s) SwissProt
 NCBInr
 contaminants
 cRAP
 Enzyme Trypsin
 Allow up to 1 missed cleavages
 Taxonomy All entries
 Fixed modifications --- none selected ---
 Display all modifications ☐
 Variable modifications --- none selected ---
 Acetyl (K)
 Acetyl (N-term)
 Acetyl (Protein N-term)
 Amidated (C-term)
 Amidated (Protein C-term)
 Ammonia-loss (N-term C)
 Biotin (K)
 Biotin (N-term)
 Carbamidomethyl (C)
 Carbamyl (K)
 Carbamyl (N-term)
 Protein mass kDa
 Peptide tol. $\pm$ Da
 Mass values ☒ MH⁺ ☐ M_r ☐ M-H⁻
 Monoisotopic ☒ Average ☐
☐ Data file No file chosen
☐ Query
 Data input
 Decoy ☐
 Report top AUTO hits

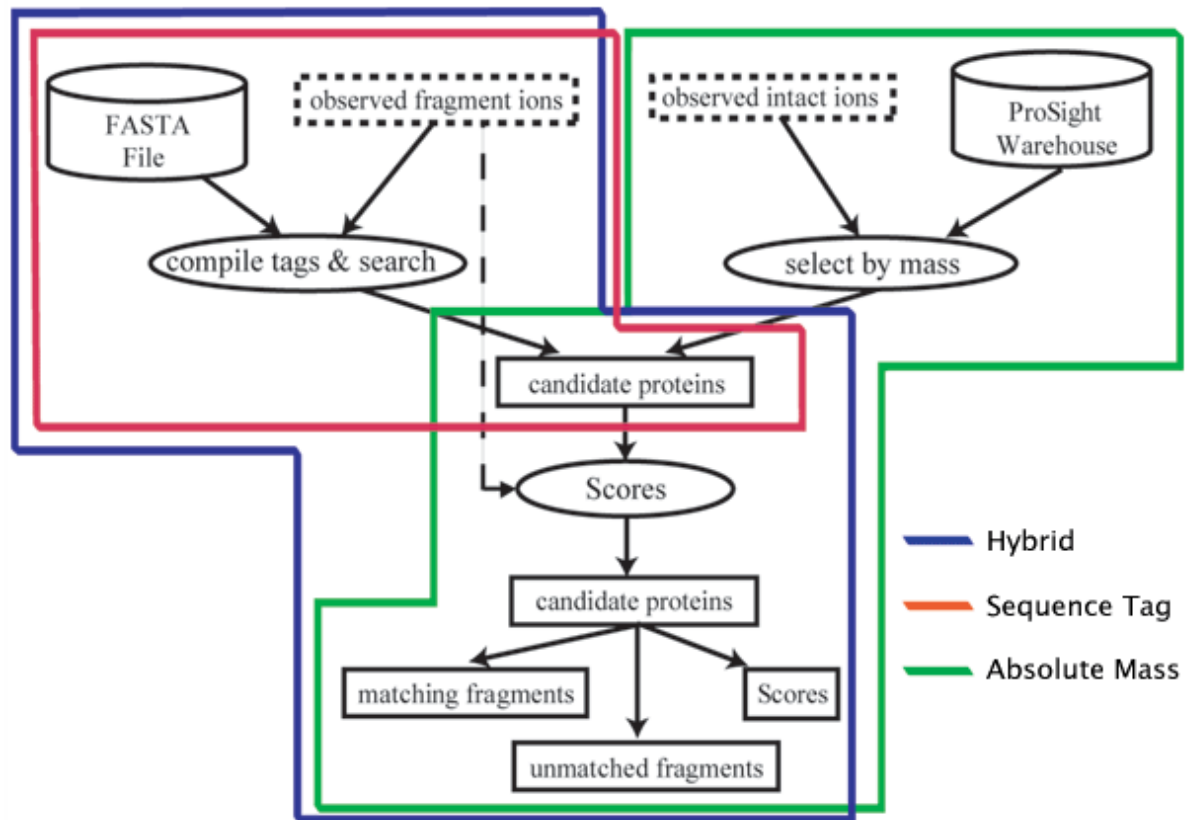
MASCOT MS/MS Ions Search

Your name Email
 Search title
 Database(s) Invertebrates_EST
 Human_EST
 Fungi_EST
 Environmental_EST
 SwissProt
 Enzyme Trypsin
 Allow up to 1 missed cleavages
 Quantitation None
 Taxonomy All entries
 Fixed modifications --- none selected ---
 Display all modifications ☐
 Variable modifications --- none selected ---
 Acetyl (K)
 Acetyl (N-term)
 Acetyl (Protein N-term)
 Amidated (C-term)
 Amidated (Protein C-term)
 Ammonia-loss (N-term C)
 Biotin (K)
 Biotin (N-term)
 Carbamidomethyl (C)
 Carbamyl (K)
 Carbamyl (N-term)
 Peptide tol. $\pm$ Da # ¹³C
 MS/MS tol. $\pm$ Da
 Peptide charge 2+
 Data file No file chosen
 Data format Mascot generic
 Instrument Default
 Decoy ☐
 Precursor m/z
 Error tolerant ☐
 Report top AUTO hits

Mascot is the most widely used online search tool for proteomics data. However, it lacks a batch processing mode. Also, it does not cater for top-down proteomics data.

Module 034: Online Proteomics Tools – ProSight PTM

Kelleher et al have developed an online Top down Proteomics Search Engine. “ProSight PTM”. ProSightPTM searches top down proteomics data and reports the precursor protein

Three different search modes

https://prosightptm.northwestern.edu/about_retriever.html

Retriever

Retriever:
Absolute Mass Search (Manual)

file input or manual input

types of searches

Absolute Mass (Files) **Absolute Mass (Manual)** **Sequence Tag** **Hybrid**

Absolute Mass Search

☒ b/y Ion Search *(threshold)*
☐ c/z Ion Search *(ECD)*

Output File(s)

outfile .massacc.X.csv

Proteins
(Limit 10)

20356.85
20396.25
20598.98

Intact Tolerance:

± 2000 Da *search window*

☐ Use Monoisotopic Masses
☒ Use Average Masses

Fragment Ions

1066.63
1179.72
1280.76
1367.8
1482.36
1501.51

Fragment Tolerance:

± 25 ☐ ppm
☐ Da

☒ Use Monoisotopic Masses
☐ Use Average Masses

☒ **Minimum Fragment Ion Matches:**
 2 *limits output file size*

https://prosightptm.northwestern.edu/about_retriever.html

Database

	<u>Organism</u>	<u>Strain</u>	<u>Annotated By</u>	<u># Protein Forms</u>	<u>Size (MB)</u>
☐	<i>M. pneumoniae</i>	M129	Herrmann	(1758)	0.9
☐	<i>E. coli</i>	K12	SWISS-Prot	(22,621)	8
☐	<i>E. coli</i>	K12	Wisconsin	(22,621)	8
☐	<i>B. subtilis</i>	168	Kunst/Danchin	(10,301)	3.8
☐	<i>S. oneidensis</i>	---	SWISS-Prot	(12,504)	4.5
☐	<i>M. jannaschii</i>	DSM2661	Olsen	(4638)	1.8

Post-Translational Modifications

General:
☒ Check For " Δm mode"

N-terminus:
☒ Met On
☒ Met Off
☐ Acetylated (Met Off)
☐ Signal Peptide (*S. cerevisiae* ONLY)
☐ Lipoylation

Other:
☐ Phosphorylation (*S. cerevisiae* ONLY)
☐ Internal Acetylation
☐ Methylation
☐ Glycosylation
☐ C-Terminus Methylation

Cystine Alkylations

☒ None
☐ Acrylamide (71)
☐ Iodoacetamide (57)

searches for one modification
of unknown type

ProSight PTM is the state of the art in top down proteomics search. Using Prosight PTM, post-translational modifications can be accurately identified.

Module 035: Example Case Study - I

For case study we follow some steps:

Step 1 – Monoisotopic Peak Detection

Natural elements occur in multiple isotopes. Isotopes differ in their masses. The abundance of each isotopic variant is unique.

Element		Abundance (%)	Mass
Hydrogen	^1H	99.99	1.007 83
	^2H	0.01	2.014 10
Carbon	^{12}C	98.91	12.000 0
	^{13}C	1.09	13.003 4
Nitrogen	^{14}N	99.6	14.003 1
	^{15}N	0.4	15.000 1
Oxygen	^{16}O	99.76	15.994 9
	^{17}O	0.04	16.999 1
	^{18}O	0.20	17.999 2
Phosphorus	^{31}P	100	30.973 8
Sulfur	^{32}S	95.02	31.972 1
	^{33}S	0.76	32.971 5
	^{34}S	4.22	33.967 6

Figure 34 Isotopic variants of natural elements

TYPES OF MASSES

- Nominal Mass
- Monoisotopic Mass
- Average Mass

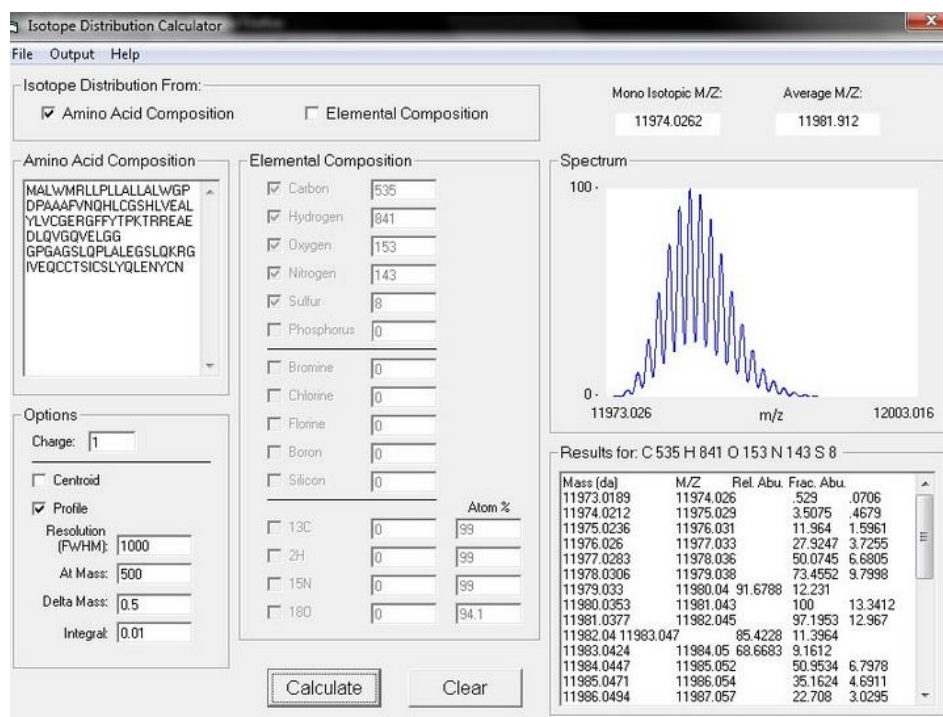


Figure 35 Detecting Monoisotopic Peaks

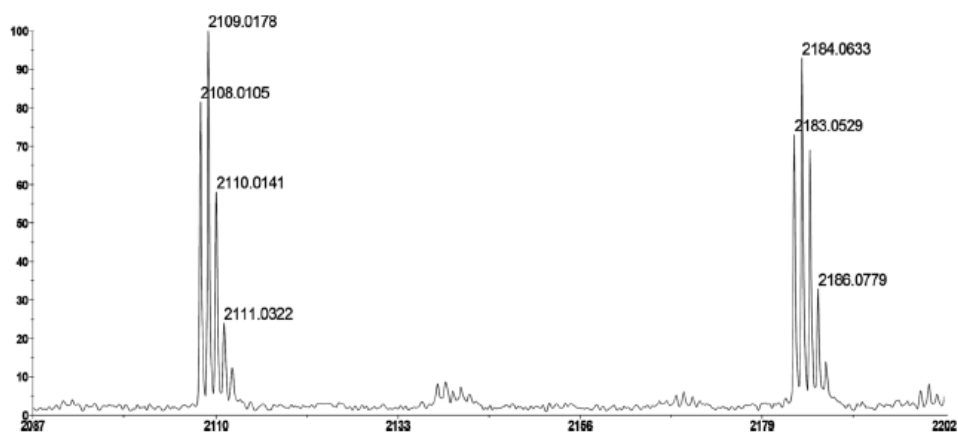


Figure 36 Detecting Monoisotopic Peaks

MS1 data reports the isotopic distribution of intact molecule's mass. Monoisotopic mass value has to be selected from this mass distribution. This value is the highest mass value in the distribution

Module 036: Example Case Study – II

The first step in protein identification and characterization using mass spectrometry involves intact protein/peptide mass measurement. Next, we fragment the protein. A protein or peptide backbone may be fragmented anywhere along the peptide backbone.

This results in formation of two fragments i.e. N-term fragment and C-Term fragment.

For possible fragments let's take an example protein with 100 residues. Such a molecule's backbone can be fragmented at 100 different locations. The total number of possible fragments is then 200

TANDEM MS

The mass of 200 fragments can then be measured by using an MS again. The necessary condition for this measurement is that all 200 fragments are ionized.

To ensure that all fragments of precursor molecule are also charged, we can use Electrospray ionization (ESI). ESI induces multiple charges on the intact molecule

Role of Electrospray Ionization

Since ESI induces multiple charges on the precursor molecule, there is a good chance that upon precursor's fragmentation, each fragment will have a portion of the charge. ESI allows for production of multiple charged ions. Tandem MS helps measure molecular weights of ionized fragments

Module 037: Example Case Study – III

Tandem MS helps measure the mass of the fragments. Those fragments which differ from each other by one amino acid's mass can provide clues on the sequence of proteins.

Consider a random protein sequence from DB:



Figure 37 Example peptide sequence tags

Peptide sequence tags help derive clues about the sequence of precursor proteins/peptides. The short peptide sequences help us in shortlisting candidate proteins from the database.

Module 038: Example Case Study – IV

MS1 helps measure the intact mass of proteins/peptides. A list of candidate proteins/peptides can be formed by comparing MS1 mass to the mass of proteins/peptides in the database. MS2 or Tandem MS was performed after fragmentation of intact proteins. MS2 helped extract peptide sequence tags from MS2 data. Candidate proteins can be further shortlisted by the PSTs

Exhaustive matching of all MS2 peaks with the theoretical fragments of candidate proteins. The set of theoretical fragments contains all possibilities of fragmentation

Theoretical vs experimental fragments comparison helps as the third stage for shortlisting candidate protein list. This shortlisting will help you arrive at a small number of proteins

Module 039: Example Case Study – V

MS1 and MS2 provide mass of intact molecules and its fragments. This information helps filter proteins from protein database. For a quantitative measure, scoring scheme is required.

MAFSAEDVLKE...

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

Figure 38 Example intact protein mass score

MAFSAEDVLKE...

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

Figure 39 Example peptide sequence tags

Three scoring schemes can be applied to score the match at each stage of protein search. These scoring elements can be integrated to arrive at an overall candidate protein score.

Module 040: Example Case Study – VI

Comparisons can be performed at various levels of information. These include MS1, MS2, PSTs and theoretical fragments comparison. Integrated scoring schemes couple these factors.

Simply sum the scores up (a linear function)

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

For comprehensive scoring

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

A comprehensive scoring scheme can combine all the scores. Several optimizations can be undertaken on the scoring scheme to further improve protein identification

Chapter 6 - Protein Structures

Module 001: PROPERTIES OF AMINO ACIDS-I

Proteins are made by polymerization of amino acids on ribosomes and proteins properties are linked to the properties of amino acids. There are 20 amino acids in nature each has different chemical composition and that's why each protein is different from other.

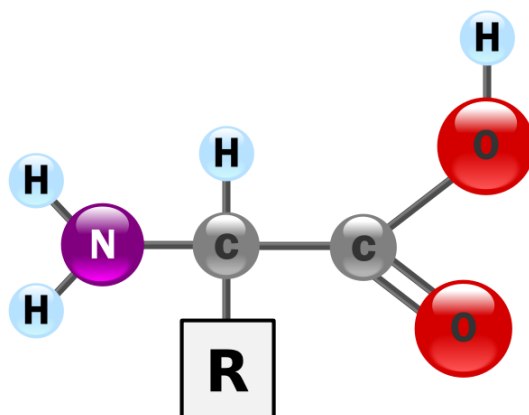


Figure 0.1 chemical structure of amino acid

Amino acid have three groups, hydroxyl group, Amine group and R group. The R group is representing any group.

Periodic Chart of Amino Acids
www.bachem.com

His 155.15 137.14 $C_6H_9N_3O_2$ Histidine					Asp 133.07 119.09 $C_4H_7NO_4$ Aspartic Acid
Arg 174.20 156.19 $C_6H_{14}N_4O_2$ Arginine	Phe 165.19 147.17 $C_9H_9NO_2$ Phenylalanine	Ala 89.09 71.08 $C_3H_7NO_2$ Alanine	Cys 121.16 103.14 $C_3H_7NO_2S$ Cysteine	Gly 75.07 57.05 $C_2H_5NO_2$ Glycine	Gln 146.15 128.13 $C_5H_9N_2O_3$ Glutamine
Lys 146.15 128.17 $C_6H_{11}N_3O_2$ Lysine	Leu 131.17 113.16 $C_6H_{13}NO_2$ Leucine	Met 140.21 131.20 $C_5H_9NO_2S$ Methionine	Asn 132.12 114.10 $C_4H_8N_2O_3$ Asparagine	Ser 105.09 87.08 $C_3H_7NO_3$ Serine	Tyr 181.19 163.17 $C_9H_9NO_3$ Tyrosine
Ile 131.17 113.16 $C_6H_{13}NO_2$ Isoleucine	Trp 204.23 186.21 $C_{11}H_{13}NO_2$ Tryptophan	Pro 115.13 97.12 $C_5H_9NO_2$ Proline	Val 117.15 99.13 $C_6H_{11}NO_2$ Valine	Ser 105.09 87.08 $C_3H_7NO_3$ Serine	

Legend:

- Basic: Blue
- Nonpolar (hydrophobic): Green
- Polar, uncharged: Yellow
- Acidic: Red

1-Letter Amino Acid Code: S
 Molecular Weight: 105.09
 MW/H₂O: 87.08
 Molecular Formula: C₃H₇NO₃
 Chemical Structure:
 Chemical Name: Serine

Created with a 102 version of BatchDraw
Registered version doesn't display this notice

Figure 0.2 periodic chart of amino acid

During polymerization of amino acids the water is formed and amino acids attached with each other.

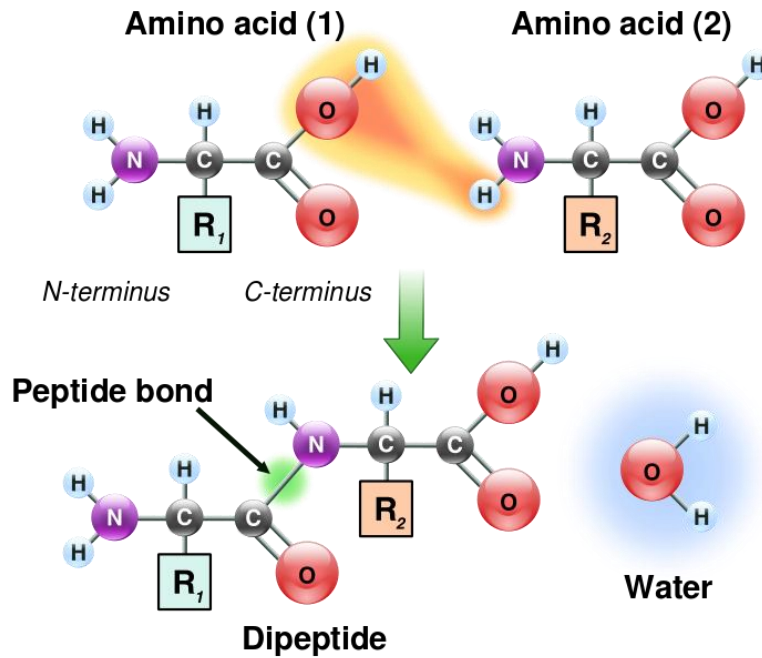


Figure 0.3 polymerization of amino acids

Amino acids have unique properties such as polarity, charge states and interactions with water. Each of these properties describes the overall characteristic of an amino acid.

Module 002: PROPERTIES OF AMINO ACIDS-II

Amino acids have characteristics like polarity, hydrophobicity, and charge states. These characteristics are governed by the elemental composition of an amino acid's side chain (R group).

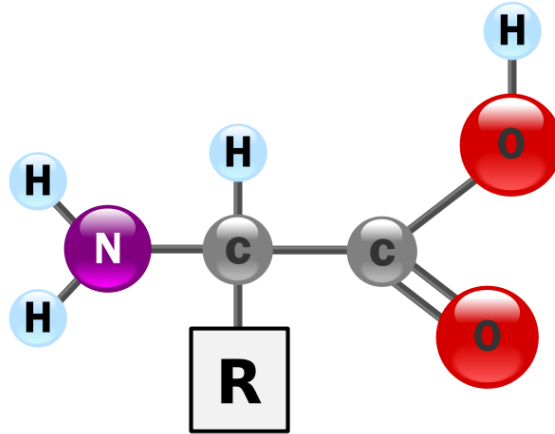


Figure 4 R group in amino acid

HYDROPHILIC AMINO ACIDS

Since H and C introduce very little dipole moments in hydrophobic amino acids, these amino acids are non-polar. Hydrophobic amino acids are mostly found at the inside of folded proteins. Hydrophilic group contain the chain of C and H group in their R group.

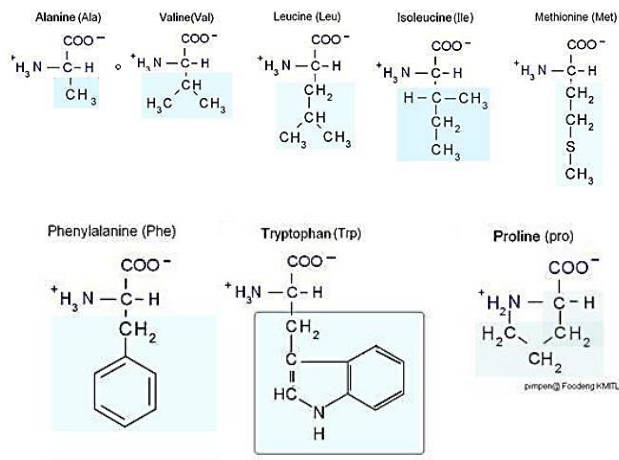


Figure 5 hydrophilic group

POLAR AMINO ACID

These amino acids are polar but are not charged i.e. no net charge on the amino acid. Prefer to reside / interact with aqueous environments. Mostly found at the surface of folded proteins.

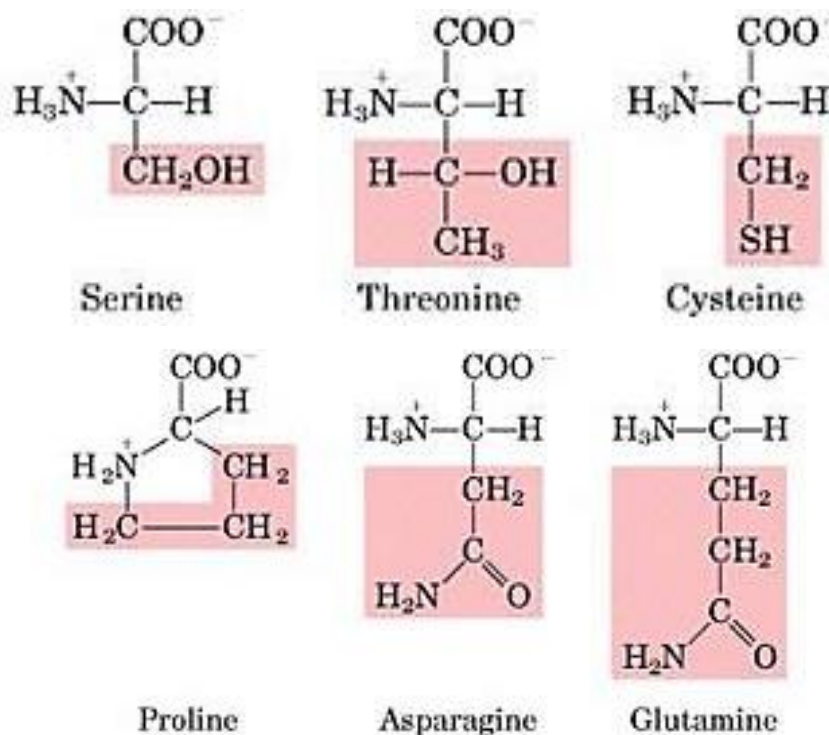


Figure 6 Polar amino acids

Amino acids have unique properties such as polarity, charge states and interactions with water. Each of these properties describes the overall characteristic of an amino acid.

Module 003: PROPERTIES OF AMINO ACIDS-III

Some amino acids are positively charged and some have negative charge.

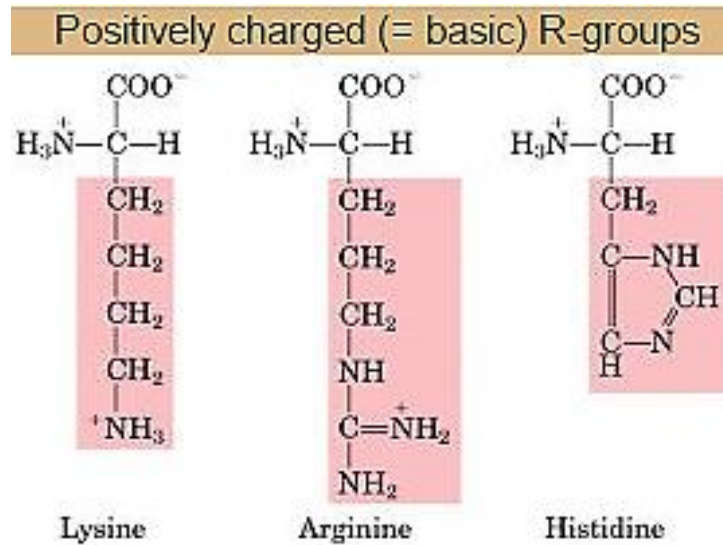


Figure 7 positively charged amino acids

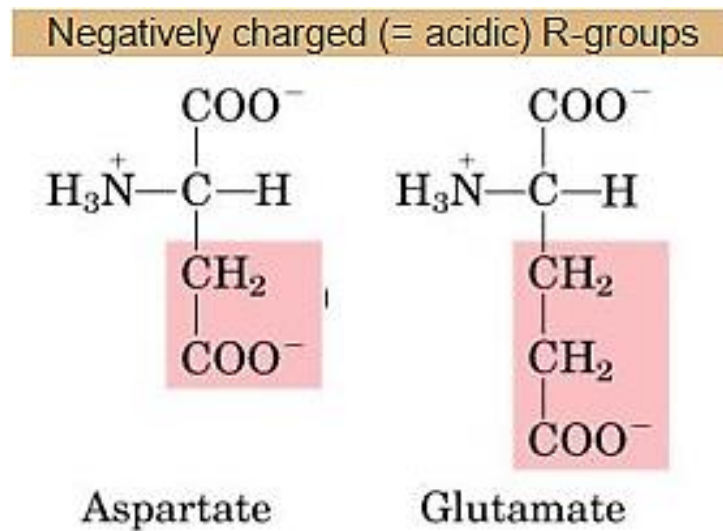


Figure 8 negative charge amino acid

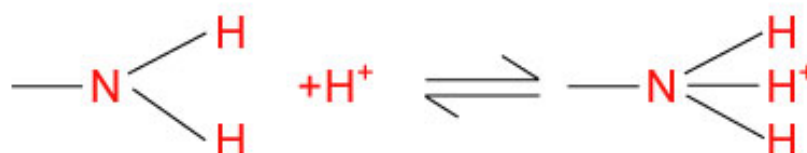
Upon polymerization of amino acids into polypeptide chains, charged amino acids get neutralized. At pH=7, five amino acids are charged, 2 negatively and 3 positively.

Module 004: PROPERTIES OF AMINO ACIDS-IV

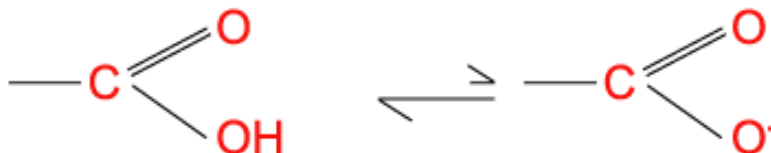
Some amino acids are positively charged and some have negative charge. pK is the values for an amino acid is the pH at which exactly half of the chargeable group is charged.

	Amino acid	pK of the side chain group
-ve	Aspartic acid	3.9
	Glutamic acid	4.2
+ve	Lysine	10.5
	Arginine	12.5
	Histidine	6.0

If $\text{pH} < \text{pK}$ for an amino acid, the amine side chains gain a proton (H^+) and become positively charged, hence basic.



If $\text{pH} > \text{pK}$ for an amino acid, the carboxyl side chains loses a proton (H^+) and become negatively charged, hence acidic.



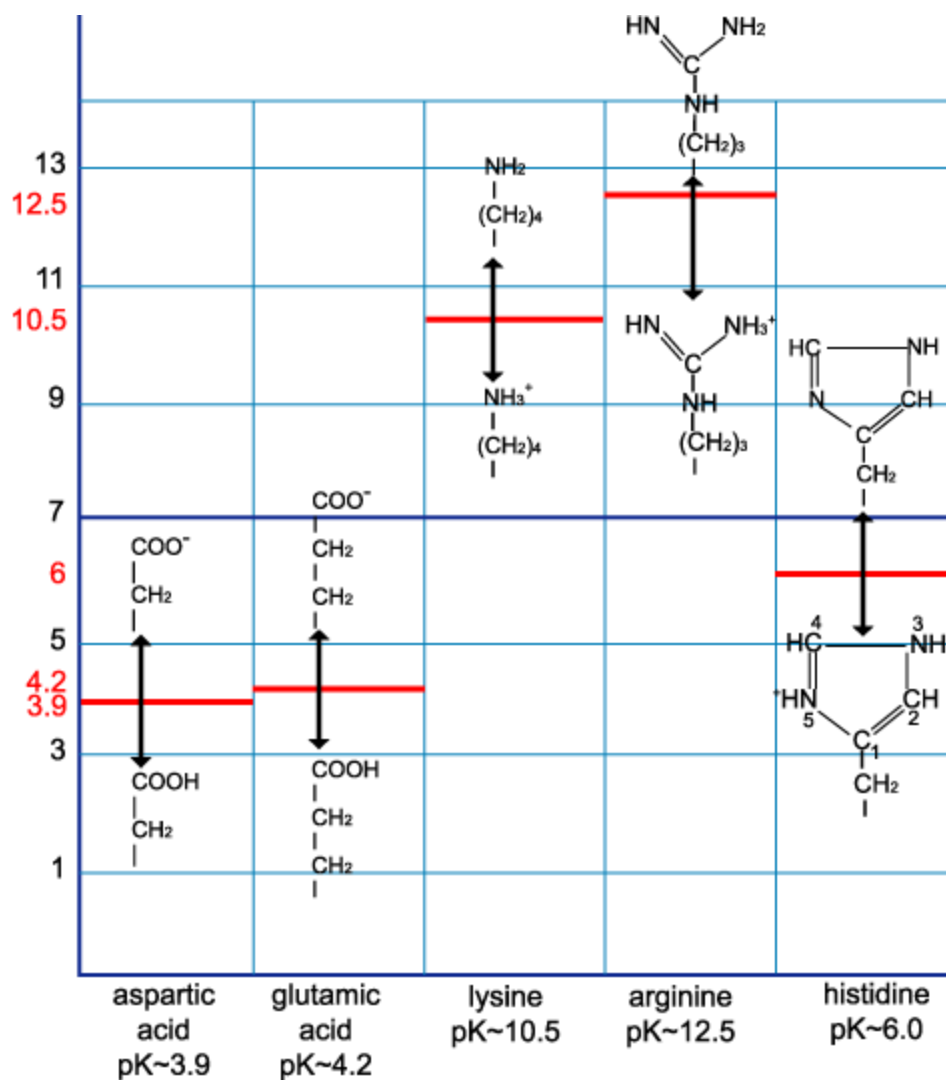


Figure 9 properties of amino acids according to pK and PH.

Depending on the pH, an amino acid may become charged. This may be positive or negative depending on the amino acid.

Module 005: PROPERTIES OF AMINO ACIDS-V

Amino acids may be charged depending on pH. This depends on the charge acceptance or donation from within an amino acid. Additionally, amino acids have structures as well.

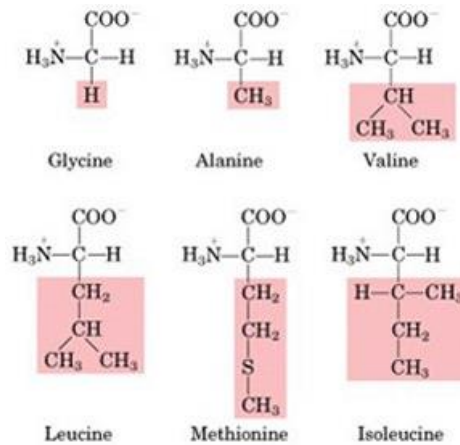


Figure 10 Aliphatic Amino Acids (Non polar C and H chains)

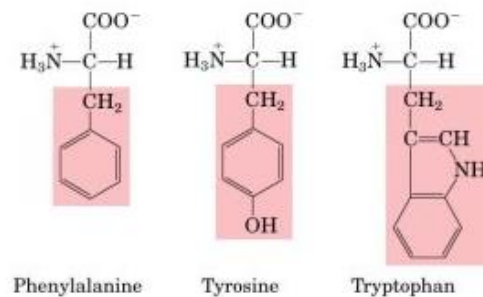


Figure 11 Aromatic R groups

Side chain also impact some properties. Side chains comprising merely of Carbon and Hydrogen are:

- Chemically inert,
- Poorly soluble in water

However, side chains containing organic acids are very different. They are chemically reactive and Soluble in water. Elemental composition plays a very important role in determining properties of amino acids. Solubility and reactivity are key factors participating in protein folding.

Module 006: STRUCTURAL TRAIT OF AMINO ACID-I

Amino acids have several properties such as charge state, polarity and hydrophobicity. It is important to note that the physical size of each amino acid also varies.

EXAMPLE-1: Glycine

Glycine residues increase backbone flexibility because they have no R group (only an H), hence agile.

(d) Glycine



G Gly, Glycine

© 1999 GARLAND PUBLISHING INC.
A member of the Taylor & Francis Group

EXAMPLE-2: Proline

Proline residues reduce the flexibility of polypeptide chains. Proline cis-trans isomerization is often a rate-limiting step in protein folding.

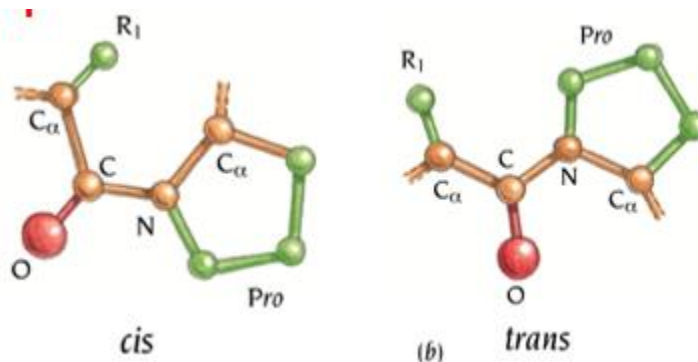


Figure 12 cis and Tran's form of proline

EXAMPLE-3: Cystine

Cysteines cement together by making disulfide bonds to stabilize 3-D protein structures. In eukaryotes, disulfide bonds can be found in secreted proteins or extracellular domains.

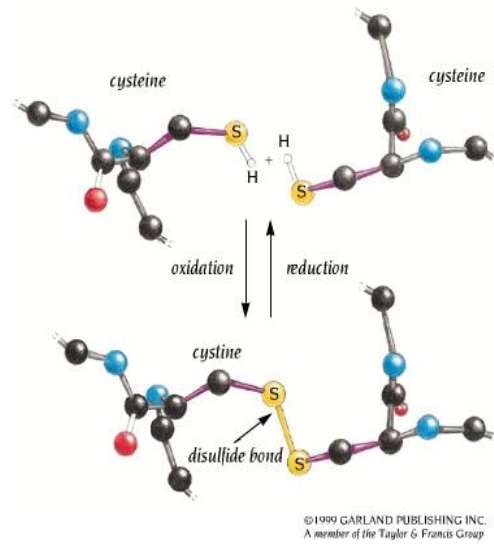


Figure 13 cystine

Amino acids not only have physical and chemical properties, but also structural properties. These structural properties are equally important in giving rise to protein structures.

Module 007: STRUCTURAL TRAIT OF AMINO ACID-II

Each amino acid has a unique set of properties such as charge state, polarity and hydrophobicity. Moreover, it may have unique structural traits as well which can help in protein folding. Since some amino acids are hydrophobic, they may be employed in forming a stable core in a protein. Also, chemically inactive amino acids reduce chances of destabilizing reactions in core.

There comes a problem in burying hydrophobic amino acids in protein core Backbone is highly polar (hydrophilic) due to polar -NH and C=O in each peptide unit; these polar groups must be neutralized.

Form regular secondary structures!

Such as:

- Alpha Helices
- Beta Sheets

Which are stabilized by H-bonds!

Module 008: STRUCTURAL TRAIT OF AMINO ACID-III

The size and structure of each amino acid is unique. Coupled with their chemical properties, each amino acid can uniquely contribute in the protein folding process.

Hydrophobic core formed by packed secondary structural elements provides compact, stable core. Upon establishment of a stable protein core, unstable or reactive groups can be added.

"Functional groups" of protein are attached to the hydrophobic core framework. Surface or a protein or its exterior must have more flexible regions (loops) and polar/charged residues.

The very few hydrophobic "patches" on protein surface are involved in protein-protein interactions. The active regions in a protein are almost all present on the surface.

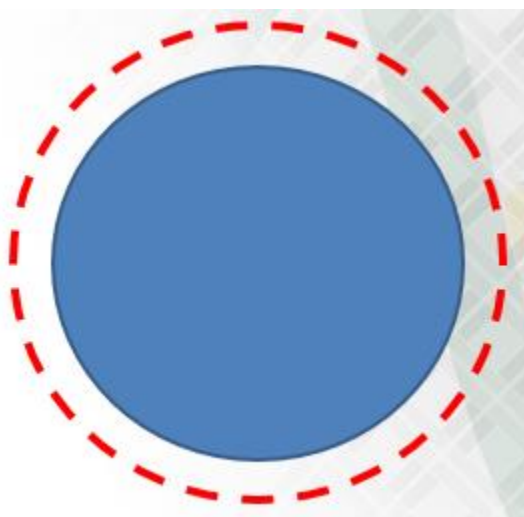


Figure 0.44 Organization of core and surface in a protein

Each component of the protein structure has a unique and precise role in the construction of proteins. Hydrophobic and hydrophilic components have equally useful roles.