AN INTRODUCTORY COURSE

BIOINFORMATICS-I

A STUDENT HANDOUT

2016

VER 1.0.0.0

LAST REVISED ON May 2, 2016

Contents

Chapter 1- Introduction to Bioinformatics

Module 001: INTRODUCTION TO BIOINFORMATICS

- **BACKGROUND**

Bioinformatics is an interdisciplinary science at the cross-roads of biology, mathematics, computer science, chemistry and physics. With the digitalization of the biological information, doors have been wide opened towards the analysis of this information using computer algorithms and software.

Now we know well that the human genome has over 25,000 genes and these genes code for thousands of different proteins which perform day-to-day functions in the living cell. Furthermore, these proteins may take on various post-translational modifications leading to a very large number of functionally unique molecules. This presents us with a huge challenge in identification of genes and proteins.

- **EXPERIMENTS IN BIOLOGY**

With the advancements in experimental protocols, now we have several next generation instruments and techniques available for obtaining digitalized biological information on genes and proteins etc. These instruments include:

1. Next Generation Sequencers (NGS) for whole genome sequencing
2. High Resolution Mass Spectrometry for whole proteome profiling
3. Nuclear Magnetic Resonance Spectroscopy for structural studies

- **DIGITALIZATION OF BIOLOGY**

In today’s world, when a biologist performs an experiment in the wet-lab, he or she in fact produces digital data which is continuously being stored on computer disks. The data may include text, numbers, symbols or images.

- **SPEED OF DATA GROWTH**

Due to advancement in instrumentation used in biological experiments, data is being accumulated at exponentially increasing rates. For example; genome sequences in genome databases are doubling every few years.

- **CONCLUSION**

Human brain is limited in recalling information from memory. First, we have to commit all information to our memory followed by its recall. To overcome our ability to memorize and recall, computers can come to our rescue. This is because computers have an infinite ability to recall this information and process it quickly towards results.

Module 002: INTRODUCTION TO BIOINFORMATICS

- **MOTIVATION**

Bioinformatics is a becoming a popular science due to several reasons.

- It is an i**nterdisciplinary field** as it covers the information of biological digital information including human, plants, animals and microorganisms.
- Although it is a new field but it is rapidly **developing field**.
- It demands a very **low cost** infrastructure and hardly any lab equipment.
- As bioinformatics data concerns a wide range of species such as humans, plants and micro-organisms, it presents us with **plenty of opportunities** in scientific discovery.

- **SCOPE OF BIOINFORMATICS**

Bioinformatics primarily deals with digitalized biological information as well as data reported from biology experiments. Computational methods, data processing techniques and algorithms are employed in addressing the following issues:

- Storage of data
- Organization data
- Analysis of many experiments
- For representation of biological information

- **ACTIVITIES**

In modern biological sciences, bioinformatics is used for activities such as:

- Developing algorithms for organizing data collected from experiments
- Writing software and tools for data analysis
- Data processing to determine the role of underlying biomolecules
- Statistical evaluation of data using methods such as t-test and ANOVA
- Data visualization for meaningful presentation of biological information

- **CONCLUSION**

In Pakistan, the field of biology is undergoing a rapid change due to the onset of bioinformatics. New research and educational programs are being constructed which is opening new door of opportunities for our future generations.

Module 003: INTRODUCTION TO BIOINFORMATICS

- **NEED FOR BIOINFORMATICS –I**

If we look at the pace of development in the area of bioinformatics then we can easily observe that from year's 2000 to 2015, the number of online tools for processing genomics and proteomics information are rapidly increasing. This is just a reflection of the need for bioinformatics in modern day biology.

The field of Bioinformatics and Computational Biology is characterized by a highly diverse confluence of traditional academic disciplines. Informatics and Bio-science are the umbrella terms given to a set of allied disciplines which make up the field, but a much larger array of traditional areas contribute to the set of tools needed by individuals training for this new and expanding interdisciplinary field. Biomedical Engineering, Electrical and Computer Engineering, Computer Science, Applied Mathematics, Genetics, Biology, Anatomy and Cell Biology, Micro Biology, and Biostatistics are the principal allied disciplines.

- **CONCLUSION**

The need for bioinformatics is on a rapid rise as biological data is rapidly increasing and becoming available online, free of any cost.

Module 004: NEED FOR BIOINFORMATICS –II

If we observe the growth of gene bank than from 1982 it comprised of 2 billion base pairs but by year 2002 it had risen to 56 billion base pairs. With the data in our hands, there is an urgent need to interpret this data. For instance, analysis of this data can help us in developing an understanding of the phylogenetic "tree of life" which consist of:

- Bacteria
- Archaea
- Eucarya

Towards exploring the possible benefits of using bioinformatics, one needs to answer the following question:

- **WHAT IS IT THAT BIOINFORMATICS CAN DELEIVER?**

The simple answer to that bioinformatics is:

- Provide us better understanding of life, evolution, molecular mechanisms as well as disease.
- Moreover, we can make better drugs with the availability of an enhanced molecular understanding of disease.

- **POSSIBLE CONTRIBUTIONS**
  - It can help us to organize the large datasets from new experiments instruments
  - Bioinformatics can help store and process this data as well.
  - It can provide insights into the meanings of our research results and findings.
  - Overall, it can help us to better understand paradoxes defining the life forms.

- **CONCLUSION**

From gene sequencing to protein sequencing, bioinformatics is providing us with an improved understanding of the genes, proteins, protein interaction and signaling pathways involved in biological functioning and disease.

Module 005: APPLICATIONS OF BIOINFORMATICS – I

When we look at bioinformatics, it seems to be a very complex and abstract field. How and where can bioinformatics be applied specifically? How does it improve the fundamental understanding of biological phenomenon? Most importantly, how can its benefits be delivered to the society at large?

The answers to these questions are categorized as follows:

- **GENOMICS**
  - ➢ Bioinformatics can help in assembling DNA sequencing data.
  - ➢ It can help in gene finding (markers).
  - ➢ Gene assembly can be performed using bioinformatics tools (nucleotide alignments)
  - ➢ It can help transcribe the gene data to RNA data
  - ➢ Also, databases can be generated from such data.

- **EVOLUTIONARY STUDIES**
  - ➢ Evolutionary relationships between different organisms can be derived from data.
  - ➢ Evolutionary distance among species can be computed by using bioinformatics tools.
  - ➢ Phylogenetic trees can be constructed to find relationships between species.
  - ➢ Ancestry can be better understood between several species and organisms.

- **PROTEOMICS**
  - ➢ Bioinformatics can help us in decoding protein sequences.
  - ➢ It can also help us in understanding protein structure.
  - ➢ We can also understand post translational changes in proteins with the help of bioinformatics.
  - ➢ We can better understand the protein-protein interaction in different biological reactions.
  - ➢ It can also help us in generating databases of these sequences and structures.

- **SYSTEMS BIOLOGY**
  - ➢ Bioinformatics can assist us in modelling regulatory mechanisms in gene and protein networks.
  - ➢ Such models can be analyzed to identify the key regulators in these networks.
  - ➢ Moreover, the models can help evaluate drugs to treat these key regulators.

- **CONCLUSION**

  Bioinformatics can be applied to life in many ways it helps us to understand the sequence and function of biomolecules and their relationships. Recent trends in bioinformatics involve development of personalized therapeutics for cancer and diabetes.

Module 006: APPLICATIONS OF BIOINFORMATICS - II

Bioinformatics is being applied in routine life in many ways like in Genomics, transcriptomics, Proteomics, Metabolomics, Structural Proteomics, Designing Drugs, System Biology and in personalization of medicines for cure.

Except these applications Bioinformatics introduced us the techniques which enabled us to generate the large data regarding biology and also its use. And step by step the applications of bioinformatics increased from genomic level to entire system level.

- **SMALL TO BIG**

- ➢ Bioinformatics helps us to understand the systems from small to big like from gene findings to entire system prediction
- ➢ In structure findings and modeling of many biological system to understand them in better ways.
- ➢ Bioinformatics helped the human to understand the protein, protein interaction in many biological systems.
- ➢ And provide us the concept how these biological process are interconnected with each other and how they affect each other.
- ➢ Now we are able to understand the modeling of molecules and genome at cell level.
- ➢ Signaling pathways are easy just because of bioinformatics.
- ➢ Now morphology of tissue can be understand by creating the models with help of bioinformatics tools.

- **CONCLUSION**

Bioinformatics not only just collect, analyze and store the data it process it in very authentic way and validates our hypothesis and very soon in future it will help us to understand that which disease is coming in future and how to tackle it with personalize medicine.

Module 007: FRONTIERS IN BIOINFORMATICS - I

- **INTROCDUCTION**

Bioinformatics is new and emerging field of science having vast opportunities and with innovation in tools it is increasing the scale of biological data, but still there are many unsolved challenges which are pending in the field of life science and for which bioinformatics is doing new innovative ideas.

- **FRONTIER IN GENOMICS**

Now we are able to sequence the whole genome with the bioinformatics tool of Next generation sequencing (NGS)

We are able to save, store and analyze the massive amount of biological data which is in (Terabyte files)

We can handle the large number of data easily and can process it as well in easy way.

Whole genome can be assemble in sequence and can flaws can be identified easily.

- **FRONTIER IN TRANSCRIPTOMICS**

Now in genomics we are able to identify those matters which are unknown yet or under discussion.

Role of RNA in making proteins and its dynamics can be understood easily now.

Interactions of RNA molecule can be easily understood by simple model.

- **FRONTIER IN PROTEOMICS**

Deficiency of low proteins in any patient tissue sample can be identify.

Expression and manufacture of protein in large molecular level in any organism can be identified.

Pathways before and after any biological reaction are easy to design.

- **CONCLUSION**

Bioinformatics is literally a science full of challenges and opportunities having a revolution in field of biology and routine life.

Module 008: FRONTIER IN BIOINFORMATICS-II

Frontier in Bioinformatics includes

- Next generation genomics
- Transcriptomics
- Proteomics

- **FRONTIER IN PROTEIN STURUCTURE**

Bioinformatics helps us to understand the layer folding of proteins that how they are processed, and helps to know that how protein interact with each other and how a drug can affect or stimulate a protein.

- **FRONTIER IN SYSTEM BIOLOGY**

It helps us to understand the whole system of a single cell, in that cell how organelles, gene, proteins and metabolites are interconnected in a single unified system (cell). And bioinformatics also give us the idea how these models can be applied to real-time.

- **FRONTIER IN PERSONALIZED MEDICINE**

This is the important thing for this century and upcoming generation that personalize the medicine for exact cure of a disease. Because all the medicine cannot work exact some effect patient badly therefor with the help of Bioinformatics we are now able to personalize some medicines for some diseases. And bioinformatics helps us to evaluate the medicine.

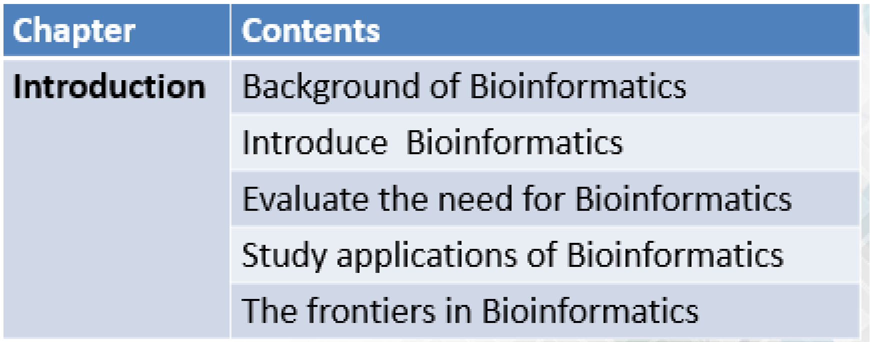
- **CONCLUSION**

If we talk about the 21st century than it’s the century of bioinformatics it will enable the human to cure many disease with one drug by personalizing it.

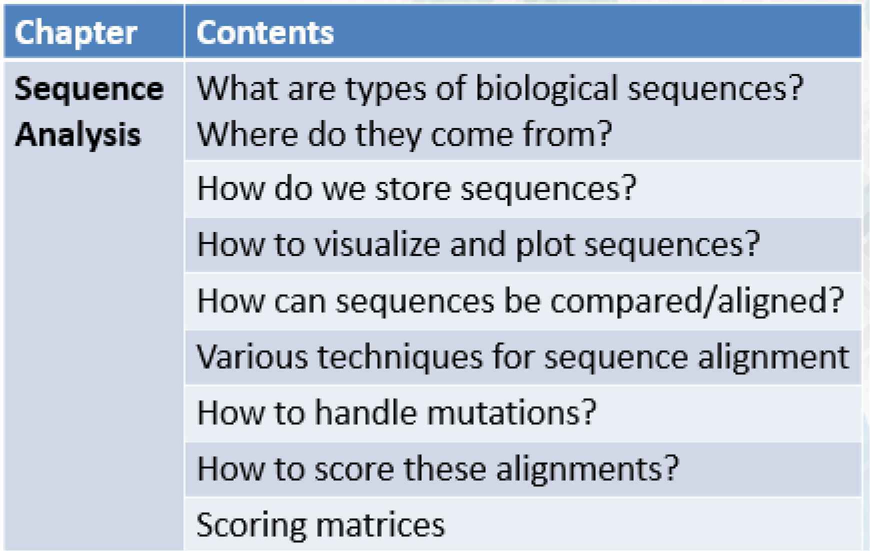
Module 009: Overview of Course Contents - I

Philosophy behind the Course Outlay

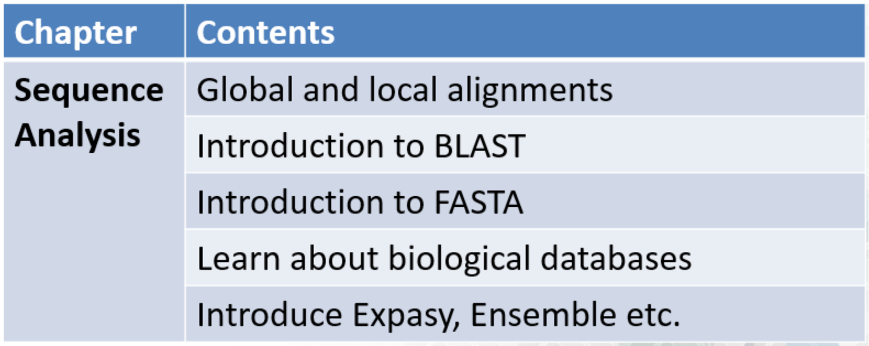
1. Introduce the classical algorithms in bioinformatics
2. Link them to latest developments in the field
3. Evaluate the future applications



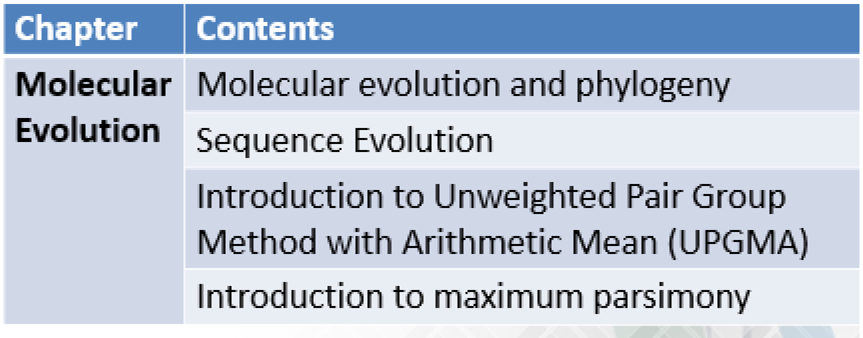
Chapter	Contents
Introduction	Background of Bioinformatics
	Introduce Bioinformatics
	Evaluate the need for Bioinformatics
	Study applications of Bioinformatics
	The frontiers in Bioinformatics



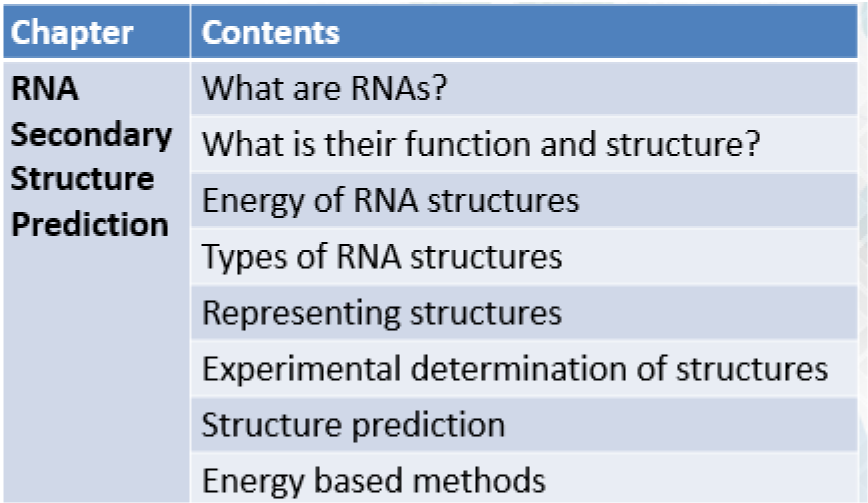
Chapter	Contents
Sequence Analysis	What are types of biological sequences? Where do they come from?
	How do we store sequences?
	How to visualize and plot sequences?
	How can sequences be compared/aligned?
	Various techniques for sequence alignment
	How to handle mutations?
	How to score these alignments?
	Scoring matrices



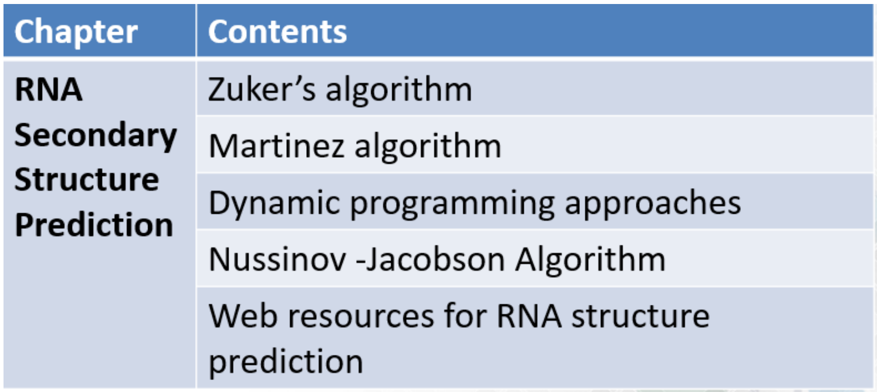
Chapter	Contents
Sequence Analysis	Global and local alignments
	Introduction to BLAST
	Introduction to FASTA
	Learn about biological databases
	Introduce Expasy, Ensemble etc.



Chapter	Contents
Molecular Evolution	Molecular evolution and phylogeny
	Sequence Evolution
	Introduction to Unweighted Pair Group Method with Arithmetic Mean (UPGMA)
	Introduction to maximum parsimony



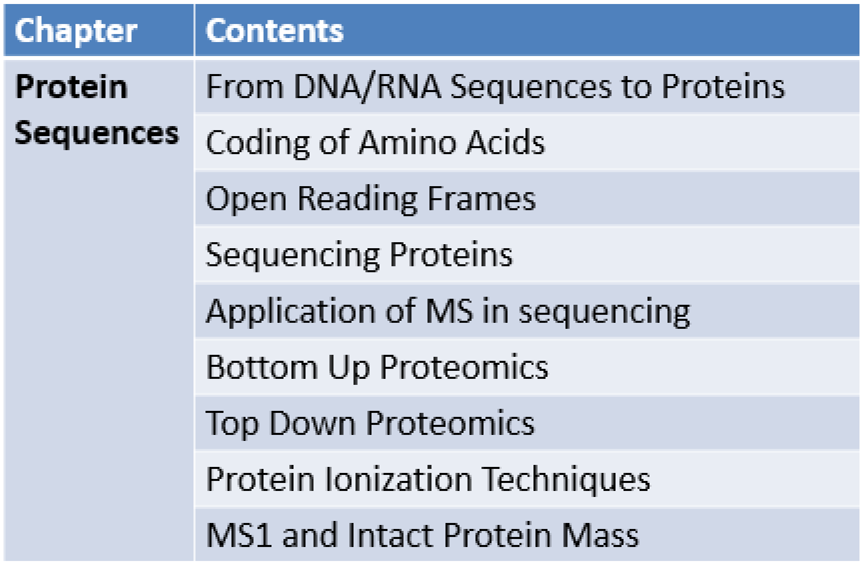
Chapter	Contents
RNA Secondary Structure Prediction	What are RNAs?
	What is their function and structure?
	Energy of RNA structures
	Types of RNA structures
	Representing structures
	Experimental determination of structures
	Structure prediction
	Energy based methods



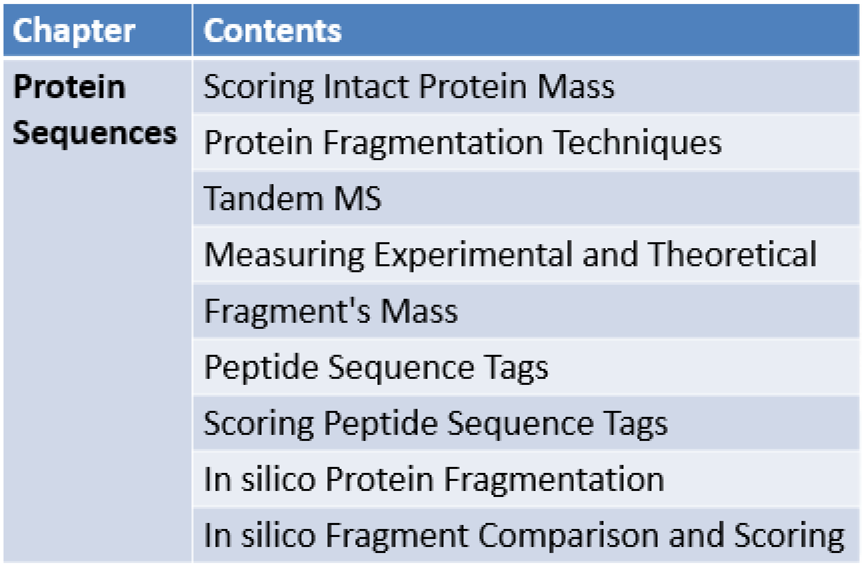
Chapter	Contents
RNA Secondary Structure Prediction	Zuker’s algorithm
	Martinez algorithm
	Dynamic programming approaches
	Nussinov -Jacobson Algorithm
	Web resources for RNA structure prediction

Sequences and operations such as alignment and comparison will be covered along with phylogenetic and RNA structure modelling. Next up we will delve into protein sequences and structures!

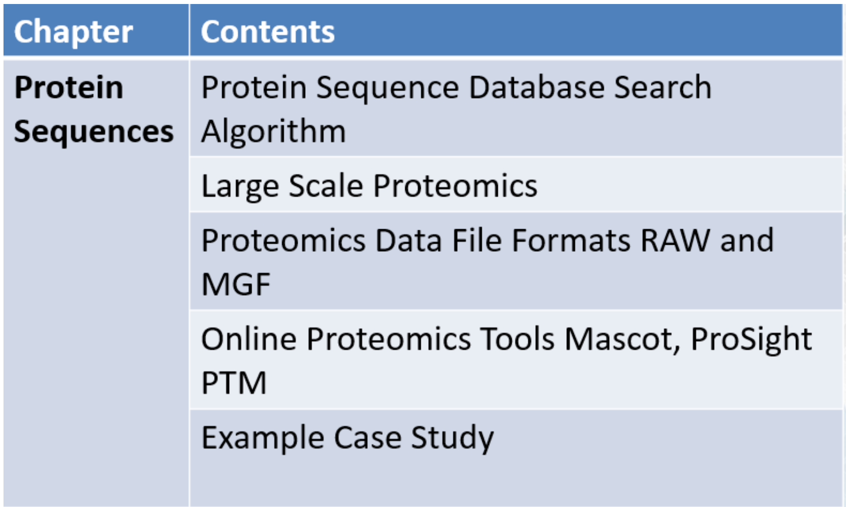
Module 010: Overview of Course Contents - II



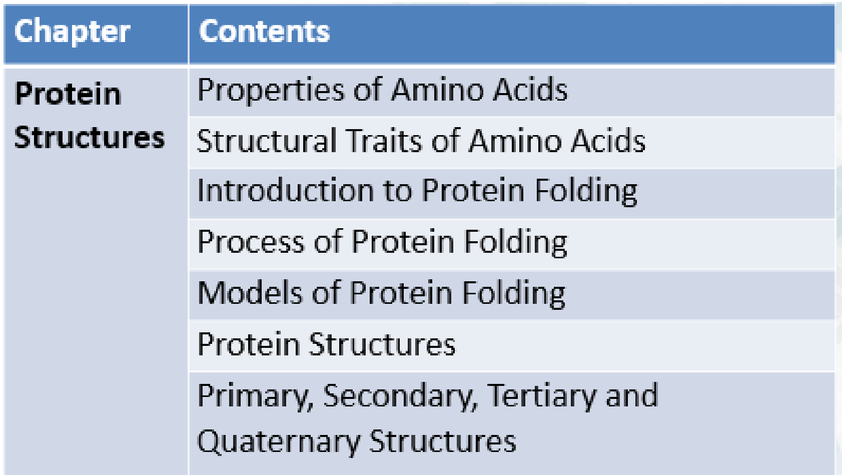
Chapter	Contents
Protein Sequences	From DNA/RNA Sequences to Proteins
	Coding of Amino Acids
	Open Reading Frames
	Sequencing Proteins
	Application of MS in sequencing
	Bottom Up Proteomics
	Top Down Proteomics
	Protein Ionization Techniques
	MS1 and Intact Protein Mass



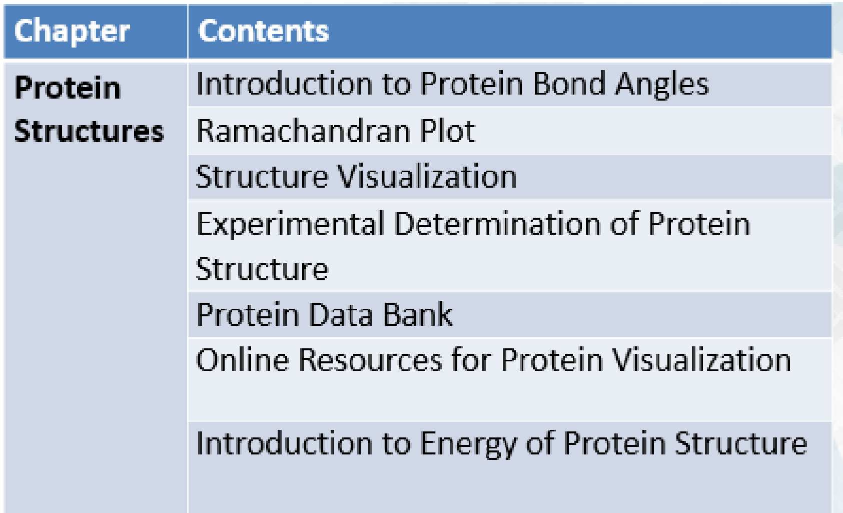
Chapter	Contents
Protein Sequences	Scoring Intact Protein Mass
	Protein Fragmentation Techniques
	Tandem MS
	Measuring Experimental and Theoretical
	Fragment's Mass
	Peptide Sequence Tags
	Scoring Peptide Sequence Tags
	In silico Protein Fragmentation
	In silico Fragment Comparison and Scoring



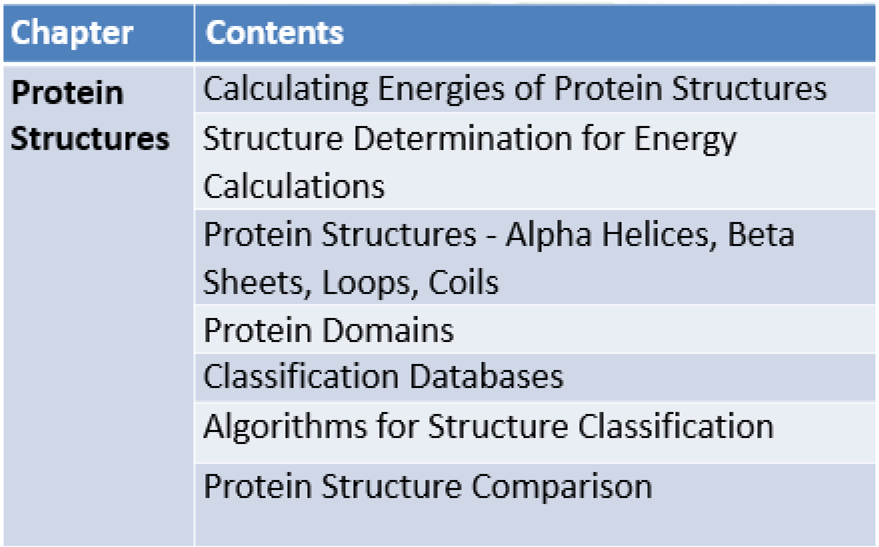
Chapter	Contents
Protein Sequences	Protein Sequence Database Search Algorithm
	Large Scale Proteomics
	Proteomics Data File Formats RAW and MGF
	Online Proteomics Tools Mascot, ProSight PTM
	Example Case Study



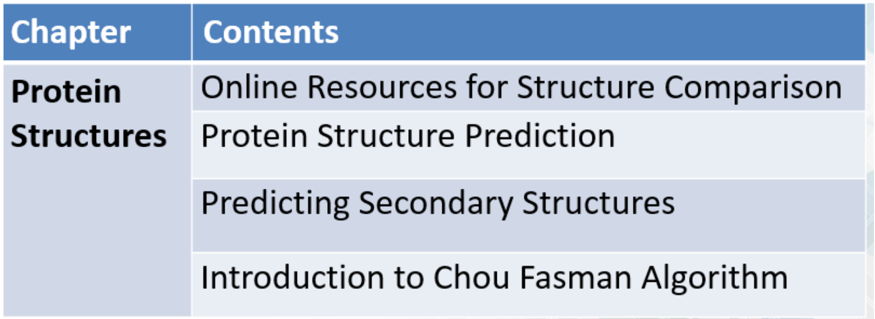
Chapter	Contents
Protein Structures	Properties of Amino Acids
	Structural Traits of Amino Acids
	Introduction to Protein Folding
	Process of Protein Folding
	Models of Protein Folding
	Protein Structures
	Primary, Secondary, Tertiary and Quaternary Structures



Chapter	Contents
Protein Structures	Introduction to Protein Bond Angles
	Ramachandran Plot
	Structure Visualization
	Experimental Determination of Protein Structure
	Protein Data Bank
	Online Resources for Protein Visualization
	Introduction to Energy of Protein Structure



Chapter	Contents
Protein Structures	Calculating Energies of Protein Structures
	Structure Determination for Energy Calculations
	Protein Structures - Alpha Helices, Beta Sheets, Loops, Coils
	Protein Domains
	Classification Databases
	Algorithms for Structure Classification
	Protein Structure Comparison

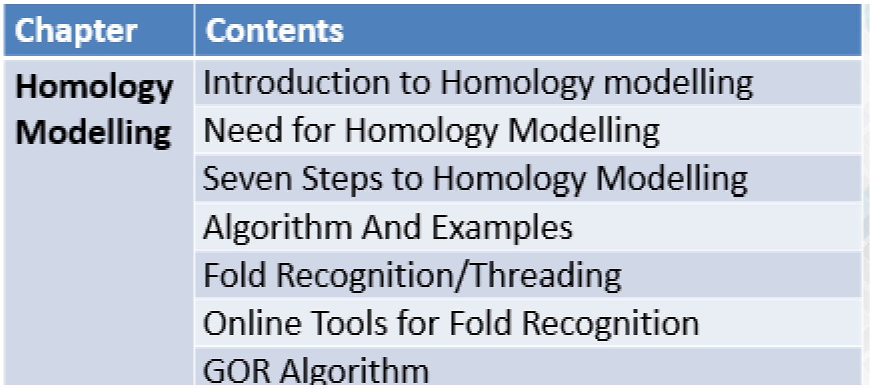


Chapter	Contents
Protein Structures	Online Resources for Structure Comparison
	Protein Structure Prediction
	Predicting Secondary Structures
	Introduction to Chou Fasman Algorithm

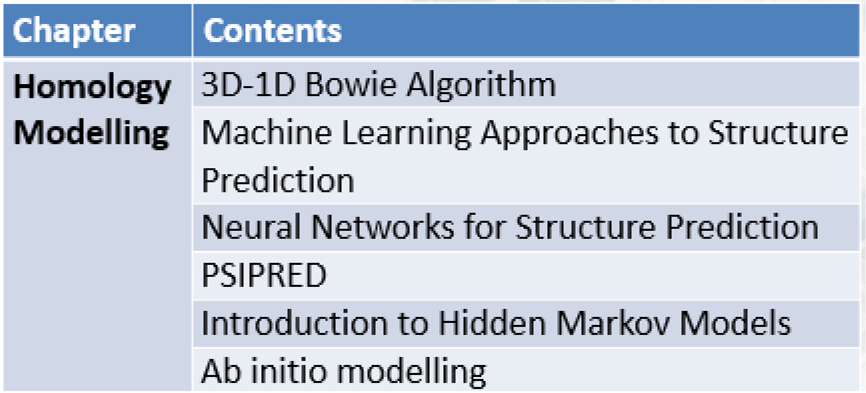
Summary

- Protein sequence and structure topics will be dealt in these modules
- Next set of modules is about the homology modelling and systems biology topics!

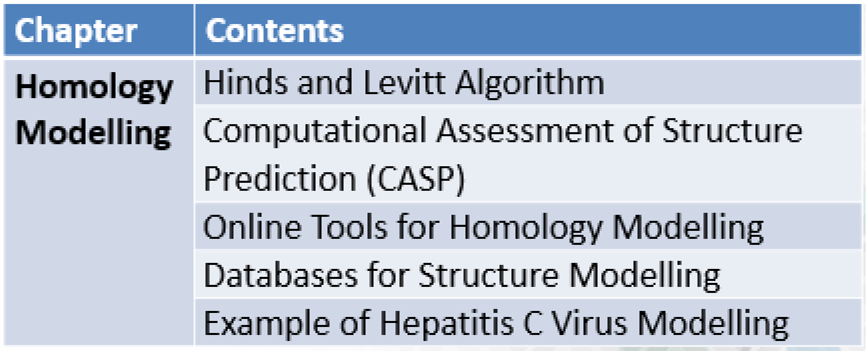
Module 011: Overview of Course Contents – III



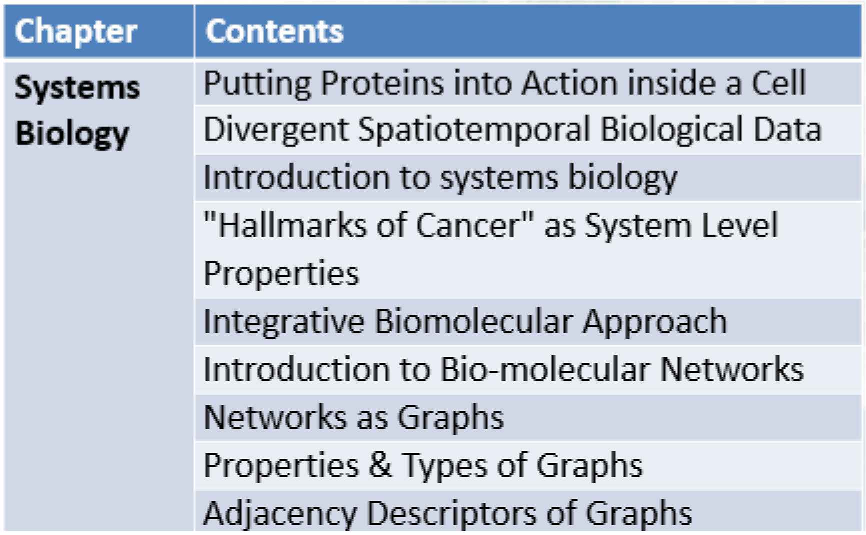
Chapter	Contents
Homology Modelling	Introduction to Homology modelling
	Need for Homology Modelling
	Seven Steps to Homology Modelling
	Algorithm And Examples
	Fold Recognition/Threading
	Online Tools for Fold Recognition
	GOR Algorithm



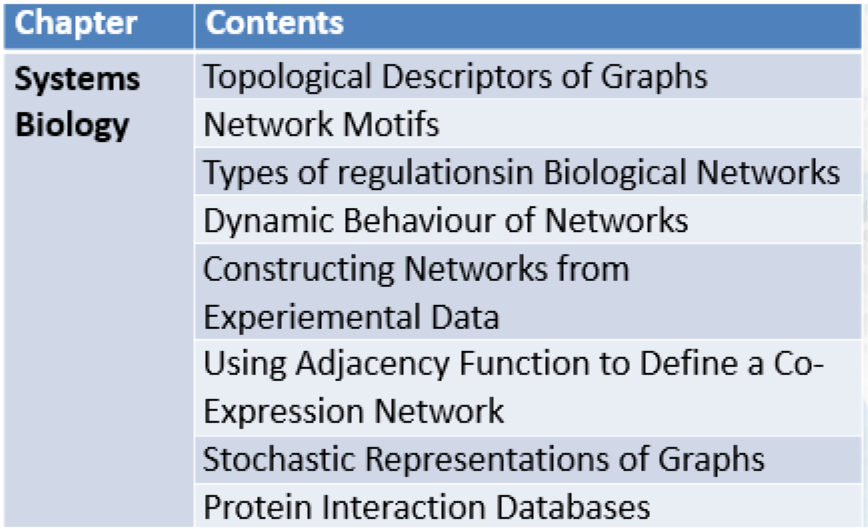
Chapter	Contents
Homology Modelling	3D-1D Bowie Algorithm
	Machine Learning Approaches to Structure Prediction
	Neural Networks for Structure Prediction
	PSIPRED
	Introduction to Hidden Markov Models
	Ab initio modelling



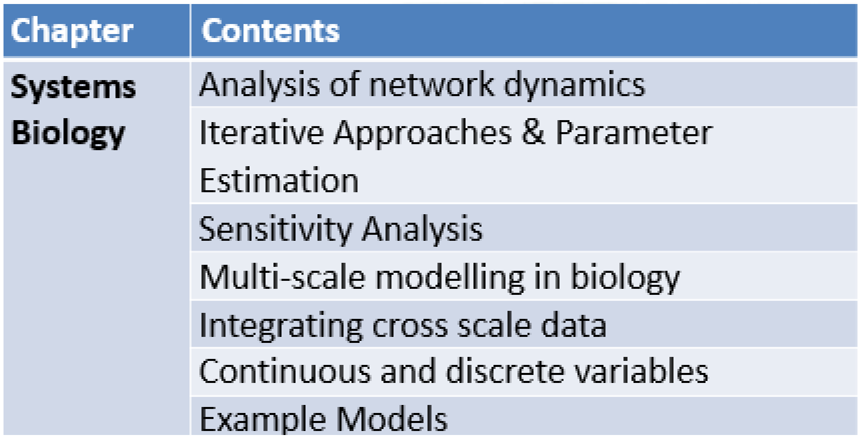
Chapter	Contents
Homology Modelling	Hinds and Levitt Algorithm
	Computational Assessment of Structure Prediction (CASP)
	Online Tools for Homology Modelling
	Databases for Structure Modelling
	Example of Hepatitis C Virus Modelling



Chapter	Contents
Systems Biology	Putting Proteins into Action inside a Cell
	Divergent Spatiotemporal Biological Data
	Introduction to systems biology
	"Hallmarks of Cancer" as System Level Properties
	Integrative Biomolecular Approach
	Introduction to Bio-molecular Networks
	Networks as Graphs
	Properties & Types of Graphs
	Adjacency Descriptors of Graphs



Chapter	Contents
Systems Biology	Topological Descriptors of Graphs
	Network Motifs
	Types of regulationsin Biological Networks
	Dynamic Behaviour of Networks
	Constructing Networks from Experiemental Data
	Using Adjacency Function to Define a Co-Expression Network
	Stochastic Representations of Graphs
	Protein Interaction Databases



Chapter	Contents
Systems Biology	Analysis of network dynamics
	Iterative Approaches & Parameter Estimation
	Sensitivity Analysis
	Multi-scale modelling in biology
	Integrating cross scale data
	Continuous and discrete variables
	Example Models

Conclusion

- These contents will give you an initial exposure to the variety of topics in bioinformatics

- After covering these topics, you should have a basic conceptual foundation for further studies into Bioinformatic

Chapter 2 - Sequence Analysis

Module 001: Gene, mRNA and Protein Sequences

- **INTRODUTION**

We all know that all the living things are composed of cells and here a question arise that how cells are made? For composition of cell DNA has blueprints for building cells along with the information of cell's protein, carbohydrate and vitamins production.

And transfer of this information from DNA to these molecules is termed as "Central Dogma" which is

DNA → RNA → Protein.

Proteins are than use in constructing the cell.

- **DNA**

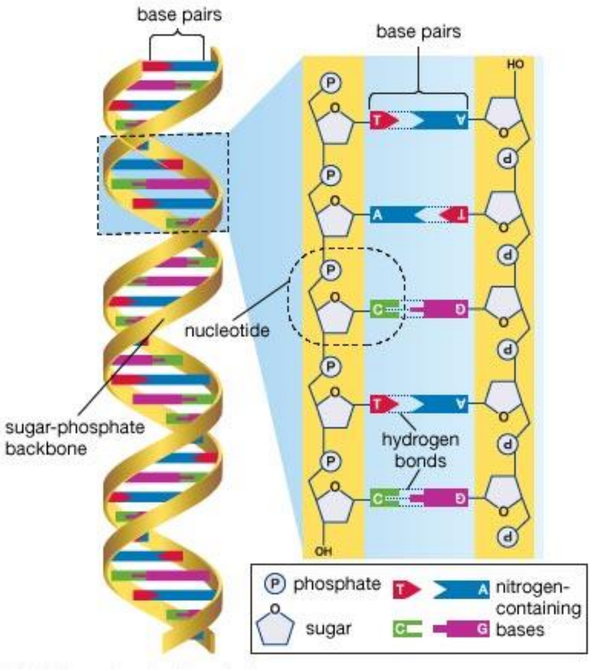


Figure 0.1 DNA Double helix

DNA molecule is double helix structure contain base pairs composed of nucleotides and these nucleotides are composed of sugar phosphate group and are bind with each other with hydrogen bonds.

Normally all the nucleotides are same in both DNA and RNA except one position in RNA which is U (Uracil) and in DNA it is T (Thiamin)

DNA sends the information to cell via mRNA and that sequence the amino acids according to coded information and protein structure is formed and that protein form a cell.

- **CONCLUSION**

According to the central dogma DNA codes information for RNA and RNA makes the Protein and that protein along with some organelles make cells and its systems.

Module 002: TRANSCRIPTION

All cells are made of carbohydrates and proteins and for these cells DNA codes the information which makes the RNA and protein both.

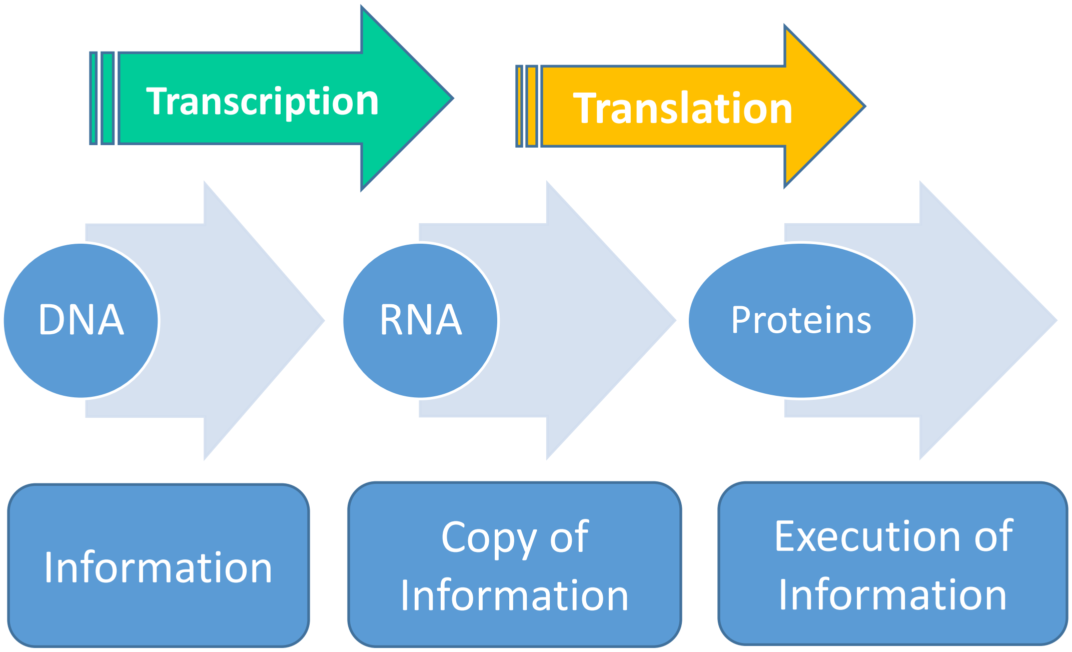


Figure 0.2 Flow of information from DNA to Proteins

The above mechanism explains the process of **transcription** in very simple way, DNA codes the information and converted into RNA where mRNA copies the information and it execute the information in cell and amino acids combine with each other according to coded information of DNA and protein formation takes place. Which is known as **Translation.**

Molecule of DNA contains only four base pairs (A, T, C, and G) which are repeated thousands of time and Adenine "A" pairs with Cytosine "C", While Thymine "T" binds with Guanine "G" and all pairings are with the help of Hydrogen bonding.

Same like DNA, the RNA contains four base pairs but Thymine is replaced with Uracil "U" and RNA is single stranded.

DNA just codes the information for protein but RNA helps in making protein.

Module 003: NUCLEOTIDES

If we talk about the composition of DNA and RNA molecule than these are composed of four other molecules which are named as Nucleotides.

These molecules are Adenine (A), Cytosine (C), Thymine (T), Uracil (U), and Guanine (G).

DNA molecule although is double stranded and RNA is single stranded but there is difference in sugar composition.

RNA has Ribose sugar and DNA has de-oxyribose sugar:

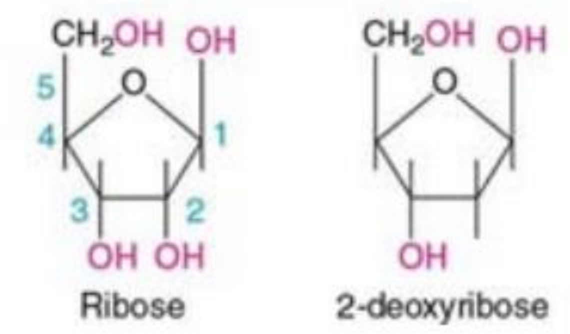
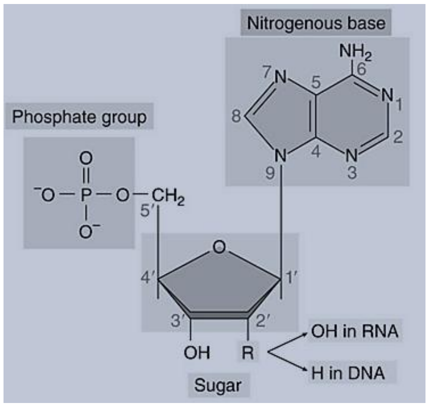


Figure 0.3 Difference between RNA and DNA sugar

RNA **DNA**

Adenine and Guanine collectively called **Purines** while Cytosine, Uracil, and Thymine are called as **Pyrimidine.**

when phosphate, nitrogen base and sugar come together if there is **(OH)** than molecule is RNA and if there is **(H)** in sugar than molecule is DNA. As figure shows.



- **CONCLUSION**

DNA molecule make RNA and RNA make the protein and DNA differ from RNA in nature due to sugar and nucleotide.

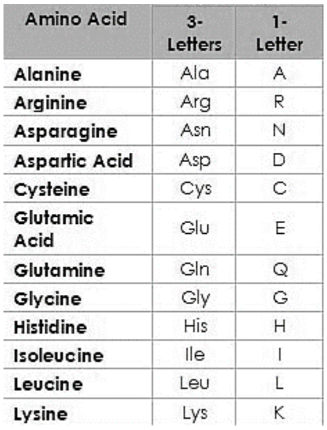
Module 004: TRANSLATION

Cells are built of proteins and carbohydrates and these proteins are made in results of transformation of RNA molecule and this transformation is called as translation.

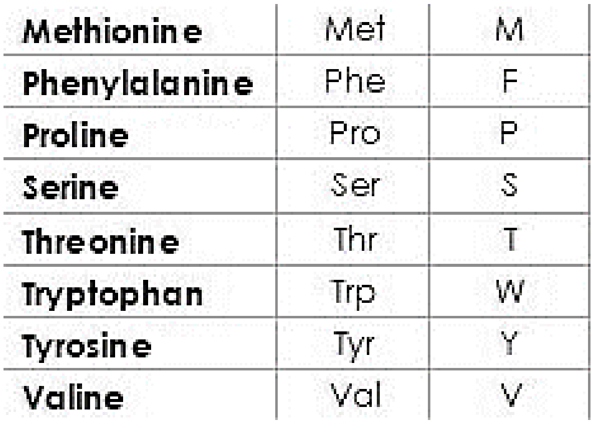
Translation takes place in ribosome of cell and ribosomes after reading the information of mRNA collects the amino acids from cell cytosol which is the part of the cytoplasm that is not held by any of the organelles in the cell.

- **MECHANISM**

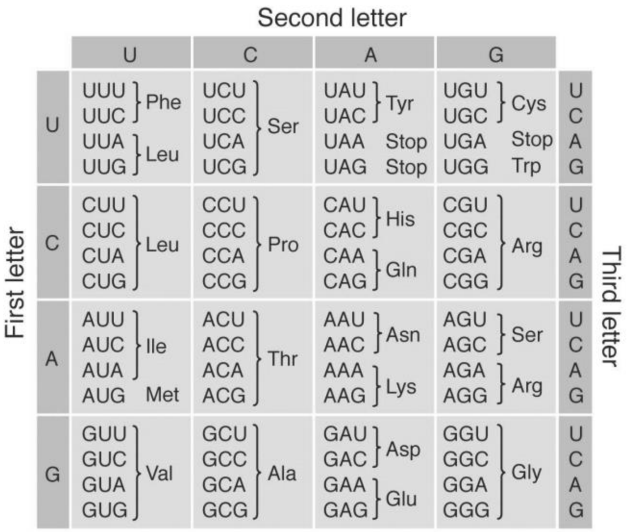
At ribosome three nucleotides are read at a time from mRNA, this set of three nucleotide is called as codon and each codon correspond to a specific amino acid.



Amino Acid	3-Letters	1-Letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic Acid	Asp	D
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K



Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V



Second letter

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

Figure 0.4 sixty four codons combinations

- **CONCLUSION**

RNA codes for protein and codons of here nucleotide code for specific amino acid on ribosomes and this process is called as translation.

Module 005: AMINO ACIDS

RNA decodes the information at ribosomes in form of Codons each codon select a specific amino acid. Because there are 20 different amino acids in nature therefore they fold together and make a protein structure by polymerizing themselves.

If we observe the structure of amino acid it contains nitrogen, hydrogen, oxygen and two carbon atoms. And a variable group R.

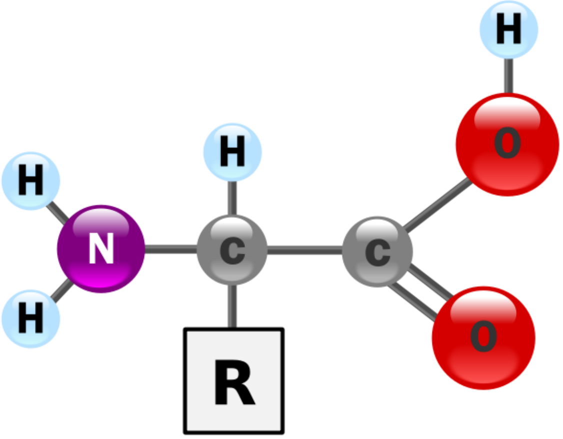
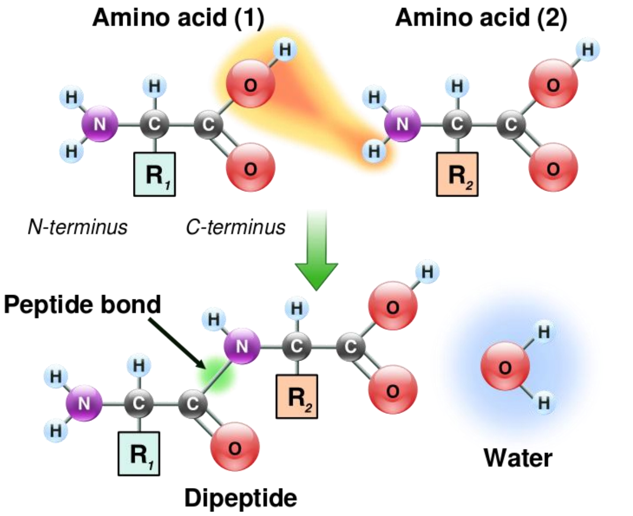


Figure 0.5 structure of amino acid

When polymerizations takes place water is formed and if any compound attached with R group than structure of protein is changed.



These amino acids are joined with each other with peptide bonds and fold with each other in 3D form they make protein structure.

Module 006: STORAGE OF BIOLOGICAL SEQUENCE INFORMATION

We know that sequence of DNA contain A,C,T&G nucleotides and sequence of RNA contains A,C,U&G while sequence of protein contain A,R,N,D,C,E,Q,G,H,I,L,K,M,F,P,S,T,W,Y&P these are actually 20 different amino acids in nature which compose a protein.

When both DNA and RNA or mRNA are sequenced in lab their sequences contains larger number of nucleotides with variety

And when we talk about protein its sequences contain large number of bases as they are complex in nature.

- **SOLUTIONS DATABASES**

This large number of sequence or bases cannot be stored in a single computer that's why solution lies in public sequence data bases for DNA & RNA the public database is **GenBank (by NIH).**

For proteins the public database is **UniProt (by Uniprot Consortium)**

Both **GenBank** and **UniProt** are online database and the DNA, RNA and Protein sequences are available here online for public and researchers.

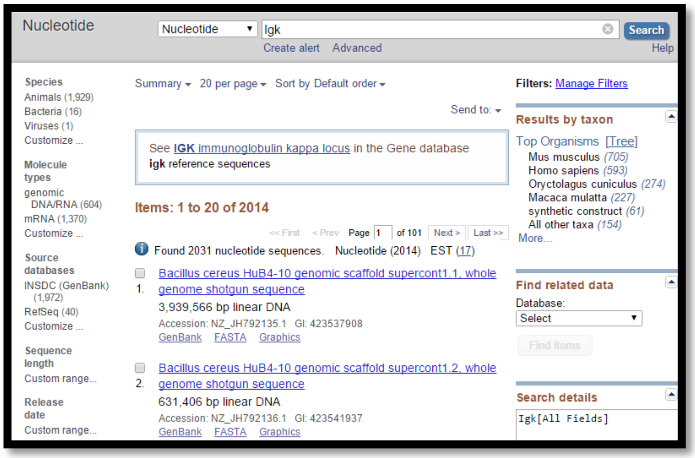
Module 007: USING GENBANK

GenBank is online database where researcher can get access to the sequences of DNA, RNA and proteins.

To find any sequence we go online to NCBI GenBank website which is Public database site. Which is;

www.ncbi.nlm.nih.gov/genbank

And for example we want to find the sequence for Immunoglobulin which is responsible for Glycoprotein antibodies in white blood cells plasma and act for immunity.



Igk

See IGK immunoglobulin kappa locus in the Gene database
igk reference sequences

Items: 1 to 20 of 2014

Page 1 of 101

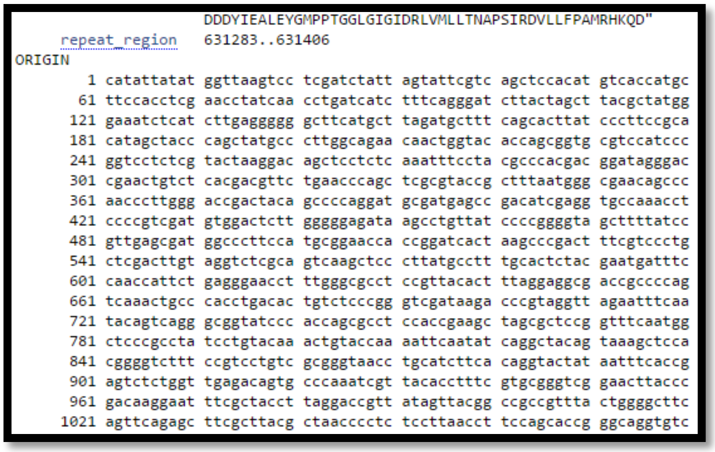
Found 2031 nucleotide sequences. Nucleotide (2014) EST (17)

□ 1. Bacillus cereus HuB4-10 genomic scaffold supercont1.1, whole genome shotgun sequence
3,939,566 bp linear DNA
Accession: NZ_JH792135.1 GI: 423537908
GenBank FASTA Graphics

□ 2. Bacillus cereus HuB4-10 genomic scaffold supercont1.2, whole genome shotgun sequence
631,406 bp linear DNA
Accession: NZ_JH792136.1 GI: 423541937
GenBank FASTA Graphics

Mus musculus	*(705)*
Homo sapiens	*(593)*
Oryctolagus cuniculus	*(274)*
Macaca mulatta	*(227)*
synthetic construct	*(61)*
All other taxa	*(154)*

```
Igk[All Fields]
```



```
                     DDDYIEALEYGMPPTGGLGIGIDRLVMLLTNAPSIRDVLLFPAMRHKQD"
     repeat_region   631283..631406
ORIGIN
        1 catattatat ggttaagtcc tcgatctatt agtattcgtc agctccacat gtcaccatgc
       61 ttccacctcg aacctatcaa cctgatcatc tttcagggat cttactagct tacgctatgg
      121 gaaatctcat cttgaggggg gcttcatgct tagatgcttt cagcacttat cccttccgca
      181 catagctacc cagctatgcc cttggcagaa caactggtac accagcggtg cgtccatccc
      241 ggtcctctcg tactaaggac agctcctctc aaatttccta cgcccacgac ggatagggac
      301 cgaactgtct cacgacgttc tgaacccagc tcgcgtaccg ctttaatggg cgaacagccc
      361 aacccttggg accgactaca gccccaggat gcgatgagcc gacatcgagg tgccaaacct
      421 ccccgtcgat gtggactctt gggggagata agcctgttat ccccggggta gcttttatcc
      481 gttgagcgat ggcccttcca tgcggaacca ccggatcact aagcccgact ttcgtccctg
      541 ctcgacttgt aggtctcgca gtcaagctcc cttatgcctt tgcactctac gaatgatttc
      601 caaccattct gagggaacct ttgggcgcct ccgttacact ttaggaggcg accgccccag
      661 tcaaactgcc cacctgacac tgtctcccgg gtcgataaga cccgtaggtt agaatttcaa
      721 tacagtcagg gcggtatccc accagcgcct ccaccgaagc tagcgctccg gtttcaatgg
      781 ctcccgccta tcctgtacaa actgtaccaa aattcaatat caggctacag taaagctcca
      841 cggggtcttt ccgtcctgtc gcgggtaacc tgcatcttca caggtactat aatttcaccg
      901 agtctctggt tgagacagtg cccaaatcgt tacacctttc gtgcgggtcg gaacttaccc
      961 gacaaggaat ttcgctacct taggaccgtt atagttacgg ccgccgttta ctggggcttc
     1021 agttcagagc ttcgcttacg ctaacccctc tccttaacct tccagcaccg ggcaggtgtc
```

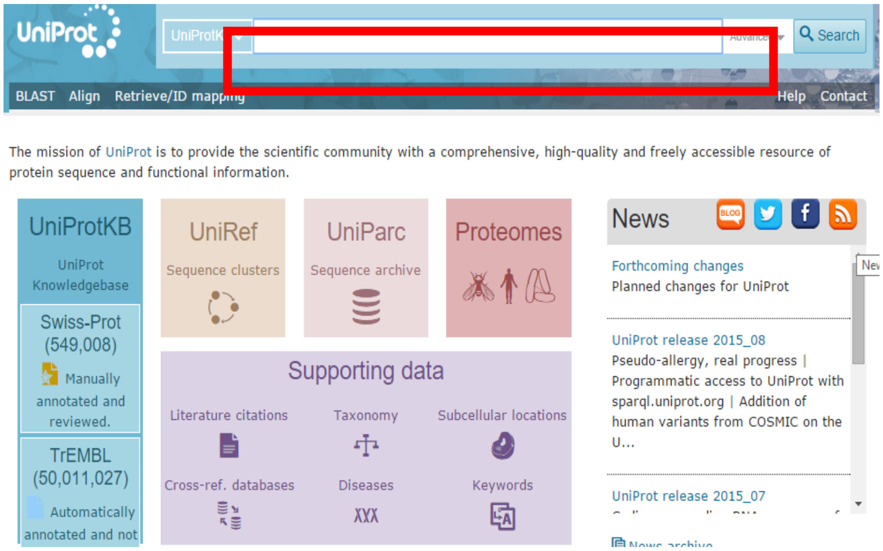
Sequences can be searched from GenBank by typing;

- Sequence name
- ID
- Name
- Species
- Locus
- Accession Number
- Author
- Journal

Module 008: USING UNIPROT

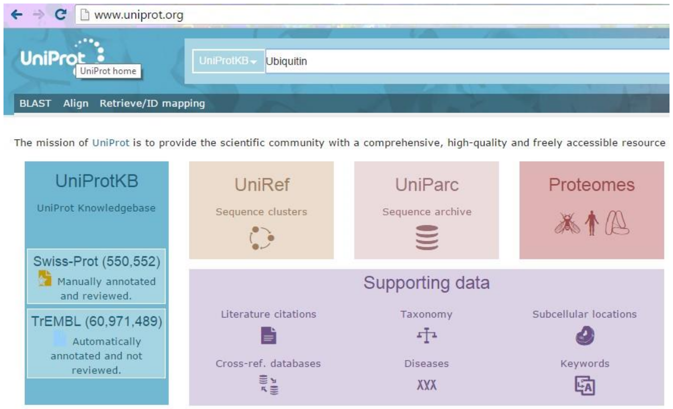
UniProt is public database which is being used to search the sequence of proteins.

www.Uniprot.org



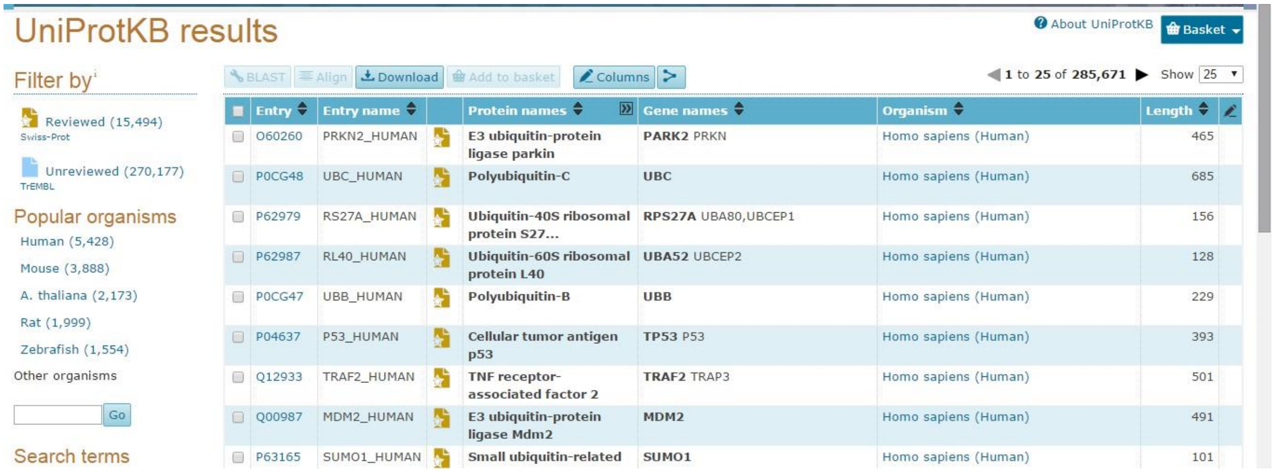
The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

For example we want to search a sequence of a protein which is Ubiquitin which plays an important role in cytosol for recycling the proteins. We have to go online to the website www.Uniprot.org and above page will appear.



The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource

We have to write the name of protein in search box and press enter. You will get the searched results like this one.



UniProtKB results

☐	Entry	Entry name		Protein names	Gene names	Organism	Length
☐	O60260	PRKN2_HUMAN		E3 ubiquitin-protein ligase parkin	**PARK2** PRKN	Homo sapiens (Human)	465
☐	P0CG48	UBC_HUMAN		Polyubiquitin-C	**UBC**	Homo sapiens (Human)	685
☐	P62979	RS27A_HUMAN		Ubiquitin-40S ribosomal protein S27...	**RPS27A** UBA80,UBCEP1	Homo sapiens (Human)	156
☐	P62987	RL40_HUMAN		Ubiquitin-60S ribosomal protein L40	**UBA52** UBCEP2	Homo sapiens (Human)	128
☐	P0CG47	UBB_HUMAN		Polyubiquitin-B	**UBB**	Homo sapiens (Human)	229
☐	P04637	P53_HUMAN		Cellular tumor antigen p53	**TP53** P53	Homo sapiens (Human)	393
☐	Q12933	TRAF2_HUMAN		TNF receptor-associated factor 2	**TRAF2** TRAP3	Homo sapiens (Human)	501
☐	Q00987	MDM2_HUMAN		E3 ubiquitin-protein ligase Mdm2	**MDM2**	Homo sapiens (Human)	491
☐	P63165	SUMO1_HUMAN		Small ubiquitin-related	**SUMO1**	Homo sapiens (Human)	101

By clicking on any result you can **download** or **Blast** the sequence.

In home page there is a box named **"Swiss Prot"** which contains human curated protein information, molecular mass, observed and predicted modifications etc.

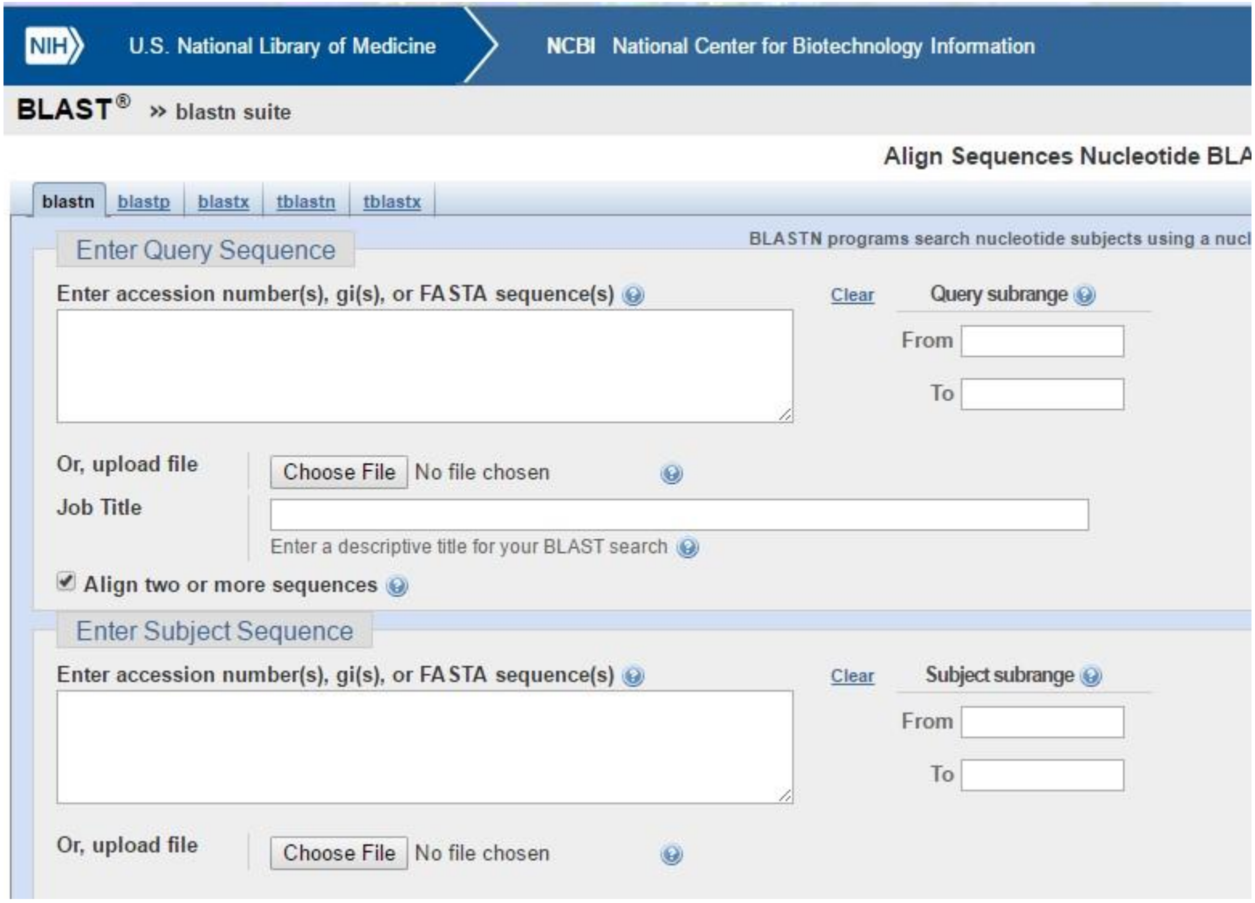
Uniprot can be searched by typing amino acid, Name, ID or sequence.

Module 009: COMPARING SEQUENCES

There are millions sequences on GenBank and UniProt what will happen if we will compare them? By comparing sequences of DNA, RNA and Proteins we can get

- Similarity among sequences
- There might be some specific difference due to some disease or mutation
- There may be some evolutionary relationship.

As there nucleotides can be similar or differ from each other



BLASTN programs search nucleotide subjects using a nucl

Enter accession number(s), gi(s), or FASTA sequence(s)

Or, upload file

Job Title

Enter a descriptive title for your BLAST search

☑ Align two or more sequences

Enter accession number(s), gi(s), or FASTA sequence(s)

Or, upload file

Figure 0.6 BLAST is used to compare the nucleotides sequences

While UniProt is used in case of amino acids sequence comparison.

By comparison of nucleotides and Amino acids of any DNA, RNA and protein sequence we can find many evolutionary facts and relations among species.

Module 010: SIMILARITIES & DIFFERENCES IN SEQUENCES

When we compare the sequences of DNA and RNA we can get the similarity and differences or relationship in evolution. And same case is with amino acids of proteins.

In compression not only they have the same number of nucleotides but they have same order or arrangements.

If some sequence are exactly similar to each other it means;

- They might have some regular expression in cell or system.
- Or they indicate some specific presence like signature of any protein or gene.
- Or they might have similar nucleotide just one or two between them are different from rest.

- **CONCLUSION**

If there is exact match in sequences it means their order or arrangement and maximum numbers of nucleotides match to each other not all of those.

While the genome of each created kind is unique, many animal kinds share some specific types of genes that are generally similar in DNA sequence. When comparing DNA sequences between animal taxa, evolutionary scientists often hand-select the genes that are commonly shared and more similar (conserved), while giving less attention to categories of DNA sequence that are dissimilar. One result of this approach is that comparing the more conserved sequences allows the scientists to include more animal taxa in their analysis, giving a broader data set so they can propose a larger evolutionary tree.

Although these types of genes can be easily aligned and compared, the overall approach is biased towards evolution. It also avoids the majority of genes and sequences that would give a better understanding of DNA similarity concepts.

http://www.icr.org/article/common-dna-sequences-evidence-evolution/

Module 011: SIMILARITIES & DIFFERENCES IN SEQUENCES

When we compare the sequences of DNA and RNA we can get the similarity and differences or relationship in evolution. And same case is with amino acids of proteins.

In compression not only they have the same number of nucleotides but they have same order or arrangements.

If some sequence are exactly similar to each other it means;

- They might have some regular expression in cell or system.
- Or they indicate some specific presence like signature of any protein or gene.
- Or they might have similar nucleotide just one or two between them are different from rest.

- **CONCLUSION**

If there is exact match in sequences it means their order or arrangement and maximum numbers of nucleotides match to each other not all of those.

While the genome of each created kind is unique, many animal kinds share some specific types of genes that are generally similar in DNA sequence. When comparing DNA sequences between animal taxa, evolutionary scientists often hand-select the genes that are commonly shared and more similar (conserved), while giving less attention to categories of DNA sequence that are dissimilar. One result of this approach is that comparing the more conserved sequences allows the scientists to include more animal taxa in their analysis, giving a broader data set so they can propose a larger evolutionary tree.

Although these types of genes can be easily aligned and compared, the overall approach is biased towards evolution. It also avoids the majority of genes and sequences that would give a better understanding of DNA similarity concepts.

http://www.icr.org/article/common-dna-sequences-evidence-evolution/

Module 012: PAIR WISE ALIGNMENT –II

In pair wise alignment of nucleotides the nucleotides comes in pairs and matching are colored while missing amino acids are indicated with “” and this empty space is called as gap.

And these Gaps are inserted for deletion or insertion of any nucleotide. Increase in Gaps can increase the chance of plenty in sequencing and less number of Gaps can increase the similarity rate of sequences.

There are two types of pair alignments.

1. **Global**
2. **Local**

In Global ways of sequence pair alignment we introduce the Gaps in all sequence to know over all matching. While in Local type of sequence pair alignment we find those regions where nucleotides are maximum matching with each other it is used to find the similarity or some nutation.

Most important the Gaps are introduced so that we may add the missing nucleotides.

Pairwise Sequence Alignment is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequences (protein or nucleic acid).

http://www.ebi.ac.uk/Tools/psa/

Module 013: PAIR WISE SEQUENCE ALIGNMENT –III

Pair wise alignment helps us to find the similarity and differences there are three ways according to which sequences can differ from each other.

Which are

- Substitutions ACGA → AGGA
- Insertions ACGA → ACCGA
- Deletions ACGA → AGA

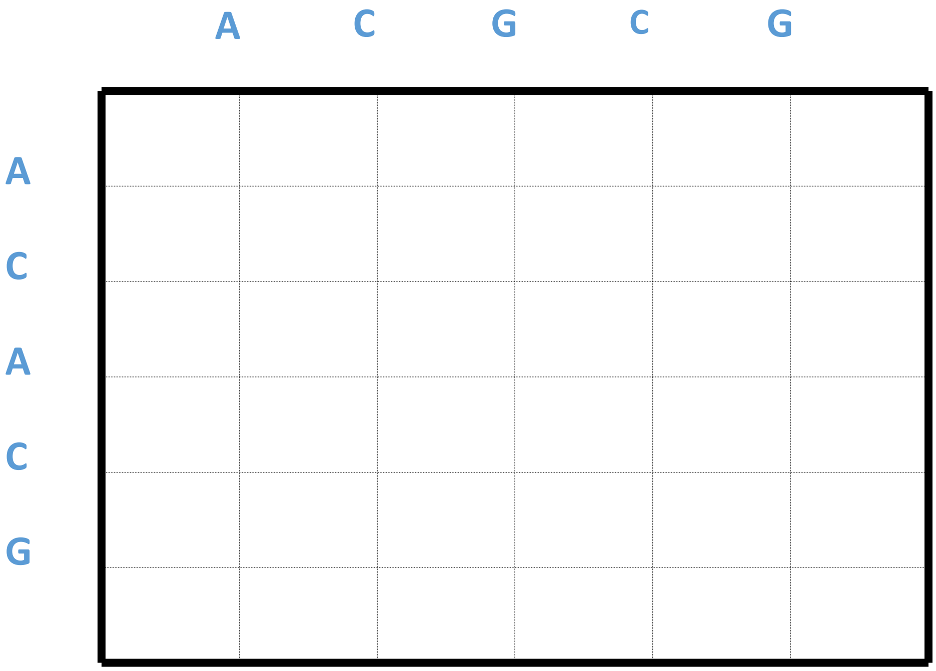
By applying all above ways to any sequence the matching and mismatching can be increased or decreased between to different comparing sequencing.

Both local and Global ways of alignments give us different results.

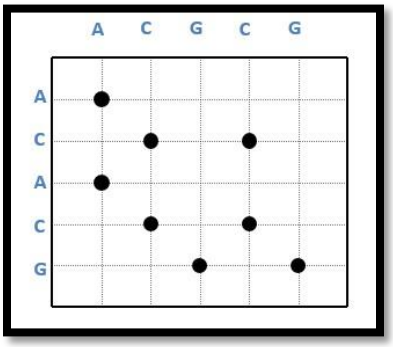
But among above Substitution increases mismatch of sequence.

Module 014: DOT PLOTS

To visualize the sequence alignment we have a method called Dot Plots in this method the sequence is written top and left side of dot matrix grid.



Where one nucleotide or amino acid match with each other the dot is placed in grid position in each row for one time.



	A	C	G	C	G
A	●				
C		●		●	
A	●				
C		●		●	
G			●		●

Similar dots are match with diagonal pattern and which remain separate differ from similar sequence

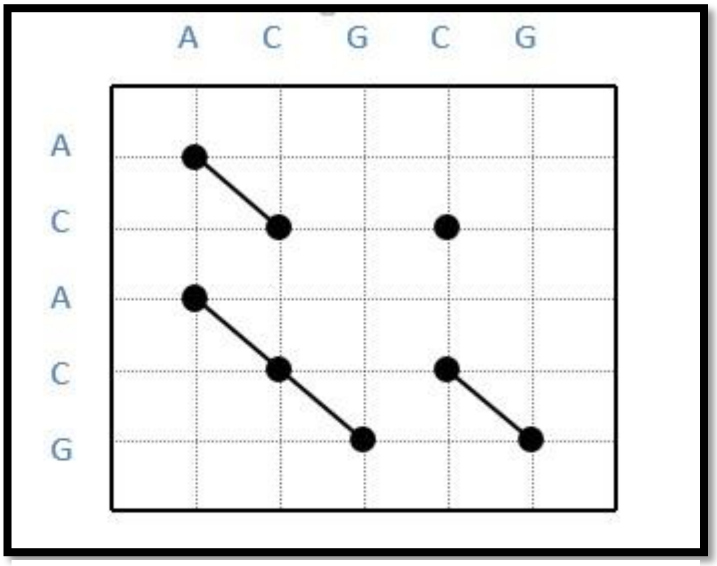


Figure 0.7 dot plot diagonal pattern

Dots on diagonal repeats the alignments and separate one give difference to the sequence.

Module 015: EXEMPLE OF DOT PLOTS

In dot plot the matching nucleotides are connected in diagonal way and represent the sequence alignments.

When we compare the **human Cytochrome** and **Tuna Fish Cytochrome** than the diagonal alignment of sequence we find is in this below diagram.

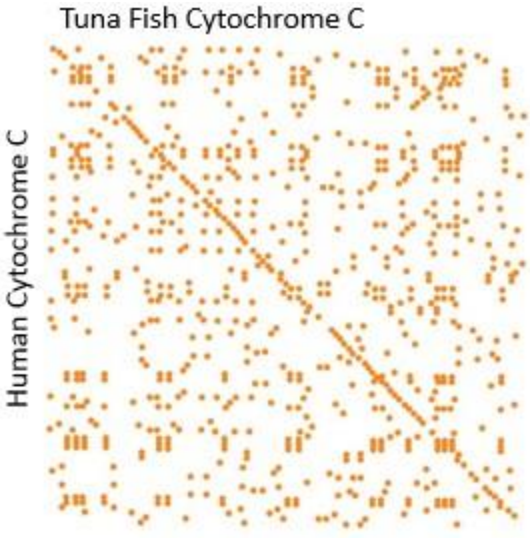


Figure 0.8 tuna fish vs Human

- **BENEFITS**

Dot plots provides us the Global similarity between the two sequences and helps us to visualize the alignments of sequences and sequence repeats appear as diagonal stacks in plot.

- **CONCLUSION**

Dot plot help us to find the threshold difference among two sequences.

Module 016: IDENTY VS SIMILARITY

When we talk about the comparison of two sequences than question arise that how we can compare the biological sequences and after comparison what will be the degree of comparison.

There are two concepts for sequence analysis

1. **Identity**
2. **Similarity**

Identity means the counting number of nucleotides or amino acids which exactly match when two biological sequences are matched.

For example:

1: CATGCTT
2: CATGC

Number of match = 5

Smaller length = 5

Sequence (1) = 7

Sequence (2) = 5

Formula for Identity:

Identity = No. of Matches / smaller length × 100

$$=\frac{5}{5} \times 100 = 100\%$$

And **Similarity** means the comparison between two different sequences calculated by alignment approach.

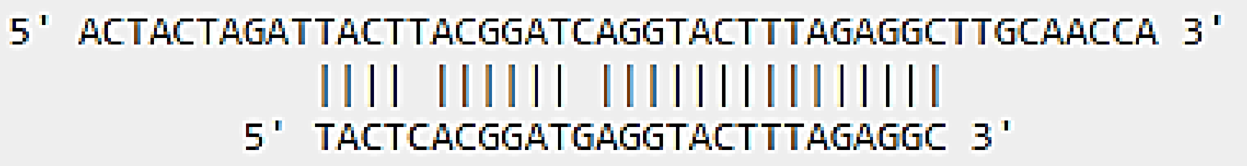
In both identity and similarity the dots are not counted.

Module 017: INTRODUCTION TO ALIGNMENT APPROACHES

When we align the sequence that may be vary due to insertion and deletion of nucleotides and to calculate the similarity we need to align the sequence first. And there are two different approaches to align the sequence.

1. **Global Alignment**
2. **Local Alignment**

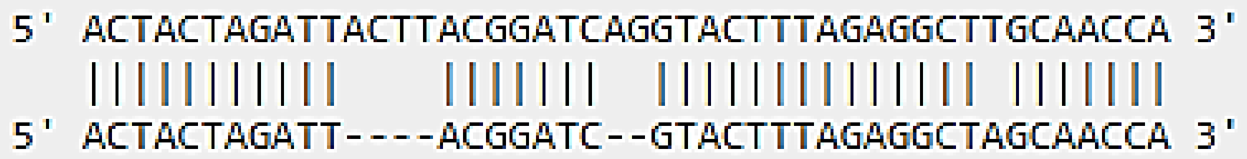
In **Local alignment** we compare one whole sequence with the one portion of other like this.



```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
             |||| |||||| |||||||||||||||
          5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Figure 0.9 local alignment

While in **Global alignment** we compare both sequence from end to end completely.



```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

Figure 0.10 Global alignment

Local alignment just focus on highly matching portions of sequence while in Global one whole sequence is compared with other one.

Module 018: WHY LOCAL ALIGNMENT

When there is Global alignment which compare the whole sequence from end to end than why local alignment is done question arise.

Because Local alignment have power to detect the smaller regions with high similarity and such matches are motifs or domains which remain hidden in case of protein function.

- **DOMAIN SHUFFLING**

Aligned portions of sequence can be considered in varying orders and this process is called as domain shuffling.

- **ADVENTAGES**

- We can compare the different length sequences
- Conserved domains can be determined from proteins
- Common function features can be identified.

- **CONCLUSION**

Local alignment is used to compare the segments for high matching sequencing.

Module 019: ALIGNING, INSERTION & DELETION

Insertion means addition of amino acids in protein sequence and addition of nucleotides in DNA sequences.

And deletion means removal of amino acids from protein sequence and removal of nucleotides from DNA or RNA sequences.

- **ALIGNING INSERTION**

For example we have following two sequences

1: A C T G A C T G

2: A C G A C T G

1: A C T G A C T G

2: A C G A C T G

To add the nucleotide in sequence 2 we will add gap first. And same happens with the deletion alignment we add gap where we delete the nucleotide from sequence. And such insertion of gap is called as –ve or plenty.

1: A C T G A C T G

2: A C . G A C T

Module 020: ALIGNING MUTATION IN SEQUENCES

Removal and addition of amino acids in proteins and nucleotides in DNA, RNA by using Gaps named as Indels.

Mutation is totally different from Indels, because in Mutation we replace the amino acid with other amino acids and replace the nucleotides with other and we don’t use Gaps is inserted in template or target for mutation.

1: A C T G A C T

2: A C G G A C T

1: A C T G A C

2: A C G G A C

- **CONCLUSION**

In identity alignment we use Gaps and in mutation we use substitution penalties and penalties depend upon the substitution.

Module 021: INTRODUCTION TO DYNAMIC PROGRAMMING

To find matching in nucleotides and amino acids of two sequences we use dot plot method. But dot plot cannot capture the insertions, deletions and gaps in the sequences.

To deal with this situation we modify the dot plot.

We represent the matching nucleotides with +1 while gaps, substitutions, insertions and mutations can be represented as -1 in dot plot.

Dynamic programming is an algorithmic technique used commonly in sequence analysis. Dynamic programming is used when recursion could be used but would be inefficient because it would repeatedly solve the same sub problems.

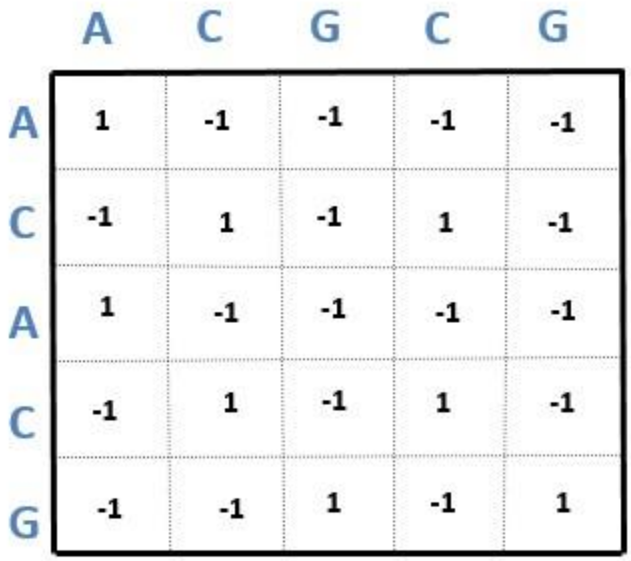
http://www.ibm.com/developerworks/library/j-seqalign/

Module 022: DYNAMIC PROGRAMMING ESSENTIALS

When we talk about the compression of two sequences one by one it need time and is computationally expensive method. That's why we need algorithm.

In algorithm we calculate the step involve in sequence compression for example if we if we compare two sequences of length **"n"** than it would be **"n^2 "**

And its order is **O (n^2)**



	A	C	G	C	G
A	1	-1	-1	-1	-1
C	-1	1	-1	1	-1
A	1	-1	-1	-1	-1
C	-1	1	-1	1	-1
G	-1	-1	1	-1	1

Figure 0.11 -1 represent deletion, insertion and gaps while +1 represent matching nucleotides or amino acids

One by one sequence compression is costly and time consuming process we minimize the cost with the help of algorithm.

Module 023: DYNAMIC PROGRAMMING METHODOLOGY

Dynamic programming helps us to reduce the computational cost in sequence comparisons and it works on the method of “scoring function”.

For example

Match = +a

Mismatch = -b

Gap = -c

Score = #match + #Mismatch +#Gaps

Match Rewards = 10
Mismatch penalty = 2
Gap penalty = 5

C	T	G	T	C	G	–	C	T	G	C
-	T	G	C	–	C	G	–	T	G	-
-5	10	10	-2	-5	-2	-5	-5	10	10	-5

Total =11

All the alignments are done in diagonal way in dot plot matrix. For total score we make calculations in diagonal way and after calculation best one is selected.

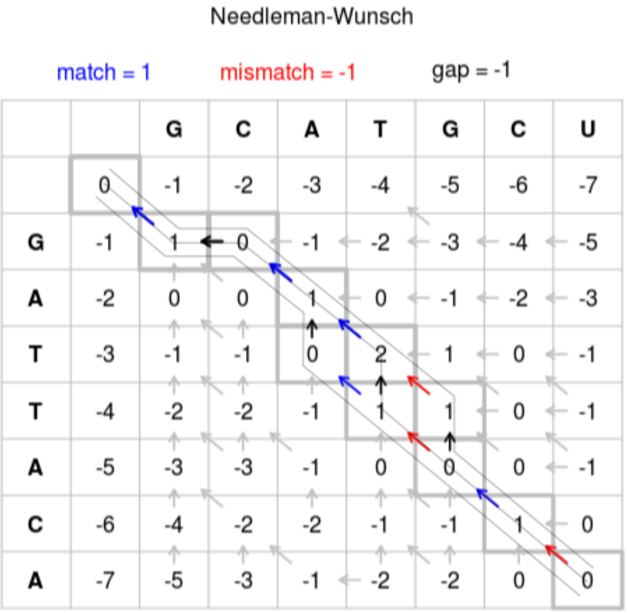
Module 024: NEEDLEMAN WUNSCH ALGORITHM-I

In two different sequences alignments are arranged in a diagonal pattern of dot matrix. Total scores are captured for each alignment and at the end the best one is selected.

Needleman Wunsch Algorithm is the way for alignment. The method is same like dot plot but it computes the scores in different way. We Start with a zero in the second row, second column. Move through the cells row by row, calculating the score for each cell.

The score is calculated as the best possible score (i.e. highest) from existing scores to the left, top or top-left (diagonal). When a score is calculated from the top, or from the left this represents an indel in our alignment. When we calculate scores from the diagonal this represents the alignment of the two letters the resulting cell matches to. Given there is no 'top' or 'top-left' cells for the second row we can only add from the existing cell to the left. Hence we add -1 for each shift to the right as this represents an indel from the previous score. This results in the first row being 0, -1, -2, -3, -4, -5, -6, and -7. The same applies to the second column as we only have existing scores above. Thus we have:

https://en.wikipedia.org/wiki/Needleman%E2%80%93Wunsch_algorithm



		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1							
A	-2							
T	-3							
T	-4							
A	-5							
C	-6							
A	-7							

Figure 11 Needleman-Wunsch pairwise sequence alignment

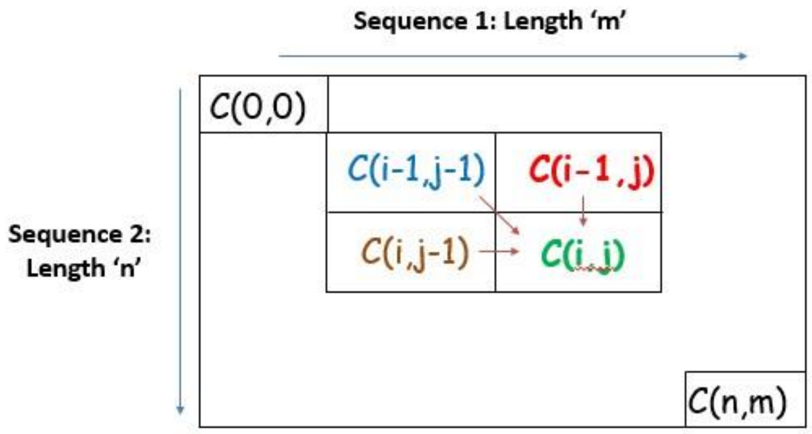


Figure 12 Needleman wunsch algorithm way of computation of nucleotides

Module 025: NEEDLEMAN WUNSCH ALGORITHM-II

Alignments are represented by unbroken diagonal dot matrix plot. In this way we can create numerous combinations.

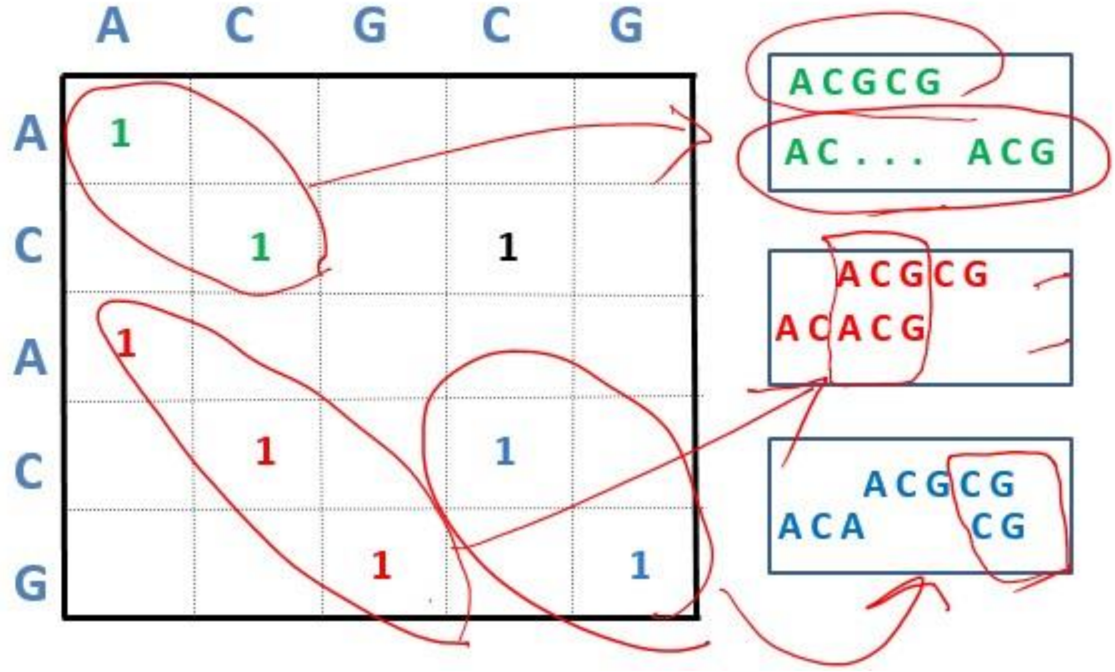
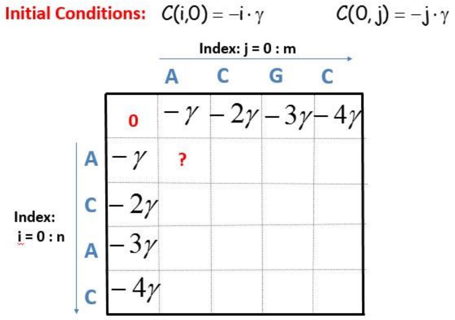


Figure 13 various combinations of sequences through dot plot

If the sequence is too long then there will be many diagonal alignments and at the end we select the best alignment by combinations of all. And for this we use Needleman Algorithm

In Needleman Algorithm we use 0, 0 in first row and first column.



Initial Conditions: $C(i,0) = -i\cdot\gamma$ $C(0,j) = -j\cdot\gamma$

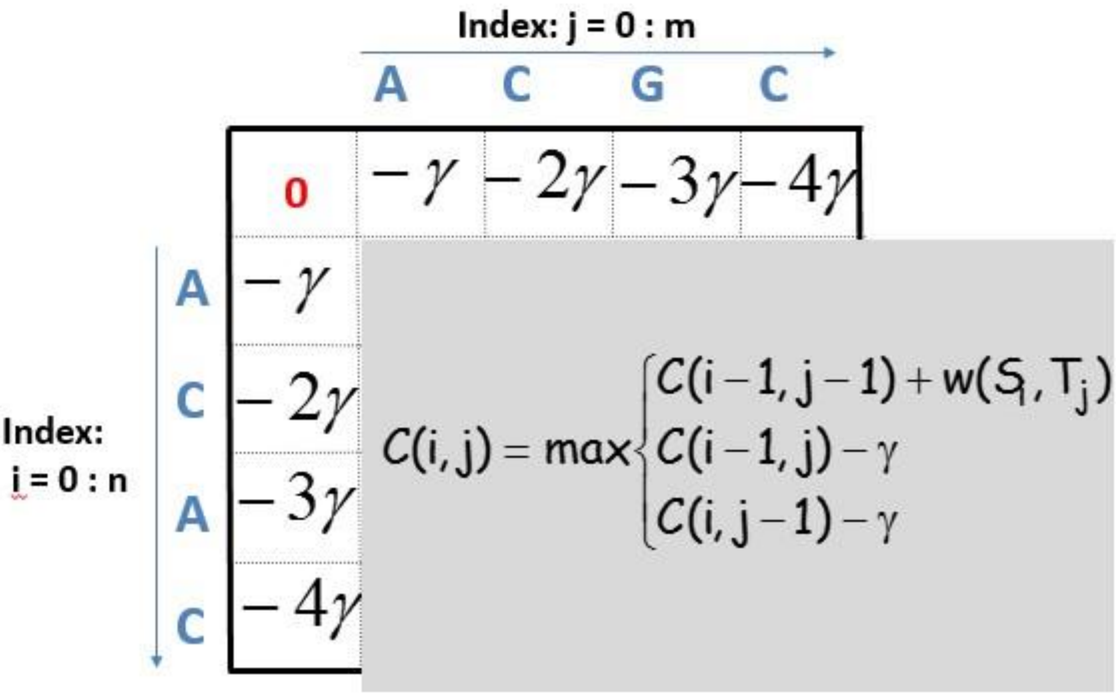
Index: j = 0 : m

Index: i = 0 : n

$$\begin{array}{c|ccccc}
 & & A & C & G & C \\
\hline
 & 0 & -\gamma & -2\gamma & -3\gamma & -4\gamma \\
A & -\gamma & ? & & & \\
C & -2\gamma & & & & \\
A & -3\gamma & & & & \\
C & -4\gamma & & & &
\end{array}$$

Figure 14 initial column and row are kept zero (0)

Left to right and top to bottom the best element (having high score) is selected.



$$C(i,j) = \max \begin{cases} C(i-1, j-1) + w(S_i, T_j) \\ C(i-1, j) - \gamma \\ C(i, j-1) - \gamma \end{cases}$$

Figure 15 maximum score element is selected from all three sides comparison

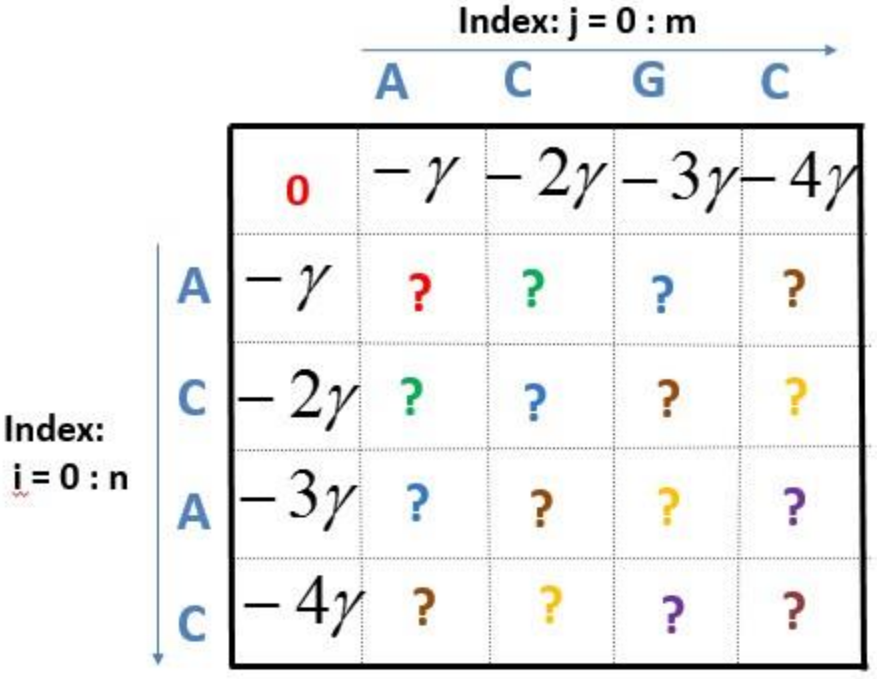
The terms for match, mismatch are:

α Alpha = Match reward
β Beta = Mismatch penalty
γ Gamma = Gap penalty

The matrix is computed progressively until the bottom right element

Module 026: NEEDLEMAN WUNSCH ALGORITHM-III

We follow the diagonal route for scoring in Needleman Wuncsh Algorithm. Left to right and top to bottom in diagonal way we select the highest score.



		A	C	G	C
	0	$-\gamma$	-2γ	-3γ	-4γ
A	$-\gamma$	?	?	?	?
C	-2γ	?	?	?	?
A	-3γ	?	?	?	?
C	-4γ	?	?	?	?

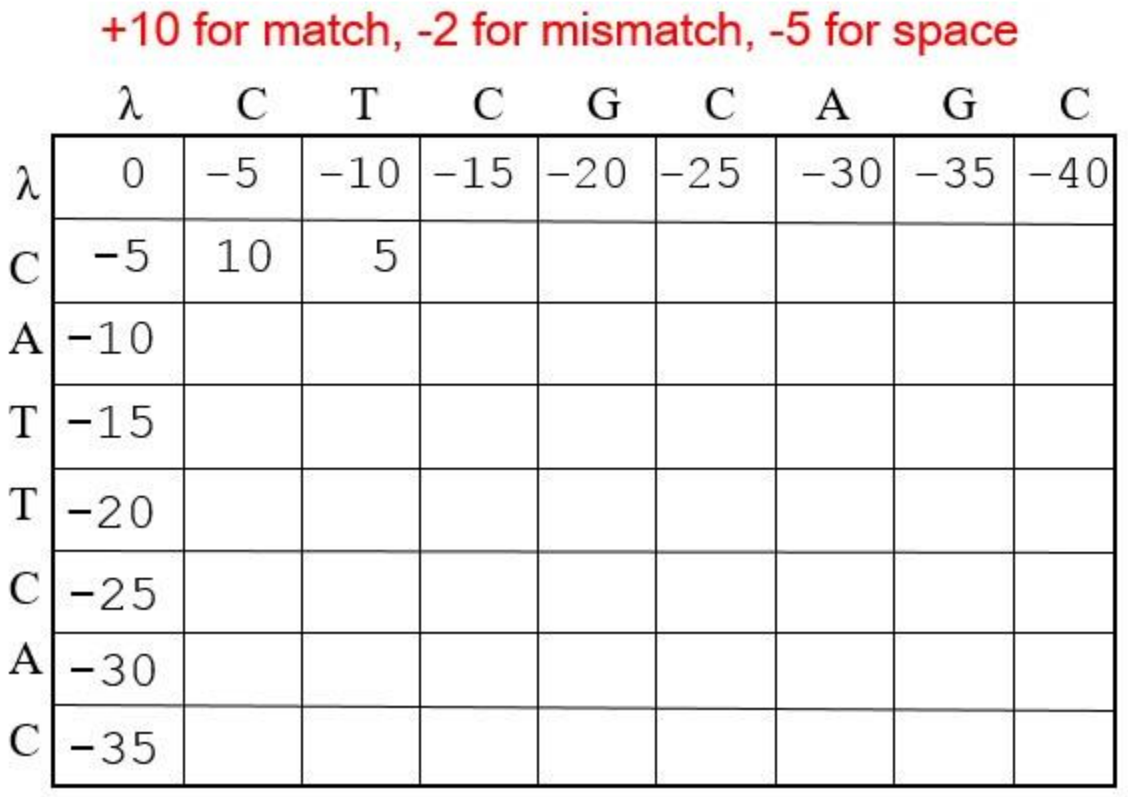
Figure 16 filling order of sequence alignment

We select the best score to make best alignment and matrix is computed progressively until we reach to the bottom right.

Module 027: NEEDLEMAN WUNSCH ALGORITHM EXAMPLE

Top left and diagonal element are considered to calculate an element in the matrix. Match, mismatch and gap penalty is computed from all there sides (Left to right) (Top to bottom) and (Diagonal).

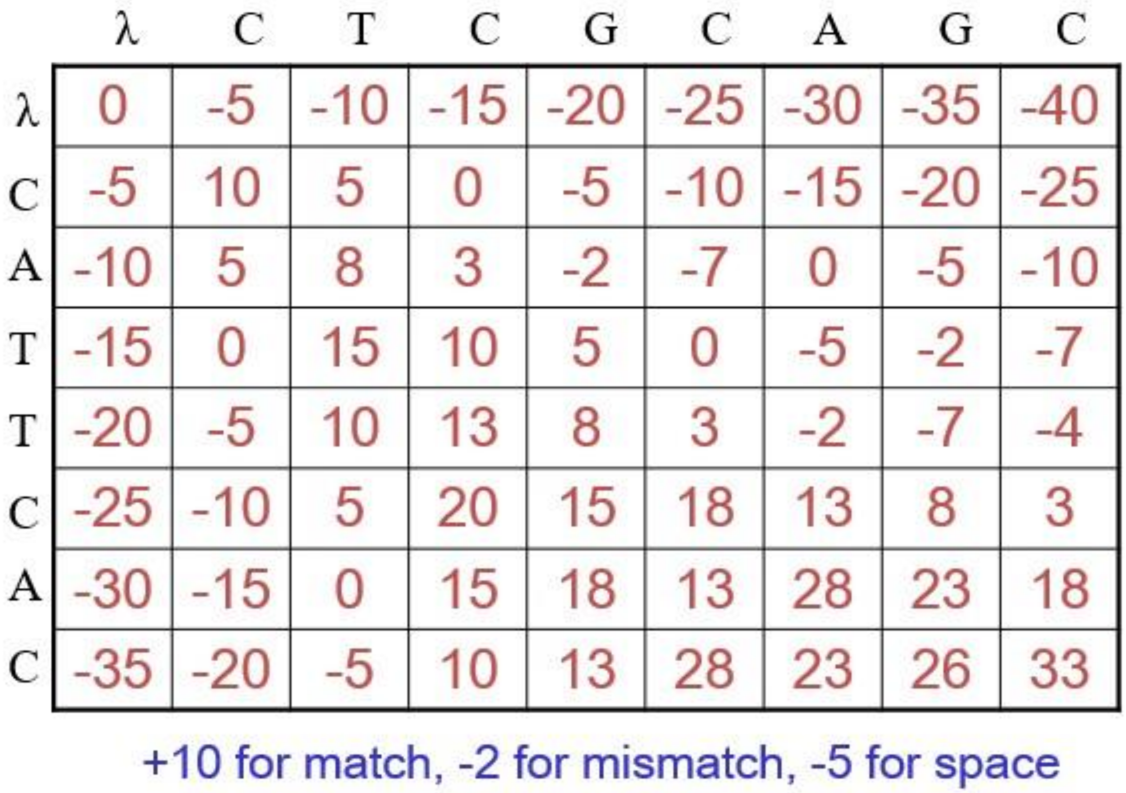
For example:



+10 for match, -2 for mismatch, -5 for space

	λ	C	T	C	G	C	A	G	C
λ	0	−5	−10	−15	−20	−25	−30	−35	−40
C	−5	10	5						
A	−10								
T	−15								
T	−20								
C	−25								
A	−30								
C	−35								

Figure 17 score computation from all sides



	λ	C	T	C	G	C	A	G	C
λ	0	-5	-10	-15	-20	-25	-30	-35	-40
C	-5	10	5	0	-5	-10	-15	-20	-25
A	-10	5	8	3	-2	-7	0	-5	-10
T	-15	0	15	10	5	0	-5	-2	-7
T	-20	-5	10	13	8	3	-2	-7	-4
C	-25	-10	5	20	15	18	13	8	3
A	-30	-15	0	15	18	13	28	23	18
C	-35	-20	-5	10	13	28	23	26	33

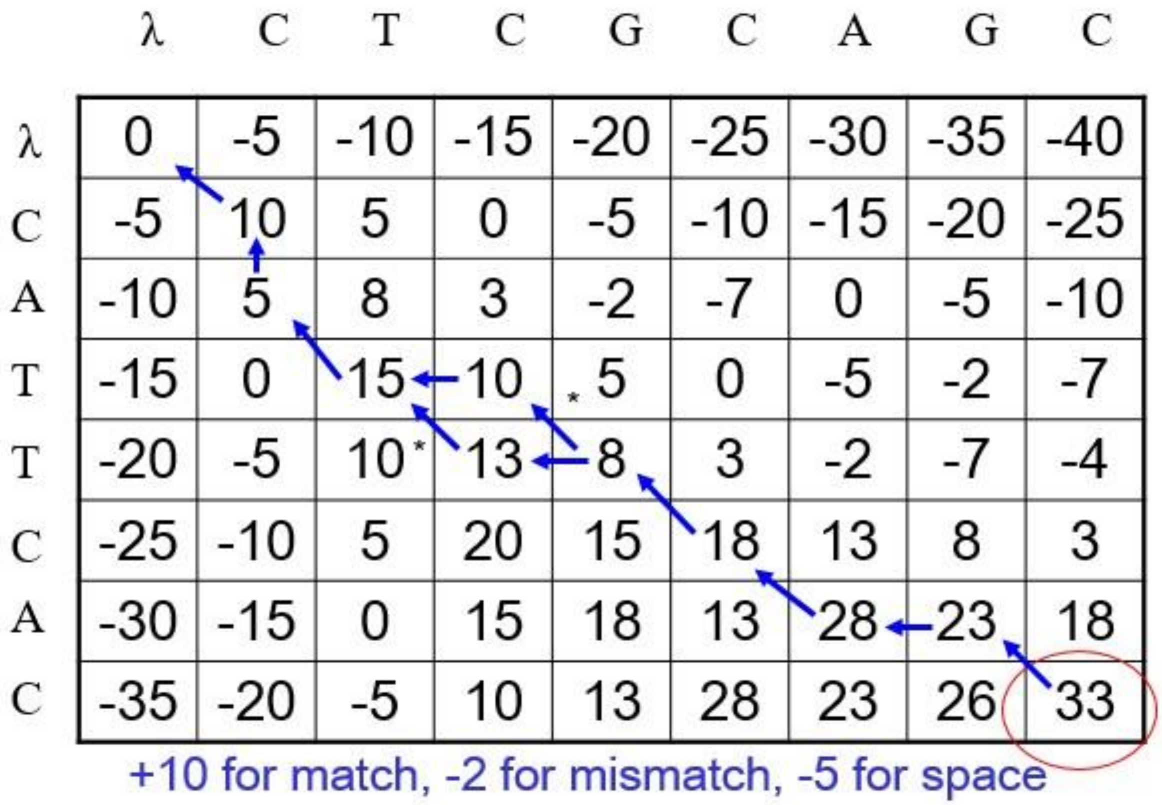
+10 for match, -2 for mismatch, -5 for space

Figure 18 the best score is computed in diagonal way

DNA, RNA and Protein sequences can be computed by using Needleman algorithm.

Module 028: BACKTRACKING ALIGNMENTS

To find an optimal alignment in Needleman Wunsch Algorithm we use traceback method.



	λ	C	T	C	G	C	A	G	C
λ	0	-5	-10	-15	-20	-25	-30	-35	-40
C	-5	10	5	0	-5	-10	-15	-20	-25
A	-10	5	8	3	-2	-7	0	-5	-10
T	-15	0	15	10	* 5	0	-5	-2	-7
T	-20	-5	10*	13	8	3	-2	-7	-4
C	-25	-10	5	20	15	18	13	8	3
A	-30	-15	0	15	18	13	28	23	18
C	-35	-20	-5	10	13	28	23	26	33

+10 for match, -2 for mismatch, -5 for space

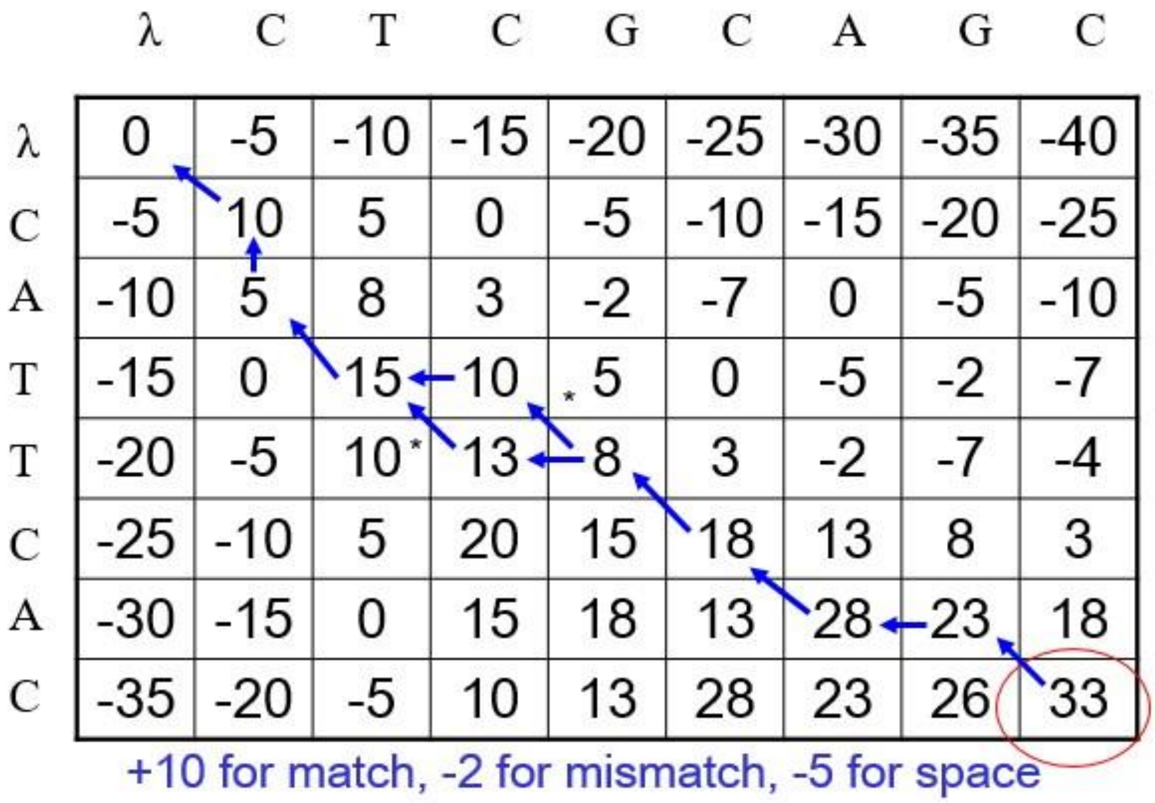
Figure 19 traceback method starts from end maximum score.

After completely matrix calculations we apply traceback to find the optimal alignment and traceback starts from bottom right (maximum score) to top side.

Module 029: REVISITING LOCAL AND GLOBAL ALIGNMENTS

We use Needleman Algorithm to align the sequences in scoring way and traceback method to find the optimal alignment.

In Needleman Algorithm we start traceback from bottom right towards top left progressively and this provides us the global alignment.



	λ	C	T	C	G	C	A	G	C
λ	0	-5	-10	-15	-20	-25	-30	-35	-40
C	-5	10	5	0	-5	-10	-15	-20	-25
A	-10	5	8	3	-2	-7	0	-5	-10
T	-15	0	15	10	* 5	0	-5	-2	-7
T	-20	-5	10*	13	8	3	-2	-7	-4
C	-25	-10	5	20	15	18	13	8	3
A	-30	-15	0	15	18	13	28	23	18
C	-35	-20	-5	10	13	28	23	26	33

+10 for match, -2 for mismatch, -5 for space

Figure 20 Traceback in Needleman Algorithm

We can start traceback from any point in matrix and smith waterman algorithm helps elicit the local alignments.

Module 030: OVERLAP MATCHES

Dot plot and Needleman wucsch are algorithm method with little difference. Dot plot help us in finding matching residues of two sequences while Needleman wunsch helps us to find the global alignments.

If some sequences have different regions of nucleotides which does not match to any other for that alignment we prefer Global alignment not local, but that does not penalize leading or trailing end.

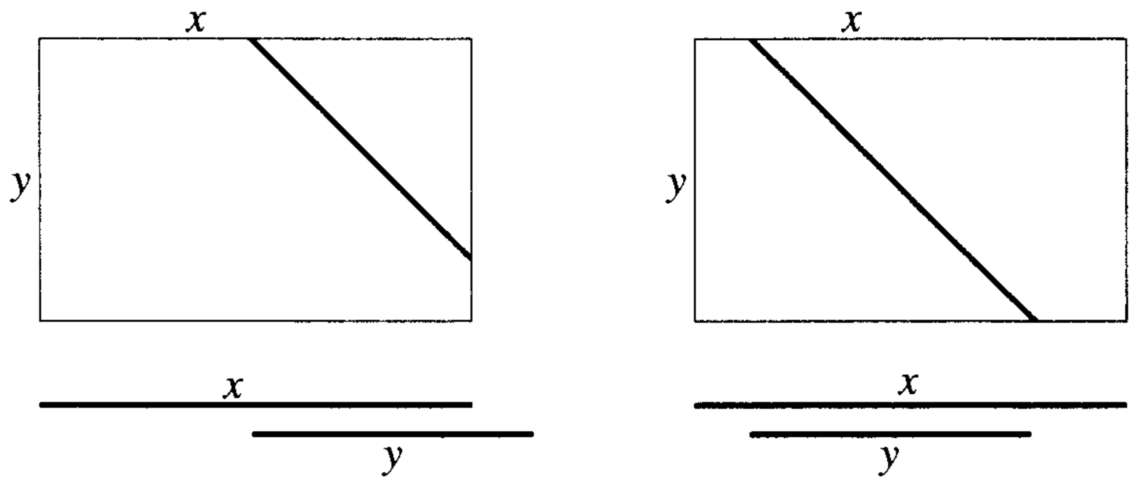
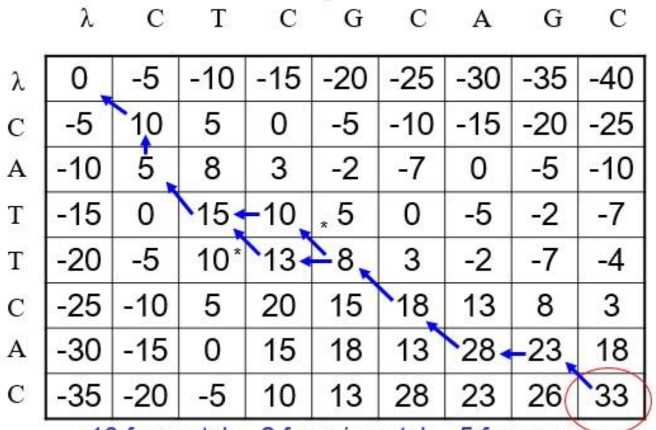


Figure 21 leading and trailing edge mismatches versus global alignment by gap-insertion (stretching) of sequences

And "Traceback" is the technique by which we can check the sequences from any end of the matrix box. And such "Tracebacks" helps us to find the overlaps in aligned sequences.



	λ	C	T	C	G	C	A	G	C
λ	0	-5	-10	-15	-20	-25	-30	-35	-40
C	-5	10	5	0	-5	-10	-15	-20	-25
A	-10	5	8	3	-2	-7	0	-5	-10
T	-15	0	15	10	*5	0	-5	-2	-7
T	-20	-5	10*	13	8	3	-2	-7	-4
C	-25	-10	5	20	15	18	13	8	3
A	-30	-15	0	15	18	13	28	23	18
C	-35	-20	-5	10	13	28	23	26	33

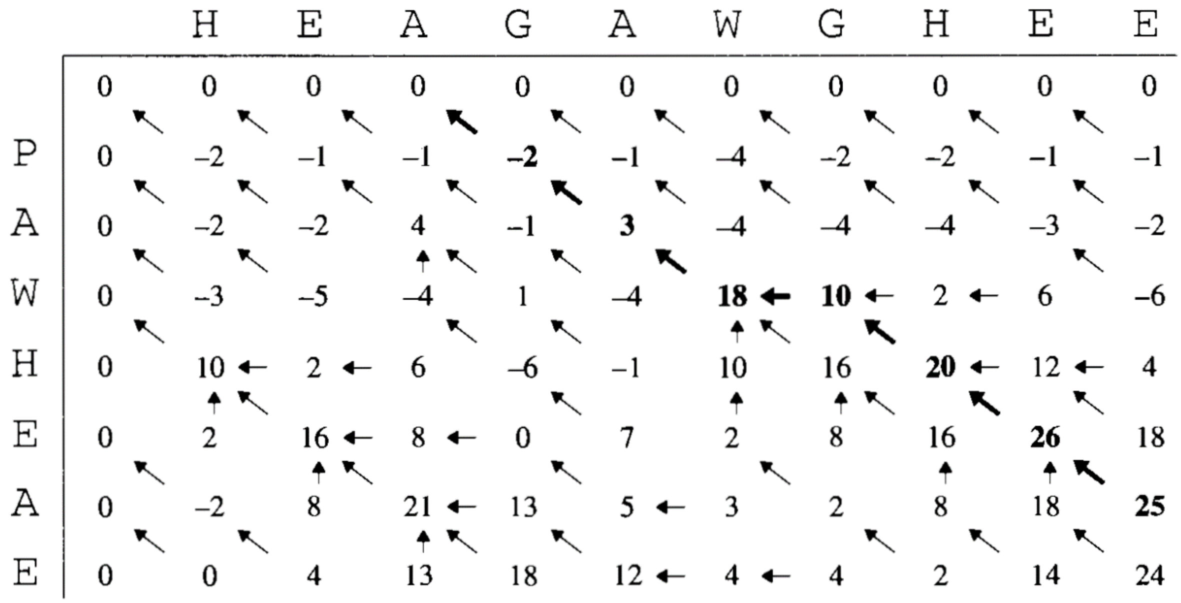
+10 for match, -2 for mismatch, -5 for space

Figure 22 Traceback method

Module 031: EXAMPLE

A slight variation in traceback can helps us to find the overlaps in sequences and can apply some interesting strategies in sequences alignments.

In following example of amino acids alignment we can understand the ways of tracback.



		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	–2	–1	–1	**–2**	–1	–4	–2	–2	–1	–1
A	0	–2	–2	4	–1	**3**	–4	–4	–4	–3	–2
W	0	–3	–5	–4	1	–4	**18** ←	**10** ←	2 ←	6	–6
H	0	10 ←	2 ←	6	–6	–1	10	16	**20** ←	12 ←	4
E	0	2	16 ←	8 ←	0	7	2	8	16	**26**	18
A	0	–2	8	21 ←	13	5 ←	3	2	8	18	**25**
E	0	0	4	13	18	12 ←	4 ←	4	2	14	24

Figure 23 Traceback in amino acid sequence alignment

Scoring stagey is:

Match = +2. Mismatch = -1, Gap = -2

Sequences are:

GAWGHEE
PAW-HEA

Module 032: MOVING FROM GLOBAL TO LOCAL ALIGNMENT

DNA has coding and noncoding regions. Coding regions are called “EXON” expressed as protein and they remain more conserved due to their role in making functional proteins.

And noncoding regions of DNA are called as “INTRONS” which are more likely involved in mutations than coding ones. It means high degree of alignment can be find among two exons.

In local alignment we use small segments of sequences and through which we can find exons. Through this we can find “functional subunits”.

However, the term *exon* is often misused to refer only to coding sequences for the final protein. This is incorrect, since many noncoding exons are known in human genes.

(Zhang 1998)

Zhang, M. Q. (1998). "Statistical features of human exons and their flanking regions." Human molecular genetics **7**(5): 919-932.

Module 033: SMITH WATERMAN ALGORITHM

In global alignment we compare the sequence from end to end but in local alignment we compare the sequences in segments.

For Global sequences we use Needleman and Wunsch algorithm while for local pairwise alignment we use Smith and waterman.

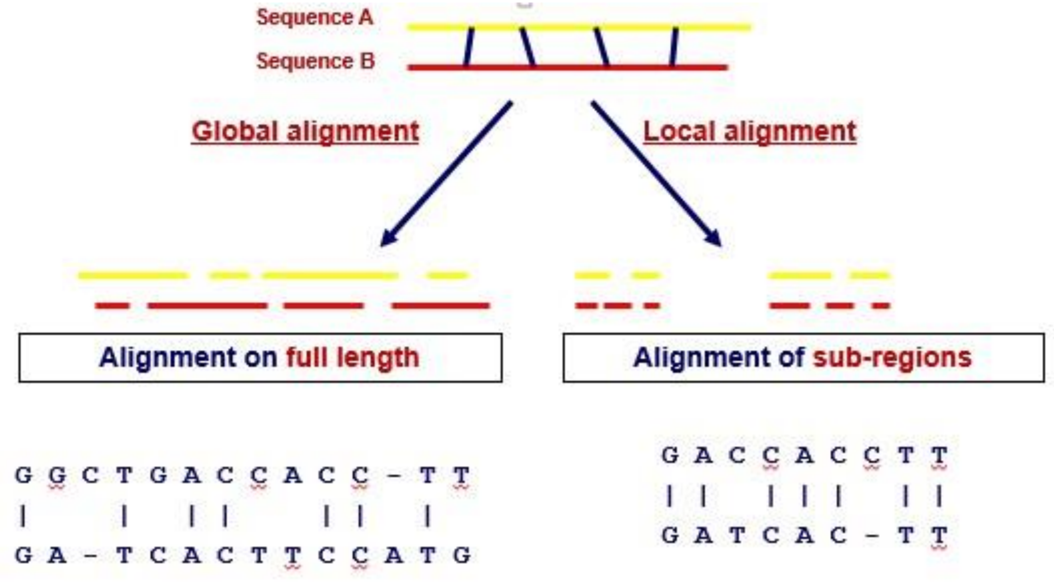


Figure 24 Global and local sequence alignment comparison.

The Smith Freshman algorithm is different from Needle man.

- Top row and Colum are set to zero.
- Alignment can end anywhere.
- Traceback starts from highest score.

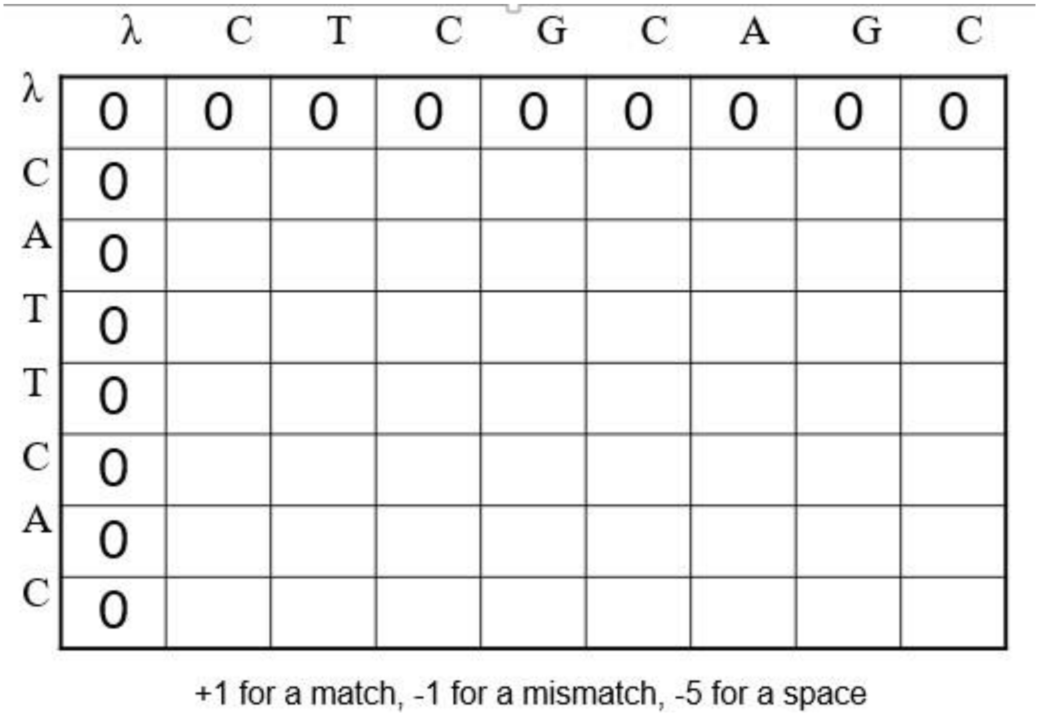
Local alignments can identify the coding portions of DNA and in this way we can find the functional domains from protein sequences.

Module 034: EXAMPLE OF SMITH WATERMAN ALGORITHM

The only difference between Needleman and Smith Waterman is that zero “0” is placed in the relationship.

$$C[i, j] = \max \begin{cases} C[i-1, j-1] + score(i, j) \\ C[i-1, j] - \gamma \\ C[i, j-1] - \gamma \\ 0 \end{cases}$$

And in the matrix we place top line of zero and first Colum of zero.



	λ	C	T	C	G	C	A	G	C
λ	0	0	0	0	0	0	0	0	0
C	0								
A	0								
T	0								
T	0								
C	0								
A	0								
C	0								

+1 for a match, -1 for a mismatch, -5 for a space

Figure 25 top line and first Colum are filled with zero in Smith Freshman Algorithm

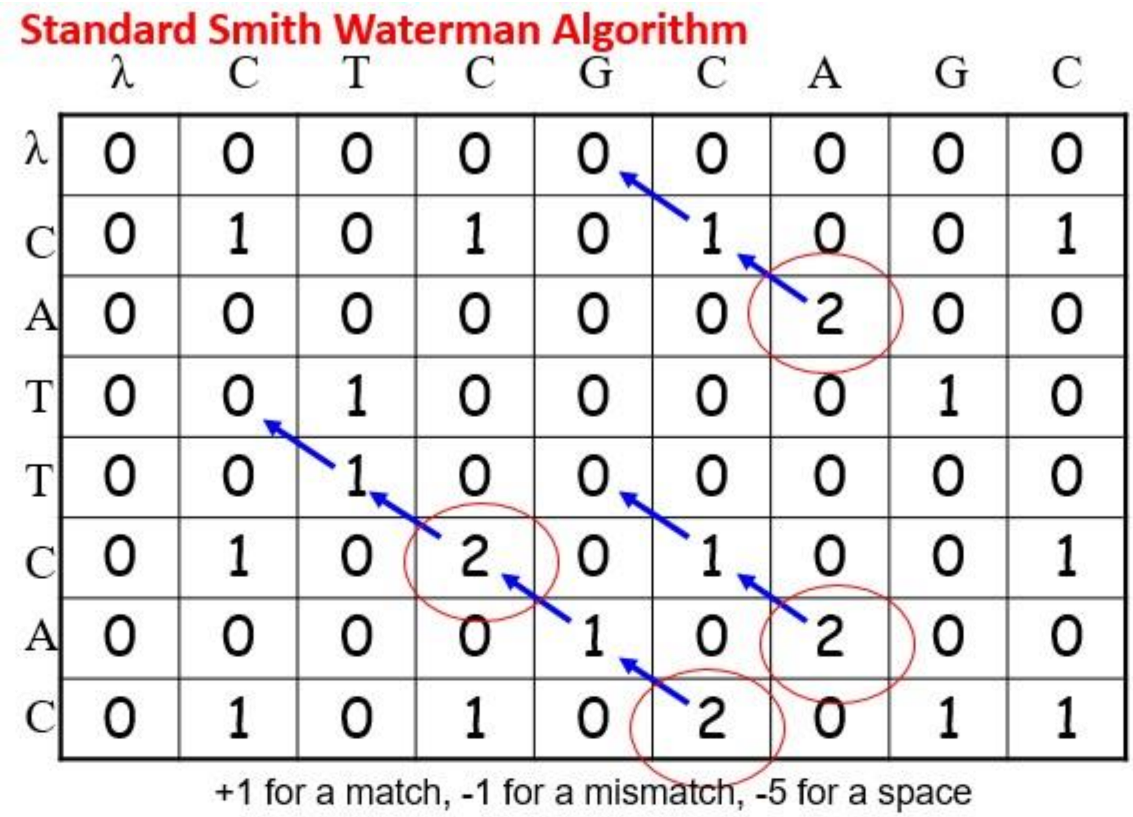
Local alignments can be extracted by starting from a high score till reaching ‘0’

Module 035: REPEATED ALIGNMENTS

We can find the best local alignments by using Smith Waterman algorithm.

By making some change in strategy of traceback we can find the repeated sequences.

We use threshold “T” score for matching and it avoids low scoring local alignment. And traceback can help us to find multiple aligned regions in multiple ways.



Standard Smith Waterman Algorithm

	λ	C	T	C	G	C	A	G	C
λ	0	0	0	0	0	0	0	0	0
C	0	1	0	1	0	1	0	0	1
A	0	0	0	0	0	0	2	0	0
T	0	0	1	0	0	0	0	1	0
T	0	0	1	0	0	0	0	0	0
C	0	1	0	2	0	1	0	0	1
A	0	0	0	0	1	0	2	0	0
C	0	1	0	1	0	2	0	1	1

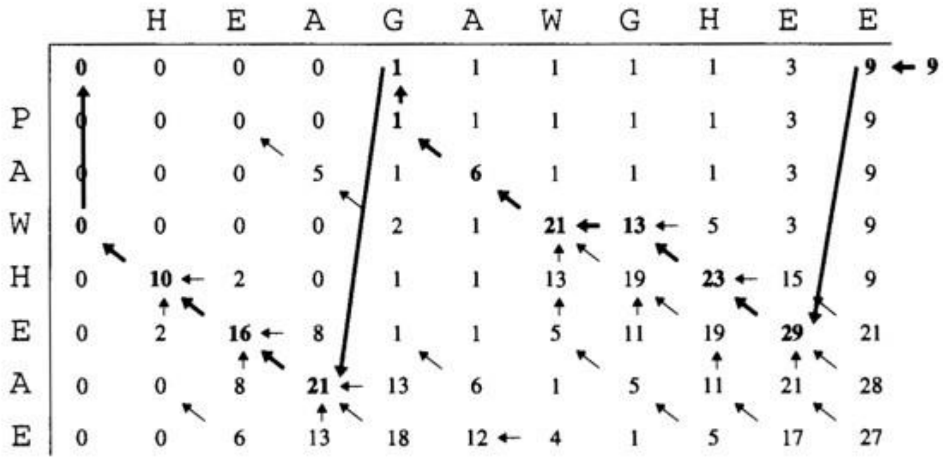
+1 for a match, -1 for a mismatch, -5 for a space

Figure 27 (-5) is threshold score in table

This threshold scoring method with some modifications in waterman algorithm can help us to find many matching sequence of amino acids or DNA.

Module 036: EXAMPLES OF REPEATED ALIGNMENTS

Slight modification in waterman model can help us to find the Exons as well as the functional units in any sequence. Matches should be end at the threshold score or we should keep track of maximum score in sequence.



		H	E	A	G	A	W	G	H	E	E	
	0	0	0	0	1	1	1	1	1	3	9	← 9
P	0	0	0	0	1	1	1	1	1	3	9	
A	0	0	0	5	1	6	1	1	1	3	9	
W	0	0	0	0	2	1	21	13	5	3	9	
H	0	10	2	0	1	1	13	19	23	15	9	
E	0	2	16	8	1	1	5	11	19	29	21	
A	0	0	8	21	13	6	1	5	11	21	28	
E	0	0	6	13	18	12	4	1	5	17	27	

Match = +2. Mismatch = -1, Gap = -2, T = 1

```
HEAGAWGHEE
HEA.AW-HE.
```

Figure 28 Trackback from different sides to find maximum or Threshold (T) Score.

Traceback should start from last element of the row and should reach at the top of row element and then move to the highest score of the Column. And this traceback is done twice and end at the point where score become “0, 0”

Module 037: INTRODUCTION TO SCORING ALIGNMENTS

There are two types of alignments;

- Optimal Alignments
- Best Alignment

Scoring scheme used in sequences matches play crucial role in producing optimal alignment. An optimal alignment should be:

- Appropriately rewarded for matches and mismatches.

• INTRODUCTION

We identify the pairs of symbol which most frequently appear in a sequences it helps us to find the substitution of specific pair of amino acid or nucleotide with other on in a sequence.

For example AA nucleotide have a specific pattern of substitution. And same pairs of amino acids does in protein sequences because it can help to preserve the function of protein.

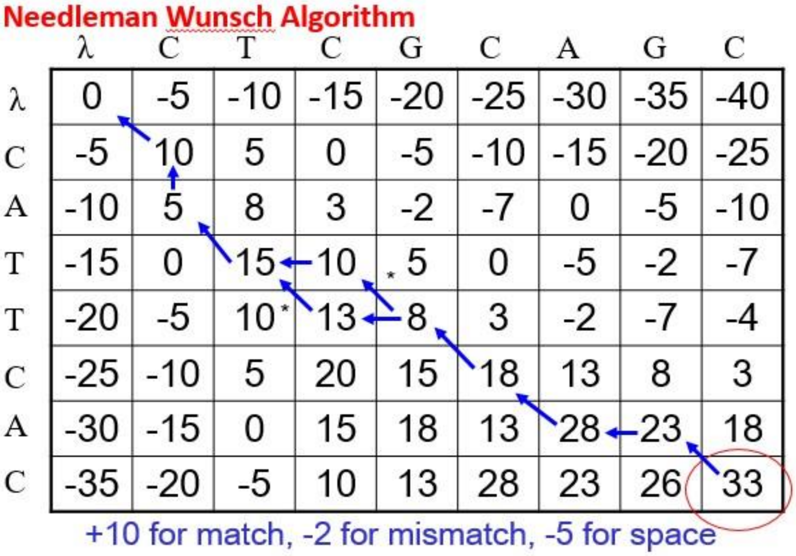
• CONCLUSION

Statically we can better align any sequence of protein or DNA, optimal gaps, penalties, insertions and deletions can be computed statically better.

Module 038: MEASURING ALIGNMENTS SCORES

Score of match and mismatch both are equally observed while sequence alignments.

For example:



Needleman Wunsch Algorithm

	λ	C	T	C	G	C	A	G	C
λ	0	-5	-10	-15	-20	-25	-30	-35	-40
C	-5	10	5	0	-5	-10	-15	-20	-25
A	-10	5	8	3	-2	-7	0	-5	-10
T	-15	0	15	10	* 5	0	-5	-2	-7
T	-20	-5	10*	13	8	3	-2	-7	-4
C	-25	-10	5	20	15	18	13	8	3
A	-30	-15	0	15	18	13	28	23	18
C	-35	-20	-5	10	13	28	23	26	33

+10 for match, -2 for mismatch, -5 for space

Figure 29 Needleman wunsch algorithm match, mismatch scoring

The matrix has positive and negative scores both, matches and mismatches therefore are all considered because it's a diagonal pattern.

If we build such scoring matrixes with matches and mismatches we can we can sequence in according to real life.

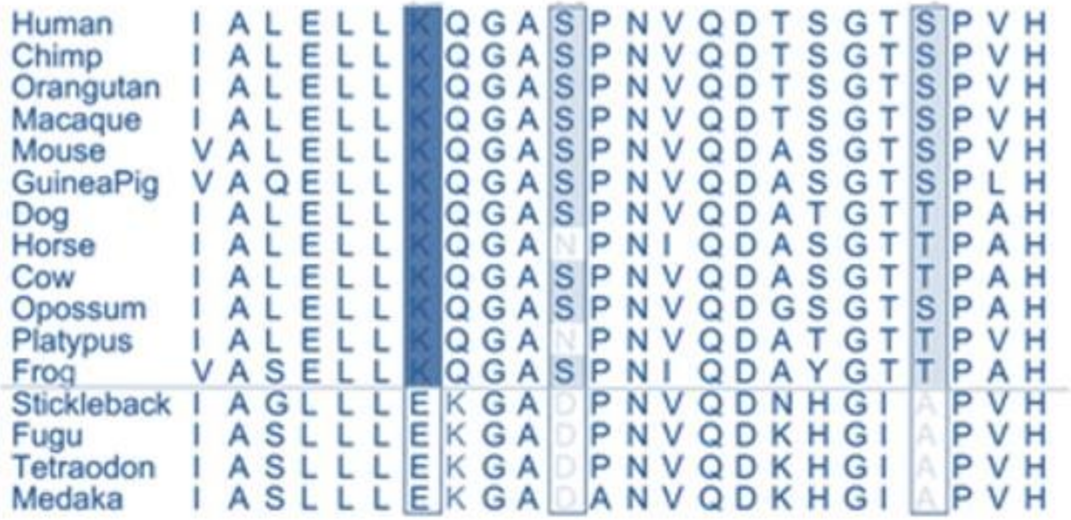
Module 039: SCORING MATRICES

Alignments are used to align the biological sequences. Amino acids and nucleotides are more easily substituted because they have similar chemical nature.

As amino acids are substituted with many probabilities that’s why we need flexible scoring. And we use **Scoring Matrices** contain such flexible scoring during alignment.

To build the Scoring Matrices we analyze the amino acids and nucleotides which are substituted in single gene and protein sequence.

Scoring Matrices have both values +ve and –ve. Positive value for matches and negative value for mismatches.



```
Human       I A L E L L K Q G A S P N V Q D T S G T S P V H
Chimp       I A L E L L K Q G A S P N V Q D T S G T S P V H
Orangutan   I A L E L L K Q G A S P N V Q D T S G T S P V H
Macaque     I A L E L L K Q G A S P N V Q D T S G T S P V H
Mouse       V A L E L L K Q G A S P N V Q D A S G T S P V H
GuineaPig   V A Q E L L K Q G A S P N V Q D A S G T S P L H
Dog         I A L E L L K Q G A S P N V Q D A T G T T P A H
Horse       I A L E L L K Q G A N P N I Q D A S G T T P A H
Cow         I A L E L L K Q G A S P N V Q D A S G T T P A H
Opossum     I A L E L L K Q G A S P N V Q D G S G T S P A H
Platypus    I A L E L L K Q G A N P N V Q D A T G T T P V H
Frog        V A S E L L K Q G A S P N I Q D A Y G T T P A H
Stickleback I A G L L L E K G A D P N V Q D N H G I A P V H
Fugu        I A S L L L E K G A D P N V Q D K H G I A P V H
Tetraodon   I A S L L L E K G A D P N V Q D K H G I A P V H
Medaka      I A S L L L E K G A D A N V Q D K H G I A P V H
```

Figure 0.12 Ubiquitin Protein where amino acids matching

Different type of scoring matrices can be developed based on underlying strategy.

Module 040: DERIVING SCORING MATRICES

Each amino acid have different property.

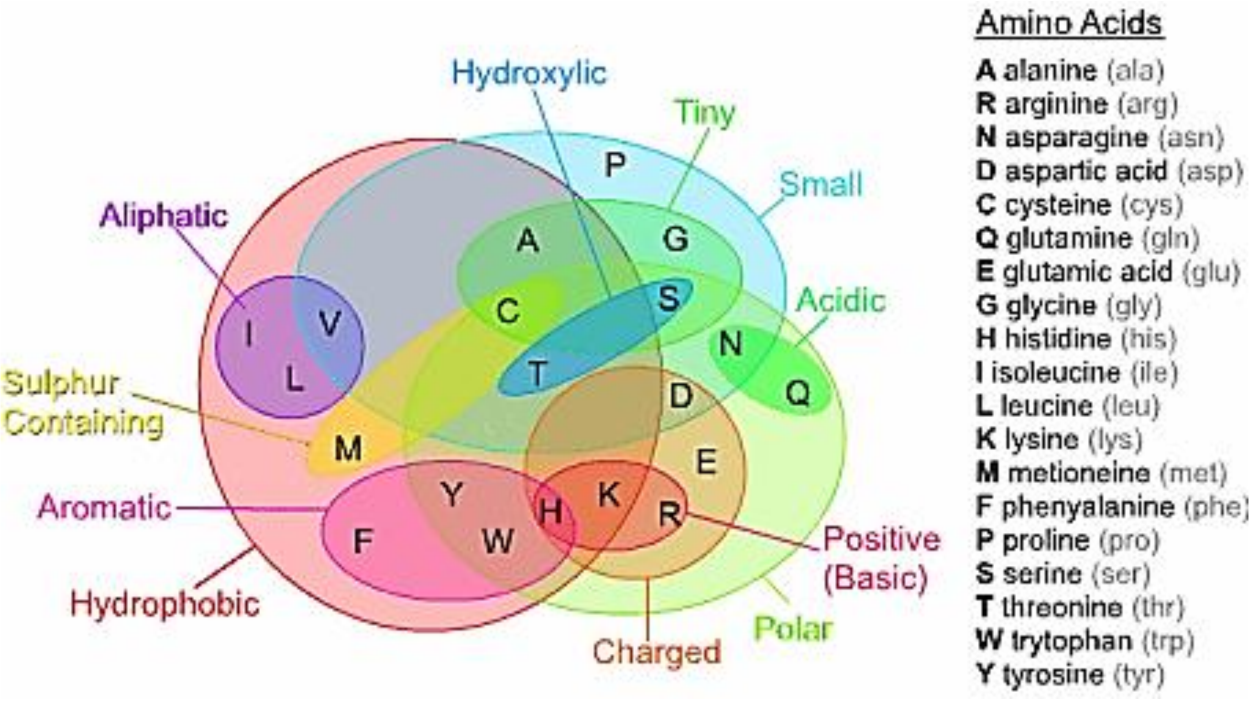


Figure 0.13 properties of amino acids (Image Esquivel et al. (2013)

And each amino acid have different frequency.

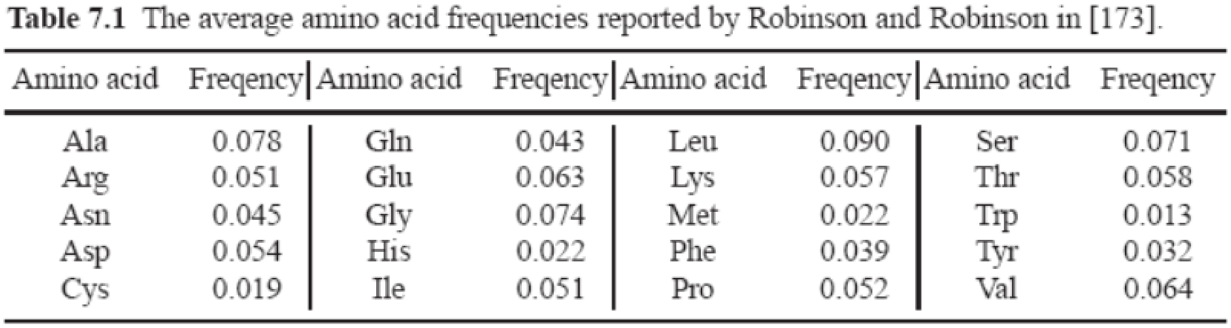


Table 7.1 The average amino acid frequencies reported by Robinson and Robinson in [173].

Amino acid	Freqency	Amino acid	Freqency	Amino acid	Freqency	Amino acid	Freqency
Ala	0.078	Gln	0.043	Leu	0.090	Ser	0.071
Arg	0.051	Glu	0.063	Lys	0.057	Thr	0.058
Asn	0.045	Gly	0.074	Met	0.022	Trp	0.013
Asp	0.054	His	0.022	Phe	0.039	Tyr	0.032
Cys	0.019	Ile	0.051	Pro	0.052	Val	0.064

When we compare the sequences they match and mismatch according to their frequency.

For example.

B	A	B	A
A	A	C	A
A	A	C	C
A	A	B	A
A	A	B	C
A	A	B	B

aligned pairs	*observed frequencies* f_{ij}
A to A	28/60
A to B	8 /60
A to C	6/60
B to B	6/60
B to C	10/60
C to C	2/60

amino acids	*observed frequencies* f_i
A	14/24
B	6/24
C	4/24

Based on frequencies we match and mismatch the sequence alignments for scoring.

Module 041: PAM MATRICES

Alignment matrices scoring is very useful method to score the sequences alignment for match and mismatch.

There are two types of scoring matrices.

- PAM
- BLOSUM

PAM means “Point Accepted Mutation”

Point accepted mutations means the substitution of one amino acid in a sequence with another that protein function remain conserved.

- **PAM UNIT**

PAM unit is actually that time during which 1% amino acid undergo for acceptable mutation. If two sequences diverge by 100 PAM units, it does not mean that they will be at totally different positions.

- **STEP TOCOMPUTE PAM MATRICES**

1. Align the protein sequence which are 1-PAM Unit diverge.
2. Let Ai,i be the number of times Ai is substituted by Ai.
3. Compute the frequency fi of amino acid Ai.

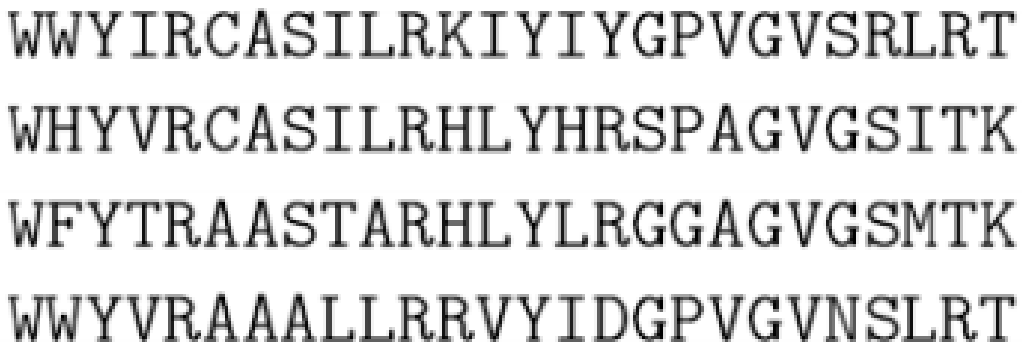
Then, PAM1=pii= $\frac{A_{ij}}{\sum_k A_{ik}}$

PAM ‘n’= $(\text{PAM1})^n$

Module 042: BLOSUM MATRICES

BLOSUM matrices can be used to align the protein sequences. BLOSUM matrices was first purposed in 1992 by Henikoff et al.

BLOSUM matrices is also called the Block substitution matrix without any gap although it has mismatches in sequences.



WWYIRCASILRKIYIYGPVGVSRLRT
WHYVRCASILRHLYHRSPAGVGSITK
WFYTRAASTARHLYLRGGAGVGSMTK
WWYVRAAALLRRVYIDGPVGVNSLRT

Figure 0.14 sequence of amino acids which have mismatch but no gap.

There are three steps to compute the BLOUSM Matrices.

Step 1: Eliminate sequences that are identical in x% positions

Step 2: Compute observed frequency f $_{i,\, j}$ of aligned pair A_i to A_j. Hence, f $_{i,j}$ becomes the probability of aligning A_i and A_j in the selected blocks.

Step 3: Compute f_i which is the frequency of observing A_i in the entire block

$$s_{ij} = \log_2 \frac{f_{ij}}{p_{ij}}, \quad p_{ij} = \begin{cases} f_i f_i & i = j \\ 2 f_i f_j & i \neq j. \end{cases}$$

Figure 0.15 formula for computation of BLOSUM MATRICES.

Typically used matrices: BLOSUM62 or PAM120 in PAMx, larger x detects more divergent sequences.

Module 043: MULTIPLE SEQUENCE ALIGNMENT

In pair wise sequence alignments we use pairs of sequence to compare them. And scoring matrices were used to score the sequence ranks.

In Multiple sequence Alignments we compare multiple number of protein and DNA sequences to identify the matches and mismatches.

```
M Q V K L F T P L H D K S D H G K Y H
M Q V K I F T P L H D K S - H G K S H
M Q V H L Y - P L H D K S - T G K S H
M Q V H L F - P L H D K S D T G K S H
```

Figure 0.16 multiple sequence alignment

For pair wise alignment we use Dynamic programming but for multiple alignment it would be very expensive computationally. So solution for this is progressive alignment.

Module 044: MORE ON MULTIPLE SEQUENCE ALIGNMENT

MSA helps compare several sequences by aligning them. MSA can extract consensus sequences from several aligned sequences. Characterize protein families based on homologous regions.

APPLICATION OF MSA

- Predict secondary and tertiary structures of new protein sequences
- Evaluate evolutionary order of species or "Phylogeny"

METHODOLOGY

- Pairwise alignment is the alignment of two sequences
- MSA can be performed by repeated application of pairwise alignment

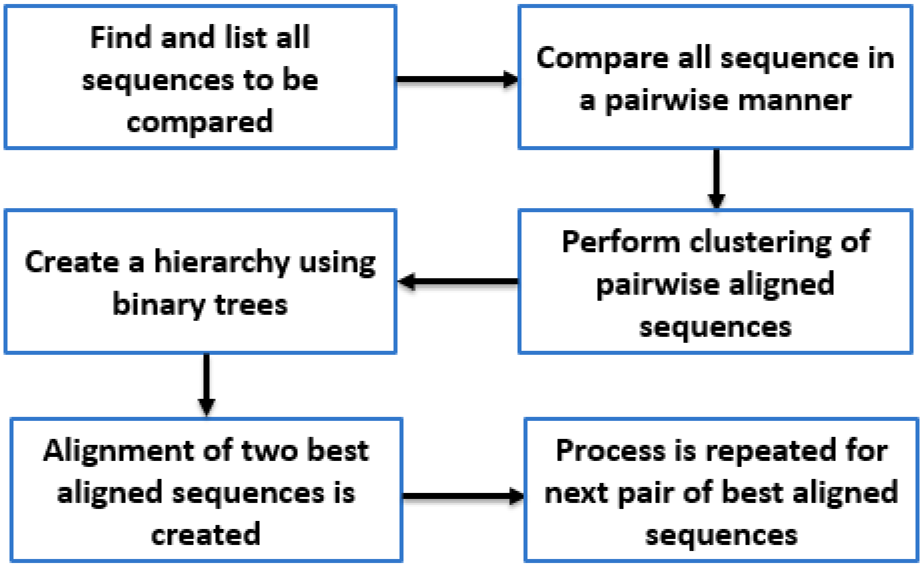

Find and list all sequences to be compared
Compare all sequence in a pairwise manner
Perform clustering of pairwise aligned sequences
Create a hierarchy using binary trees
Alignment of two best aligned sequences is created
Process is repeated for next pair of best aligned sequences

Figure 0.17 Methodology

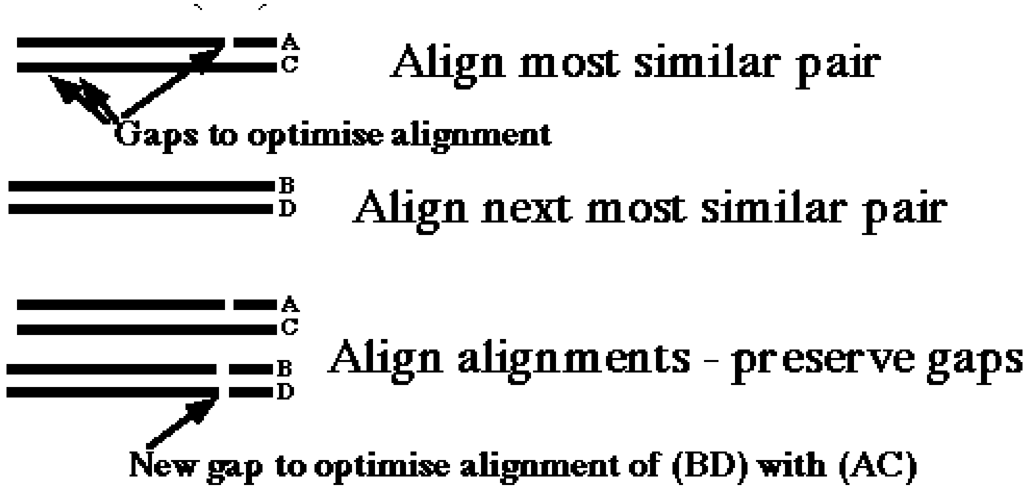
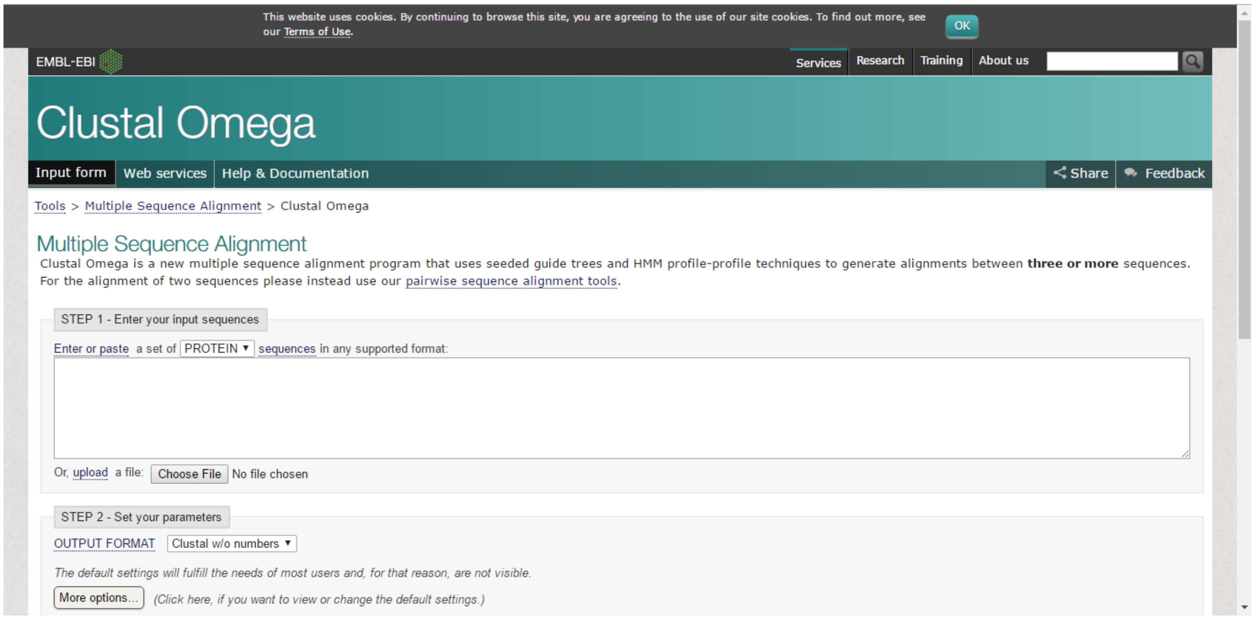

A
C
Align most similar pair
Gaps to optimise alignment
B
D
Align next most similar pair
A
C
B
D
Align alignments - preserve gaps
New gap to optimise alignment of (BD) with (AC)

Figure 0.18 sequence alignments

CONCLUSION

MSA can help align multiple sequences. Progressive alignment can help perform MSA. Need to remove sequences with >80% similarity.



This website uses cookies. By continuing to browse this site, you are agreeing to the use of our site cookies. To find out more, see our Terms of Use.

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

STEP 1 - Enter your input sequences

Enter or paste a set of PROTEIN ▾ sequences in any supported format:

Or, upload a file: Choose File No file chosen

STEP 2 - Set your parameters

OUTPUT FORMAT Clustal w/o numbers ▾

The default settings will fulfill the needs of most users and, for that reason, are not visible.

More options... *(Click here, if you want to view or change the default settings.)*

Figure 0.19 CLUSTAL – Online tool

http://www.ebi.ac.uk/Tools/msa/clustalo/

Module 045: PROGRESSIVE ALIGNMENT FOR MSA

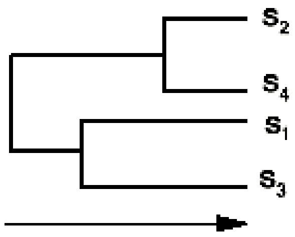
MSA involves progressive alignment of sequences. Doing so many progressive alignments can be slow.

STEPS:

- Step 1 : Pairwise Alignment of all sequences

Example: S_1, S_2, S_3, $S_{4,}$ so that is 6 pairwise comparisons.

- Step 2: Construct a Guide Tree (Dendogram) using a *Distance Matrix.*



$$\mathit{Distance} = 1 - \frac{\mathit{Matches}}{\mathit{Comparisons}}$$

$$\mathit{Similarity} = \frac{\mathit{Matches}}{\mathit{Comparisons}}$$

- Step 3: Progressive alignment following branching order in tree.

$$\mathit{Similarity} = \frac{\mathit{Matches}}{\mathit{Comparisons}}$$

$$\begin{array}{c|cccc}
 & v_1 & v_2 & v_3 & v_4 \\
\hline
v_1 & - & & & \\
v_2 & .17 & - & & \\
v_3 & .87 & .28 & - & \\
v_4 & .59 & .33 & .62 & -
\end{array}$$

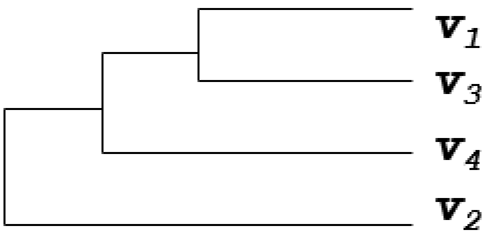


Figure 0.20 Similarity Matrix

- **SHORTCOMING OF THIS APPROACH**

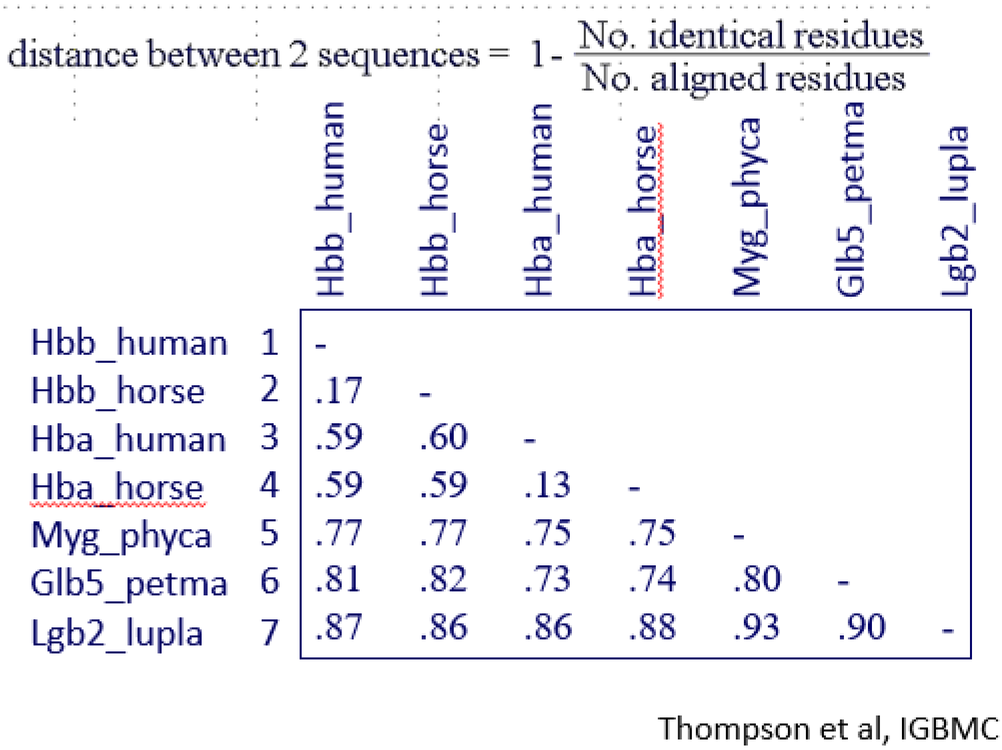
- Dependence upon initial alignments
- If sequences are dissimilar, errors in alignment are propagated
- Solution: Begin by using an initial alignment, and refine it repeatedly

Progressive alignments are used in aligning multiple sequences. Iterative approaches can help refine results from progressive alignments.

Module 046: MSA-EXAMPLE

MSA involves progressive alignment of sequences. Doing so many progressive alignments can be slow.

For example:

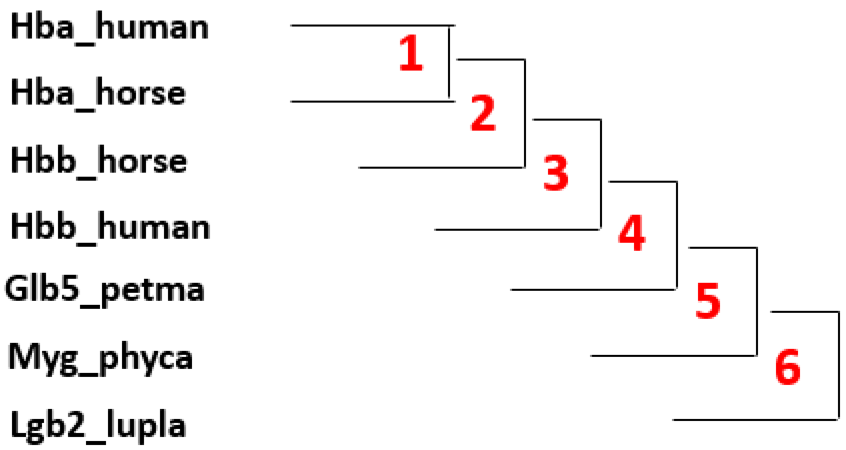


$$\text{distance between 2 sequences} = 1 - \frac{\text{No. identical residues}}{\text{No. aligned residues}}$$

		Hbb_human	Hbb_horse	Hba_human	Hba_horse	Myg_phyca	Glb5_petma	Lgb2_lupla
Hbb_human	1	-						
Hbb_horse	2	.17	-					
Hba_human	3	.59	.60	-				
Hba_horse	4	.59	.59	.13	-			
Myg_phyca	5	.77	.77	.75	.75	-		
Glb5_petma	6	.81	.82	.73	.74	.80	-	
Lgb2_lupla	7	.87	.86	.86	.88	.93	.90	-

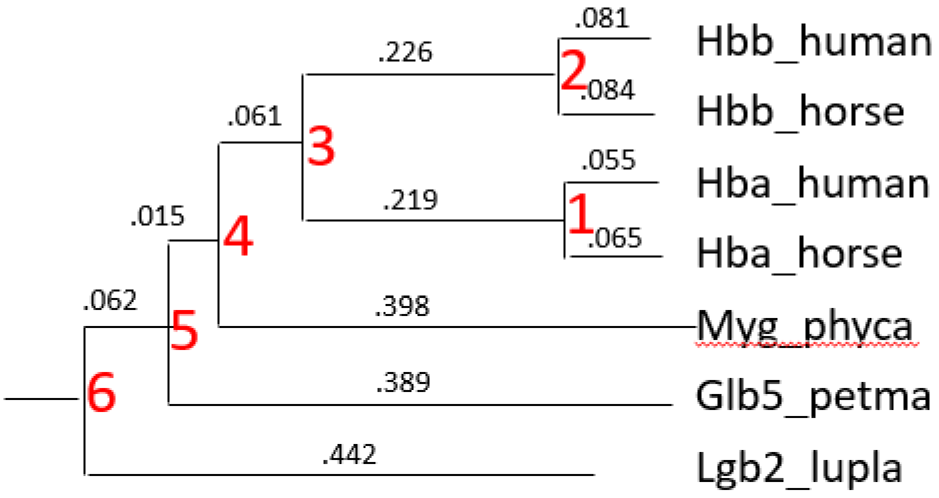
Thompson et al, IGBMC

Figure 0.21 MSA on globin sequences



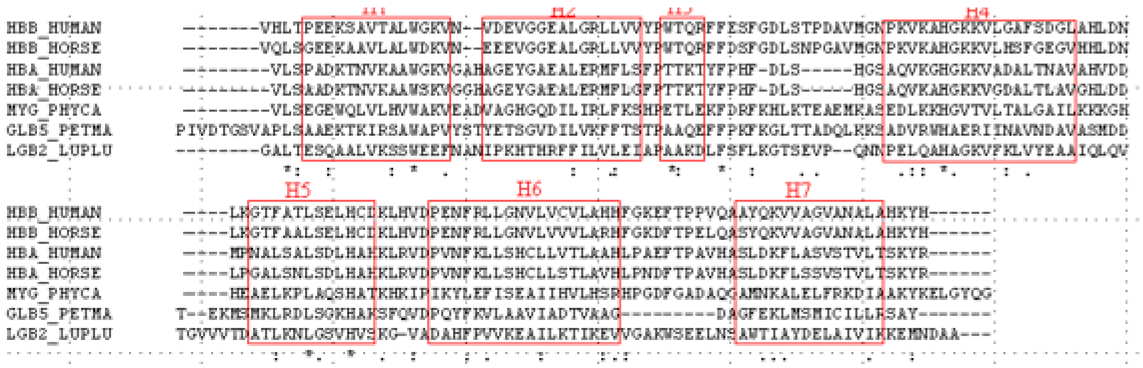
Thompson et al, IGBMC

Figure 0.22 Progressive alignment using sequential branching



Thompson et al, IGBMC

Figure 0.23 Progressive alignment following a guide tree



```
HBB_HUMAN      --------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDN
HBB_HORSE      --------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNPKVKAHGKKVLHSFGEGVHHLDN
HBA_HUMAN      ---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSAQVKGHGKKVADALTNAVAHVDD
HBA_HORSE      ---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHF-DLS-----HGSAQVKAHGKKVGDALTLAVGHLDD
MYG_PHYCA      ---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGH
GLB5_PETMA     PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTTADQLKKSADVRWHAERIINAVNDAVASMDD
LGB2_LUPLU     --------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSEVP--QNNPELQAHAGKVFKLVYEAAIQLQV

HBB_HUMAN      ----LKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
HBB_HORSE      ----LKGTFAALSELHCDKLHVDPENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
HBA_HUMAN      ----MPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
HBA_HORSE      ----LPGALSNLSDLHAHKLRVDPVNFKLLSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
MYG_PHYCA      ----HEAELKPLAQSHATKHKIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
GLB5_PETMA     T---EKMSMKLRDLSGKHAKSFQVDPQYFKVLAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
LGB2_LUPLU     TGVVVTDATLKNLGSVHVSKG--VADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
```

Figure 0.24 Alignment results

MSA can be better performed using clustering strategies followed by alignment of the alignments later. CLUSTAL is a free online tool that does all of this for us!

Module 047: CLUSTALW

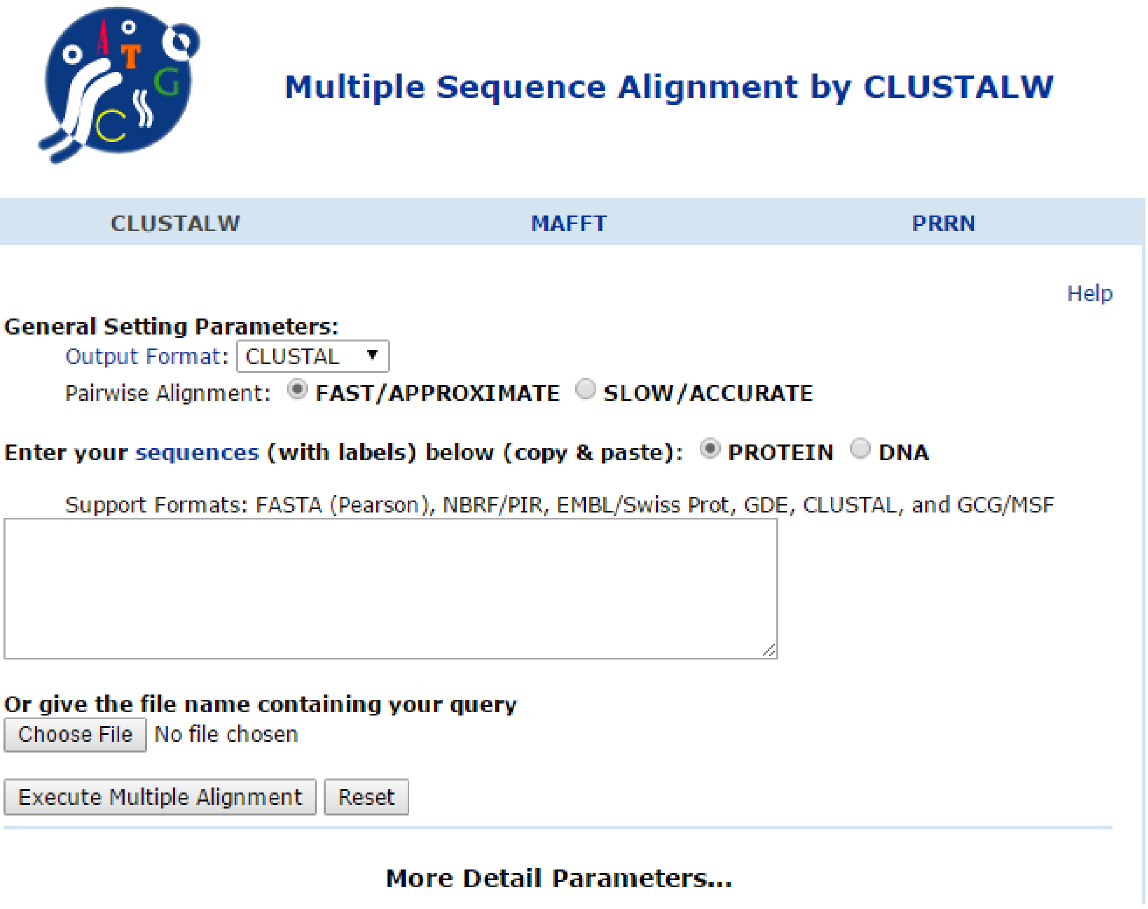
MSA involves progressive alignment of sequences. Doing so many progressive alignments can be slow. CLUSTALW is an online tool to perform MSA.

Developed by European Molecular Biology Laboratory & European Bioinformatics Institute. Performs alignment in:

- slow/accurate
- fast/approximate

SCOPE

- create multiple alignments,
- optimize existing alignments,
- profile analysis &
- create phylogenetic trees



Multiple Sequence Alignment by CLUSTALW

General Setting Parameters:

Output Format: CLUSTAL ▼

Pairwise Alignment: ● FAST/APPROXIMATE ○ SLOW/ACCURATE

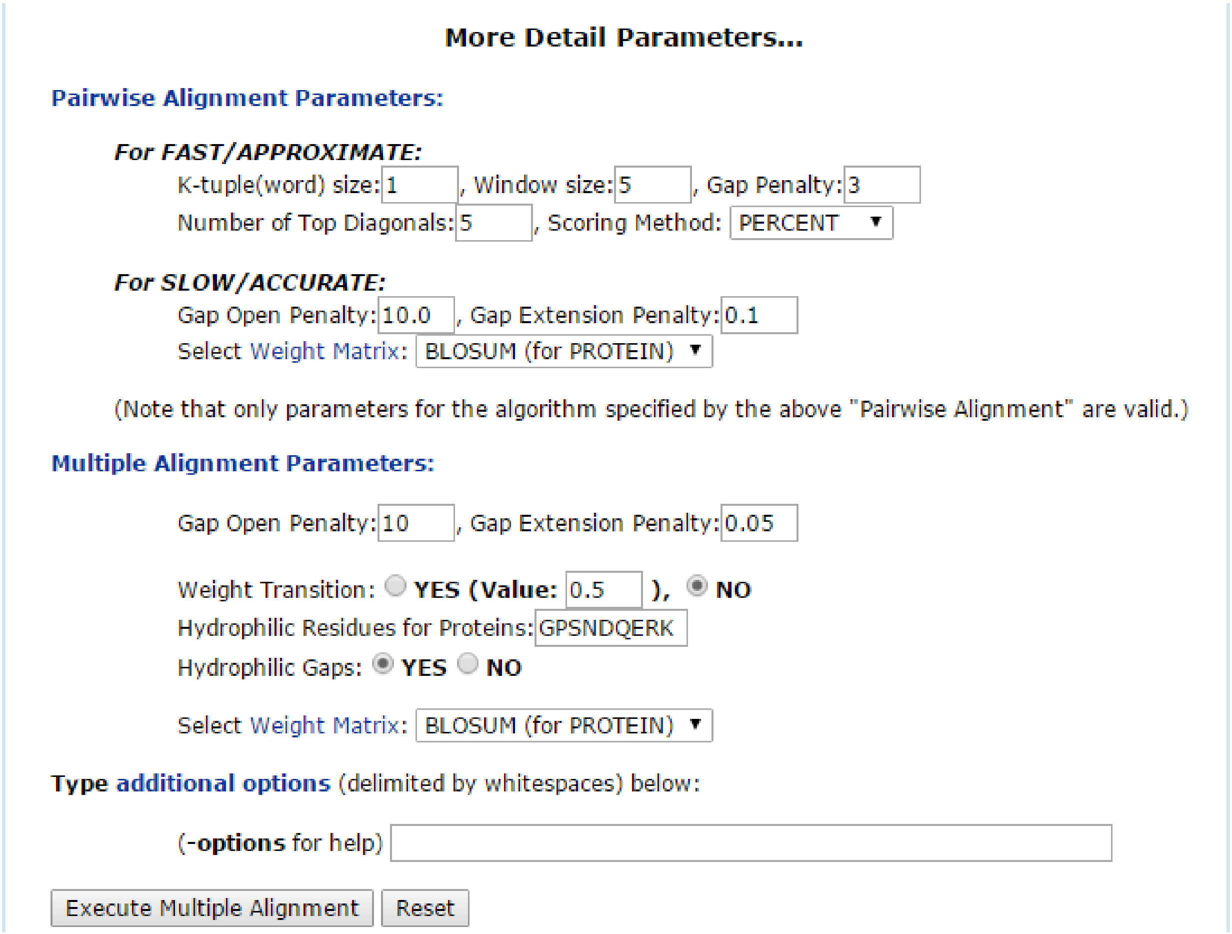
Enter your sequences (with labels) below (copy & paste): ● PROTEIN ○ DNA

Support Formats: FASTA (Pearson), NBRF/PIR, EMBL/Swiss Prot, GDE, CLUSTAL, and GCG/MSF

Or give the file name containing your query

Choose File No file chosen

http://www.genome.jp/tools/clustalw



More Detail Parameters...

Pairwise Alignment Parameters:

For FAST/APPROXIMATE:

K-tuple(word) size: 1 , Window size: 5 , Gap Penalty: 3

Number of Top Diagonals: 5 , Scoring Method: PERCENT ▼

For SLOW/ACCURATE:

Gap Open Penalty: 10.0 , Gap Extension Penalty: 0.1

Select Weight Matrix: BLOSUM (for PROTEIN) ▼

(Note that only parameters for the algorithm specified by the above "Pairwise Alignment" are valid.)

Multiple Alignment Parameters:

Gap Open Penalty: 10 , Gap Extension Penalty: 0.05

Weight Transition: ○ **YES (Value:** 0.5 **),** ● **NO**

Hydrophilic Residues for Proteins: GPSNDQERK

Hydrophilic Gaps: ● **YES** ○ **NO**

Select Weight Matrix: BLOSUM (for PROTEIN) ▼

Type additional options (delimited by whitespaces) below:

(**-options** for help)

Execute Multiple Alignment
Reset

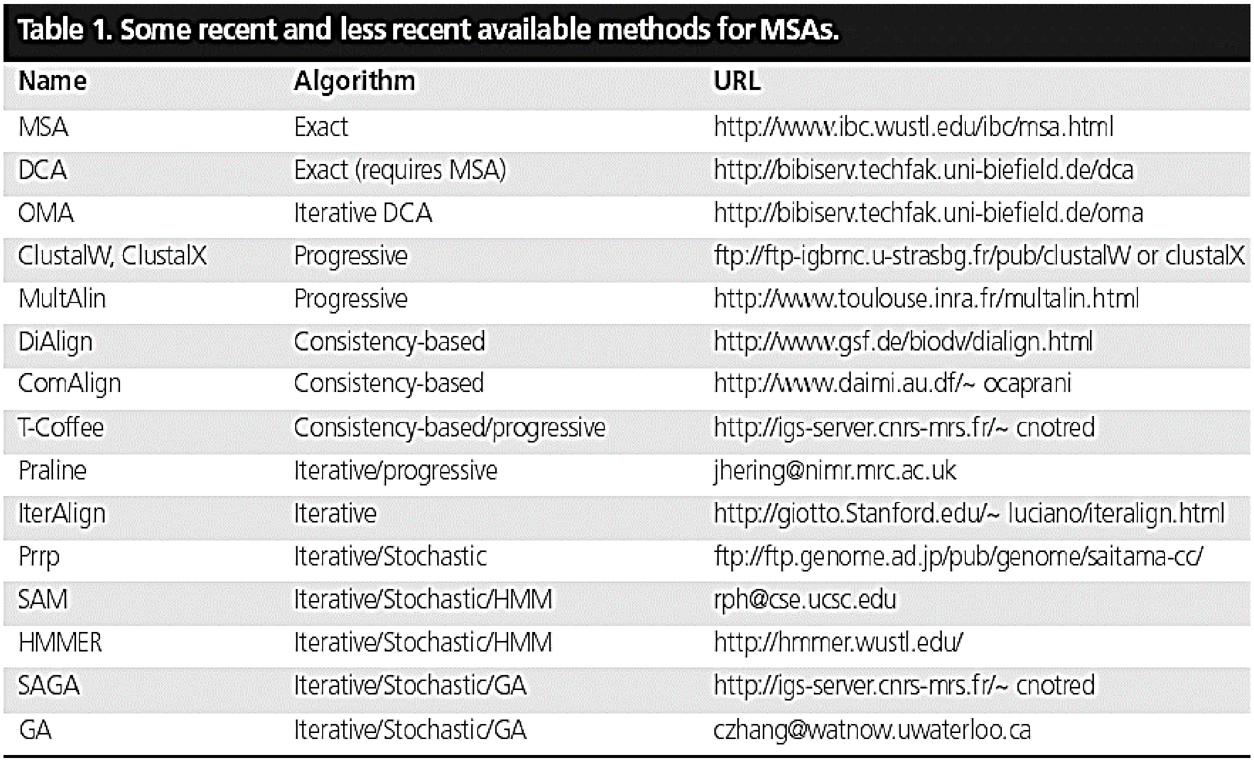


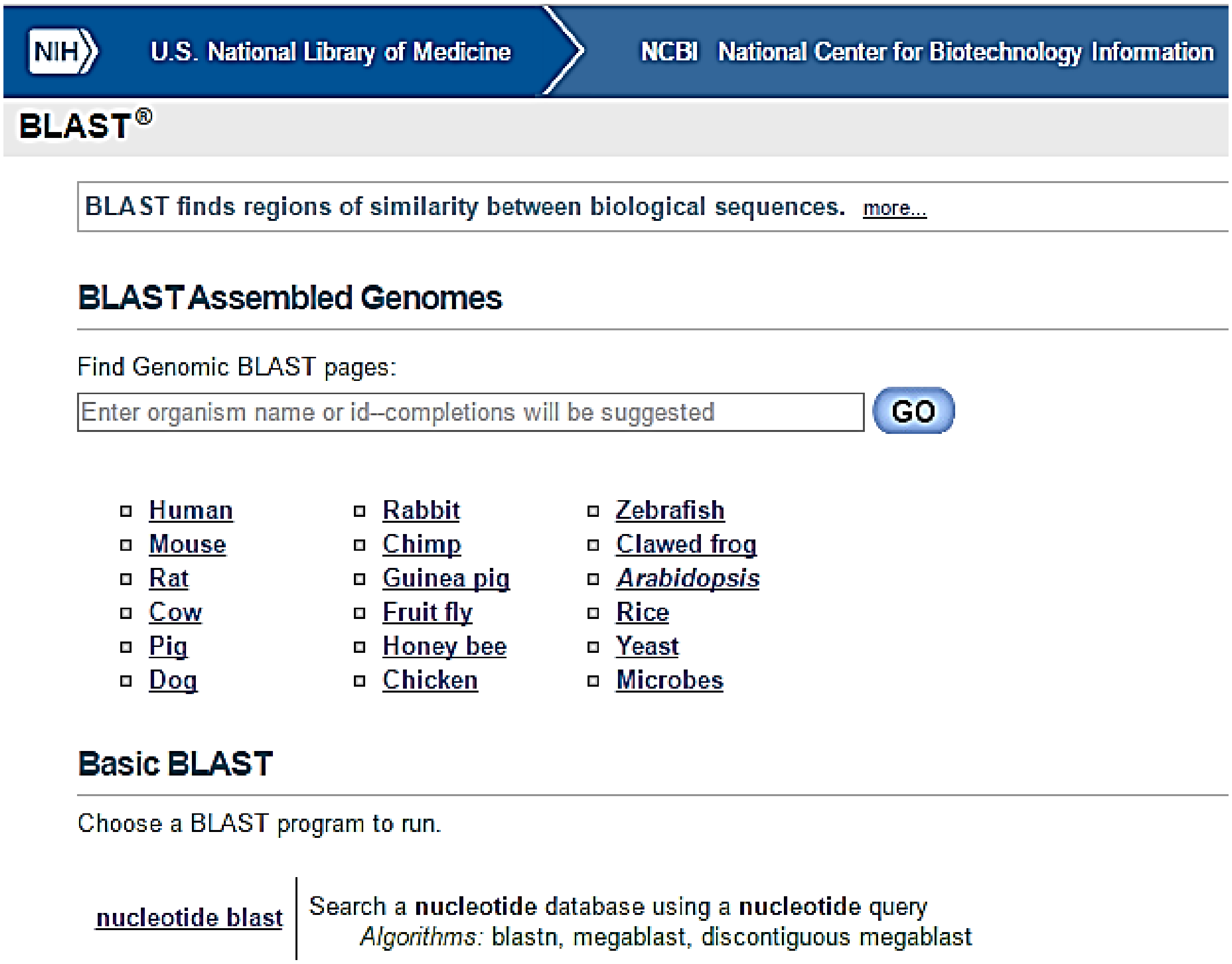
Table 1. Some recent and less recent available methods for MSAs.

Name	Algorithm	URL
MSA	Exact	http://www.ibc.wustl.edu/ibc/msa.html
DCA	Exact (requires MSA)	http://bibiserv.techfak.uni-biefield.de/dca
OMA	Iterative DCA	http://bibiserv.techfak.uni-biefield.de/oma
ClustalW, ClustalX	Progressive	ftp://ftp-igbmc.u-strasbg.fr/pub/clustalW or clustalX
MultAlin	Progressive	http://www.toulouse.inra.fr/multalin.html
DiAlign	Consistency-based	http://www.gsf.de/biodv/dialign.html
ComAlign	Consistency-based	http://www.daimi.au.df/~ ocaprani
T-Coffee	Consistency-based/progressive	http://igs-server.cnrs-mrs.fr/~ cnotred
Praline	Iterative/progressive	jhering@nimr.mrc.ac.uk
IterAlign	Iterative	http://giotto.Stanford.edu/~ luciano/iteralign.html
Prrp	Iterative/Stochastic	ftp://ftp.genome.ad.jp/pub/genome/saitama-cc/
SAM	Iterative/Stochastic/HMM	rph@cse.ucsc.edu
HMMER	Iterative/Stochastic/HMM	http://hmmer.wustl.edu/
SAGA	Iterative/Stochastic/GA	http://igs-server.cnrs-mrs.fr/~ cnotred
GA	Iterative/Stochastic/GA	czhang@watnow.uwaterloo.ca

Module 048: INTRODUCTION TO BLAST-I

National Center for the Biotechnology Information (NCBI) – USA. BLAST developed in 1990. "Basic Local Alignment Search Tool". Searches databases for query protein and nucleotide sequences. Also searches for translational products etc. Online availability

www.blast.ncbi.nlm.nih.gov/Blast.cgi


NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
BLAST®
GO

BLAST finds regions of similarity between biological sequences. more...

BLAST Assembled Genomes

Find Genomic BLAST pages:

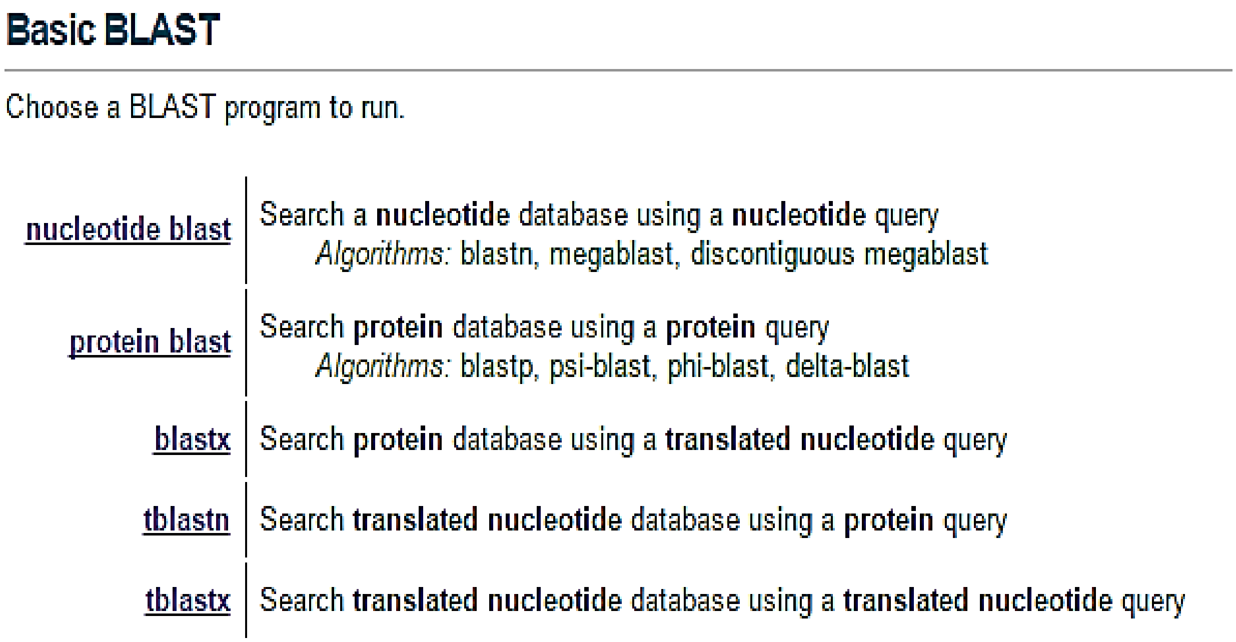
Enter organism name or id--completions will be suggested

- Human
- Mouse
- Rat
- Cow
- Pig
- Dog
- Rabbit
- Chimp
- Guinea pig
- Fruit fly
- Honey bee
- Chicken
- Zebrafish
- Clawed frog
- *Arabidopsis*
- Rice
- Yeast
- Microbes

Basic BLAST

Choose a BLAST program to run.

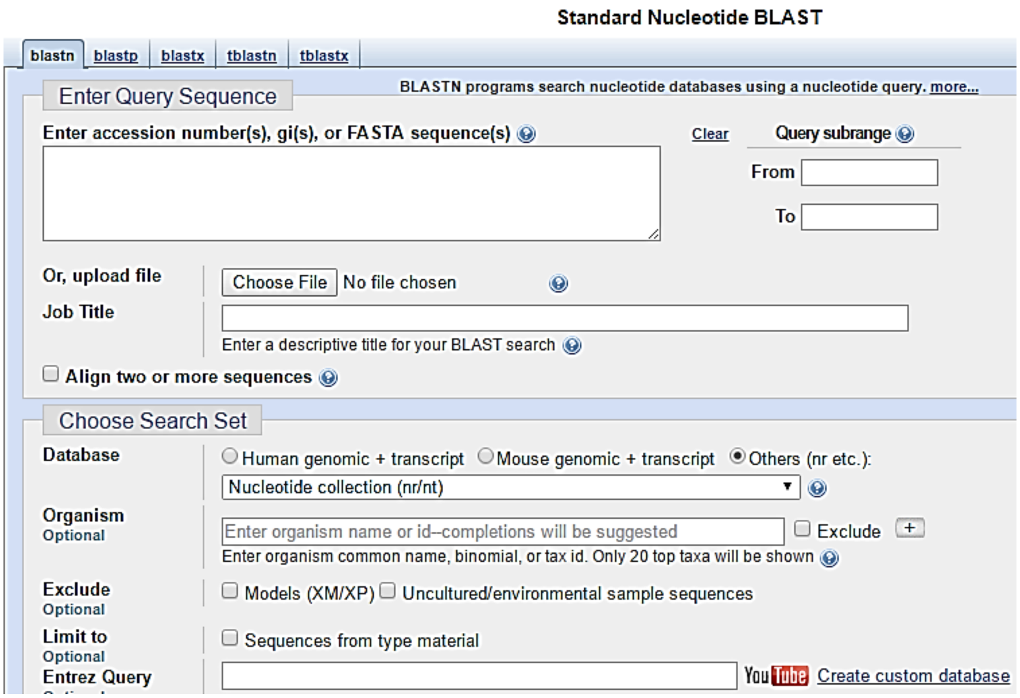
nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast



Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query



Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Query subrange

From

To

Or, upload file

No file chosen

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):

Nucleotide collection (nr/nt) ▼

Organism
Optional

Enter organism name or id--completions will be suggested □ Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude
Optional

□ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to
Optional

□ Sequences from type material

Entrez Query

BLAST can be used to search for local alignment of protein and nucleotide sequences. It is available online. Can perform searches across species and organisms

Module 049: INTRODUCTION TO BLAST-II

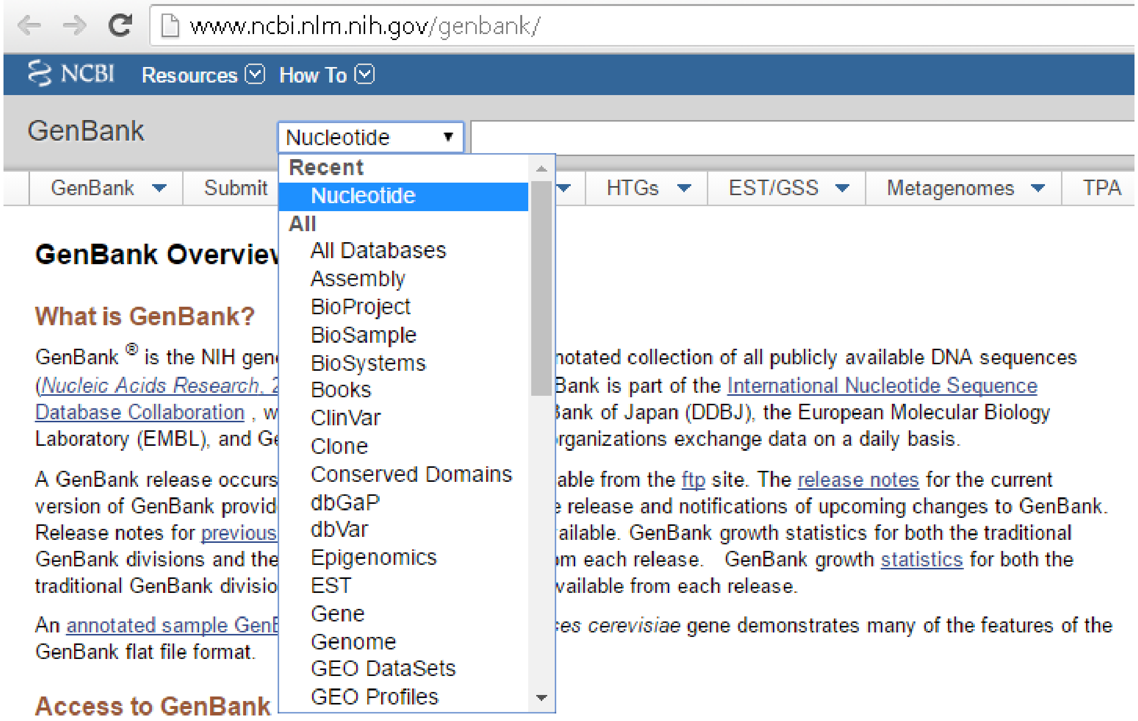
National Center for the Biotechnology Information (NCBI) – USA. BLAST developed in 1990. “Basic Local Alignment Search Tool”. Searches databases for query protein and nucleotide sequences. Also searches for translational products etc. Online availability

www.blast.ncbi.nlm.nih.gov/Blast.cgi

Smith Waterman can align complete sequences. BLAST does it in an approximate way. Hence, BLAST is faster BUT does not ensure optimal alignment. BLAST provides for approximate sequence matching. Input to BLAST is a FASTA formatted sequence and a set of search parameters

OUTPUT OF BLAST

Results are shown in HTML, plain text, and XML formats. A table lists the sequence hits found along with scores. Users can read this table off and evaluate results



GenBank Overvie

What is GenBank?

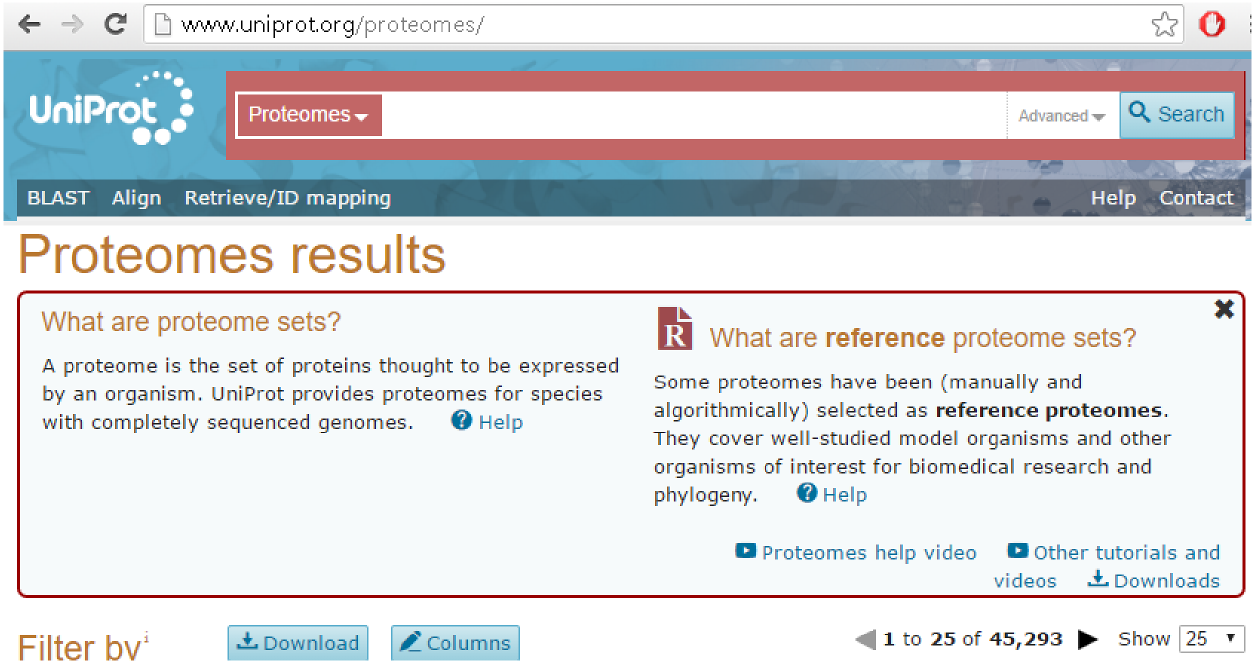
GenBank ® is the NIH gen notated collection of all publicly available DNA sequences (Nucleic Acids Research, 2 Bank is part of the International Nucleotide Sequence Database Collaboration , w Bank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and Ge rganizations exchange data on a daily basis.

A GenBank release occurs able from the ftp site. The release notes for the current version of GenBank provid e release and notifications of upcoming changes to GenBank. Release notes for previous ailable. GenBank growth statistics for both the traditional GenBank divisions and the m each release. GenBank growth statistics for both the traditional GenBank divisio vailable from each release.

An annotated sample GenB es *cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

Figure 0.25Input to BLAST: Gene IDs



Proteomes results

What are proteome sets?

A proteome is the set of proteins thought to be expressed by an organism. UniProt provides proteomes for species with completely sequenced genomes. Help

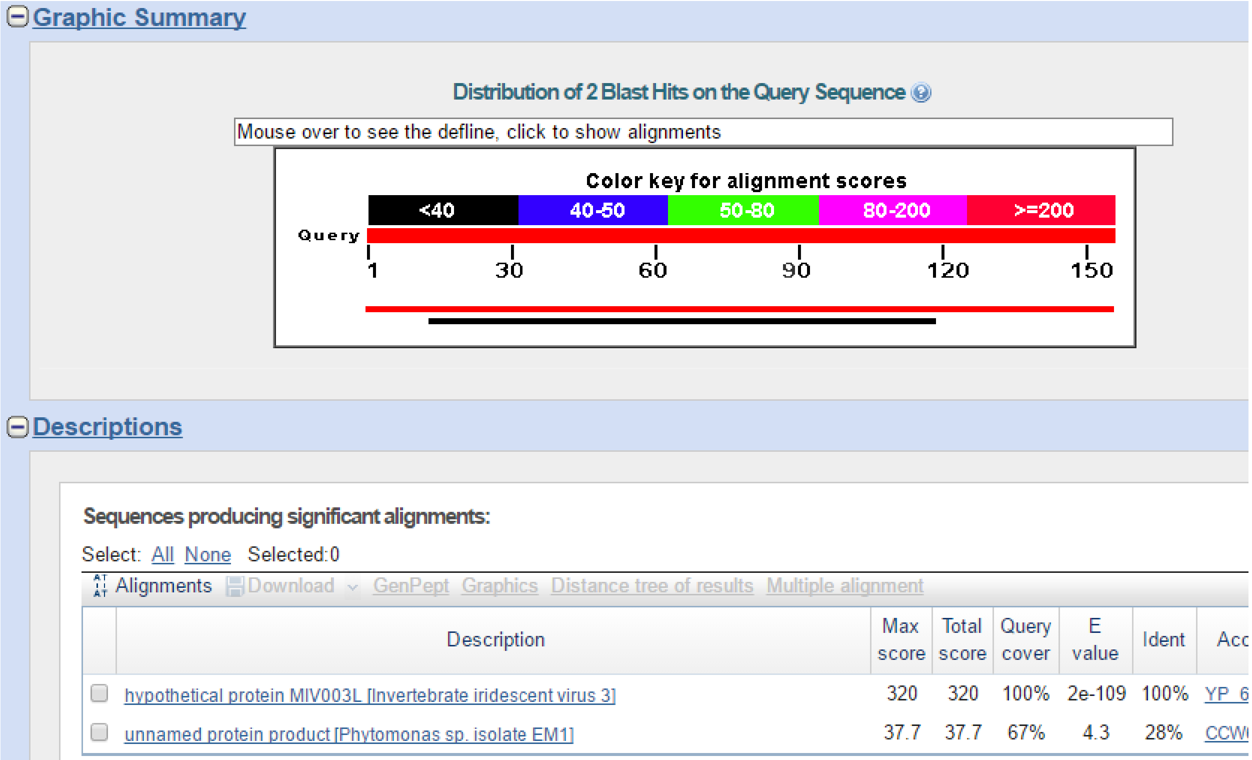
What are **reference** proteome sets?

Some proteomes have been (manually and algorithmically) selected as **reference proteomes**. They cover well-studied model organisms and other organisms of interest for biomedical research and phylogeny. Help

Proteomes help video Other tutorials and videos Downloads

1 to **25** of **45,293**

Figure 0.26 Input to BLAST: Protein IDs


Graphic Summary
Distribution of 2 Blast Hits on the Query Sequence
Mouse over to see the defline, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
1
30
60
90
120
150
Descriptions
Select: All None Selected:0
Alignments
Download
GenPept
Graphics
Distance tree of results
Multiple alignment

Sequences producing significant alignments:

	Description	Max score	Total score	Query cover	E value	Ident	Acc
□	hypothetical protein MIV003L [Invertebrate iridescent virus 3]	320	320	100%	2e-109	100%	YP_6
□	unnamed protein product [Phytomonas sp. isolate EM1]	37.7	37.7	67%	4.3	28%	CCW

Figure 0.27 Results from BLAST

Module 050: BLAST ALGORITHM

BLAST can search sequence databases and identify unknown sequences by comparing them to the known sequences. This can help identify the parent organism, function and evolutionary history.

For example:

Query sequence: PQGELV

Make list of all possible worlds (length 3 for proteins)

PQG (score 15)

QGE (score 9)

GEL (score 12)

ELV (score 10)

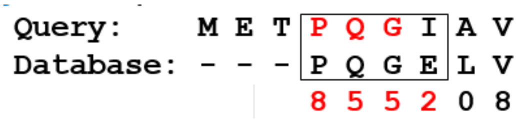
Assign scores from Blosum62, use those with score> 11: PQG & GEL

Mutate words such that score still > 11

PQG (score 15) similar to PEG (score 13)

At the end, we get: PQG, GEL and PEG

Find all database sequences that have at least 2 matches among our 3 words: PQG, GEL & PEG. Find database hits and extend alignment (High-scoring Segment Pair):



```
Query:     M E T P Q G I A V
Database:  - - - P Q G E L V
                 8 5 5 2 0 8
```

High Scoring Pair: PQGI (score 8+5+5+2)

If 2 HSP in query sequence are < 40 positions away

Full dynamic alignment on query and hit sequences

BLAST performs quick alignments on sequences. The results are tabulated with alignment regions overlapping each other. Statistical evaluation is also provided alongside

Module 051: TYPES OF BLAST

BLAST can search sequence databases and identify unknown sequences by comparing them to the known sequences. This can help identify the parent organism, function and evolutionary history.

There are two main types of BLAST.

Nucleotides

- **Blastn:** Compares a nucleotide query sequence against a nucleotide database.

Proteins

- **Blastp:** Compares an amino acid query sequence against a protein database.

There are also many other types of BLAST:

- **Blastx:**

Compares a nucleotide query sequence against a protein sequence database.

Helps find potential translation products of unknown nucleotide sequences

- **tblastn:**

Compares a protein query sequence against a nucleotide sequence database

Nucleotide sequence dynamically translated into all reading frames

- **tblastx:**

Compares the six-frame translated proteins of a nucleotide query sequence against the six-frame translated proteins of a nucleotide sequence database.

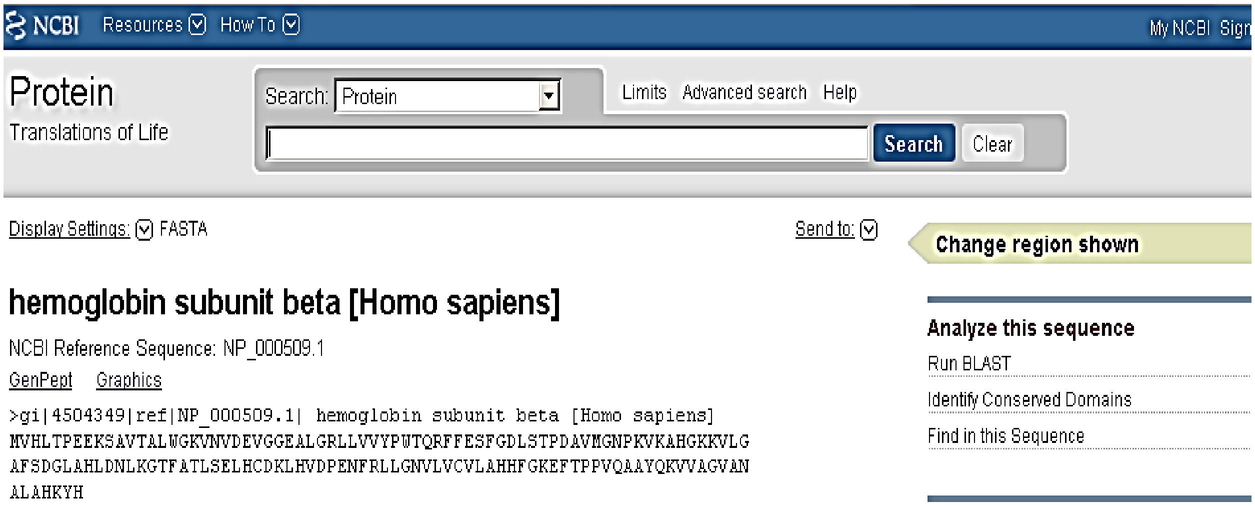
- BLAST performs quick alignments on biological sequences

- Several types of BLAST exist which can assist in comparing nucleotide sequences with amino acids and vice versa

Module 052: SUMMERY OF BLAST

BLAST can search sequence databases and identify unknown sequences by comparing them to the known sequences. This can help identify the parent organism, function and evolutionary history.

Step1: obtain a query of sequence



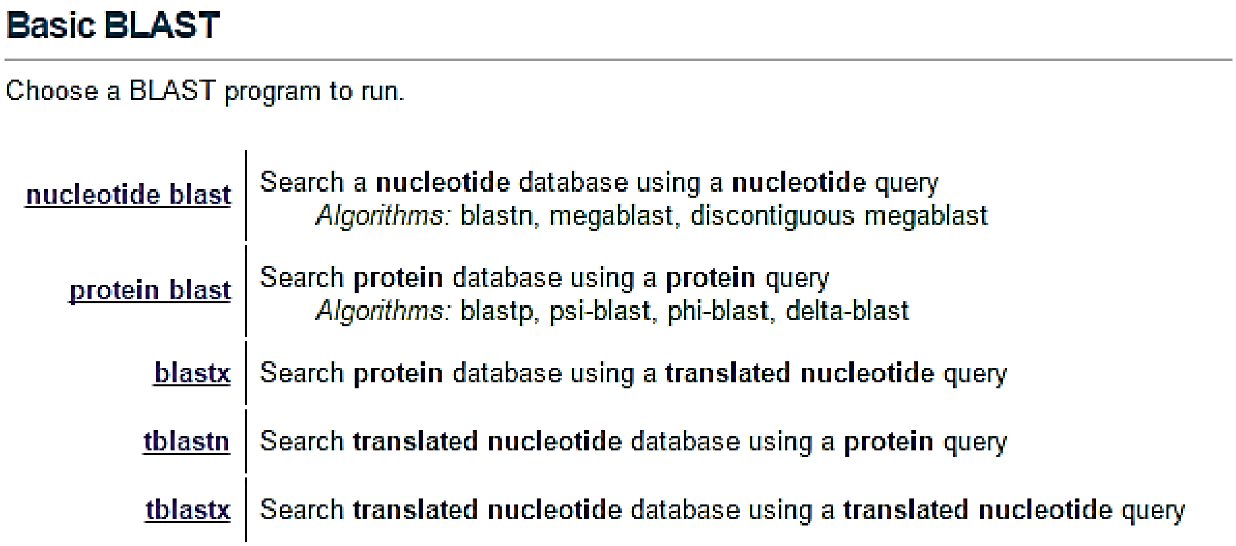
hemoglobin subunit beta [Homo sapiens]

NCBI Reference Sequence: NP_000509.1

GenPept Graphics

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

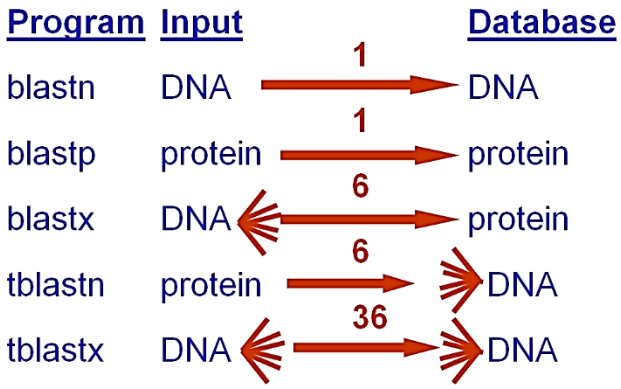
Step2: choose a type of BLAST



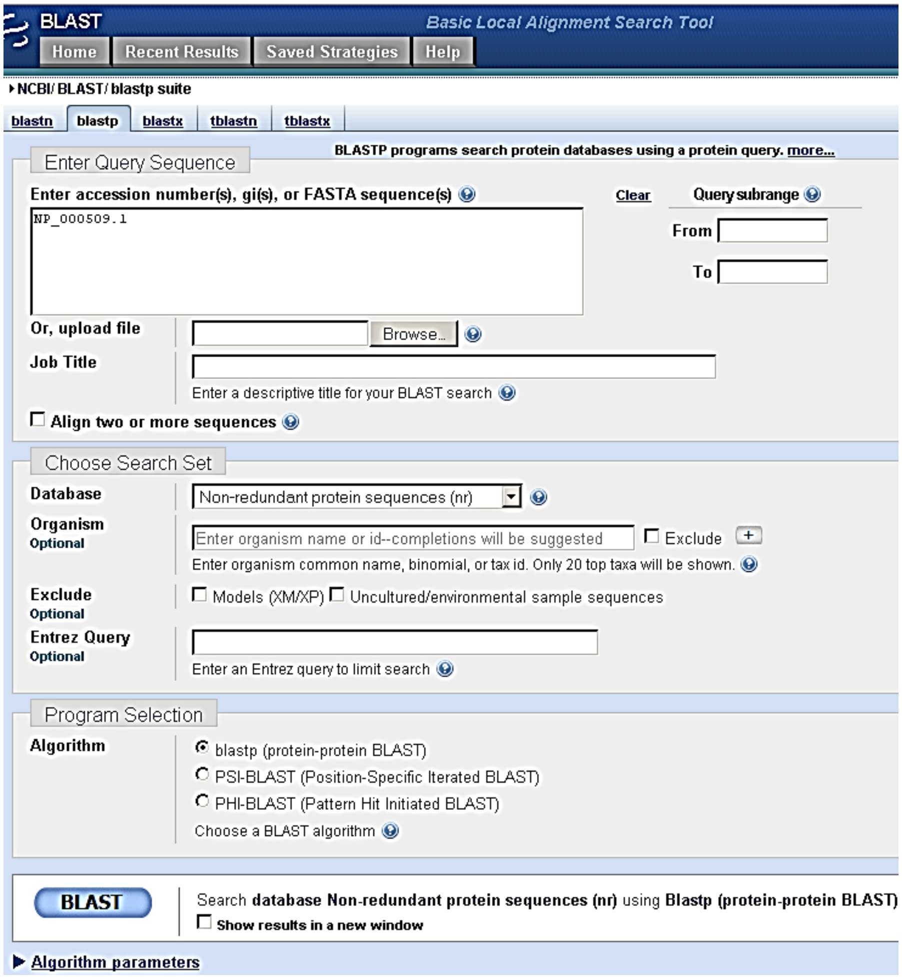
Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query



Step3: search parameter



BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
NP_000509.1
```

Clear

Query subrange

From

To

Or, upload file Browse...

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query Optional:

Enter an Entrez query to limit search

Program Selection

Algorithm

● blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

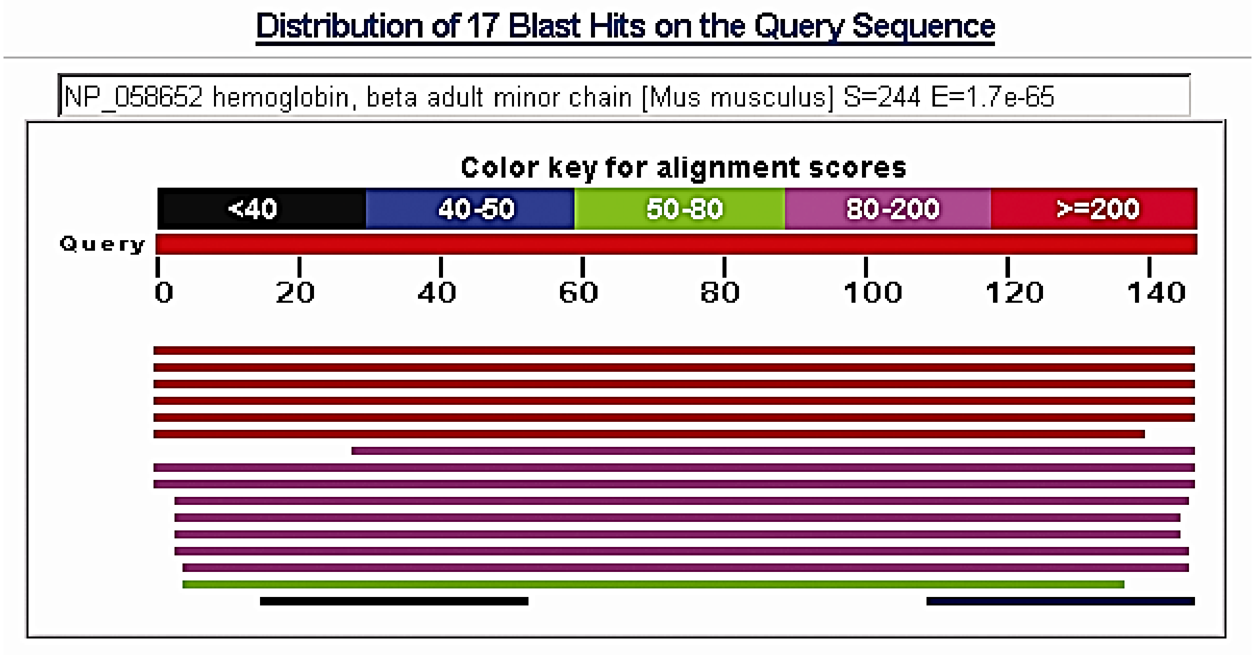
BLAST

Search **database Non-redundant protein sequences (nr)** using **Blastp (protein-protein BLAST)**

□ **Show results in a new window**

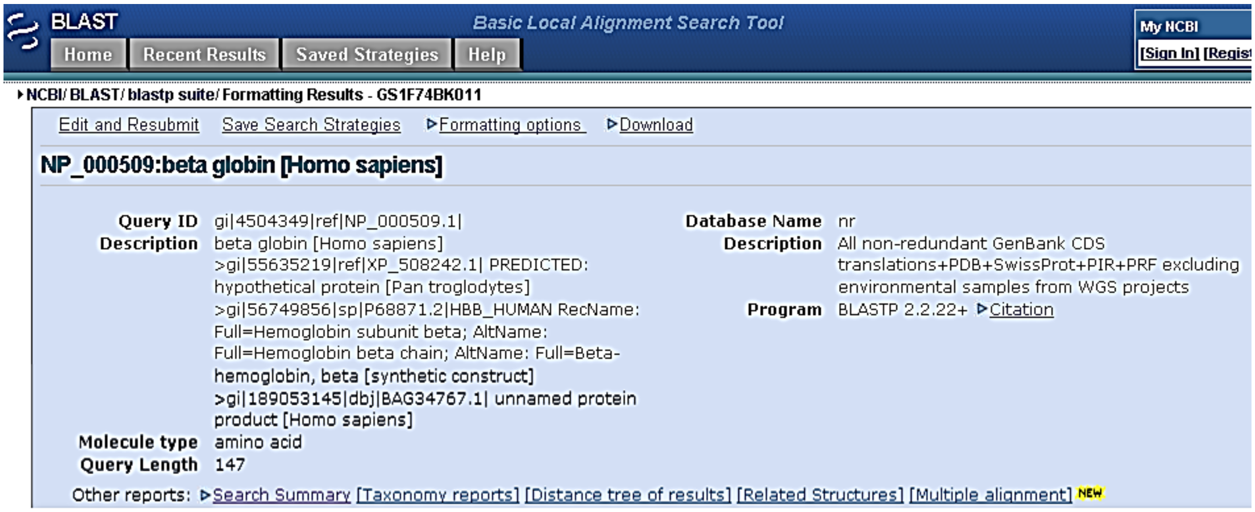
▶ Algorithm parameters

Step4: tabulated search results



Distribution of 17 Blast Hits on the Query Sequence

NP_058652 hemoglobin, beta adult minor chain [Mus musculus] S=244 E=1.7e-65



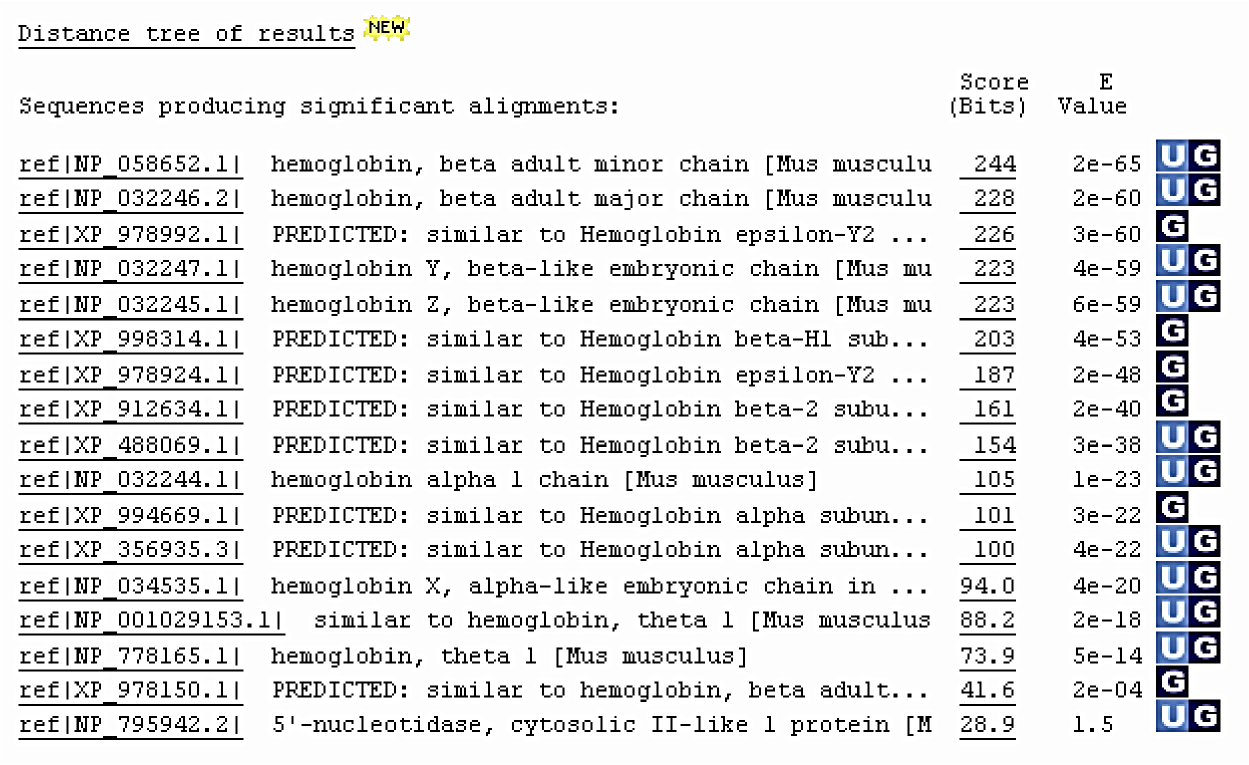
▸NCBI/ BLAST/ blastp suite/ Formatting Results - GS1F74BK011

Edit and Resubmit Save Search Strategies ▷Formatting options ▷Download

NP_000509:beta globin [Homo sapiens]

Query ID	gi|4504349|ref|NP_000509.1|	**Database Name**	nr
Description	beta globin [Homo sapiens] >gi|55635219|ref|XP_508242.1| PREDICTED: hypothetical protein [Pan troglodytes] >gi|56749856|sp|P68871.2|HBB_HUMAN RecName: Full=Hemoglobin subunit beta; AltName: Full=Hemoglobin beta chain; AltName: Full=Beta-hemoglobin, beta [synthetic construct] >gi|189053145|dbj|BAG34767.1| unnamed protein product [Homo sapiens]	**Description**	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
		Program	BLASTP 2.2.22+ ▷Citation
Molecule type	amino acid		
Query Length	147		

Other reports: ▷Search Summary [Taxonomy reports] [Distance tree of results] [Related Structures] [Multiple alignment] NEW



```
Distance tree of results NEW

                                                                 Score     E
Sequences producing significant alignments:                      (Bits)  Value

ref|NP_058652.1|  hemoglobin, beta adult minor chain [Mus musculu   244    2e-65 U G
ref|NP_032246.2|  hemoglobin, beta adult major chain [Mus musculu   228    2e-60 U G
ref|XP_978992.1|  PREDICTED: similar to Hemoglobin epsilon-Y2 ...   226    3e-60 G
ref|NP_032247.1|  hemoglobin Y, beta-like embryonic chain [Mus mu   223    4e-59 U G
ref|NP_032245.1|  hemoglobin Z, beta-like embryonic chain [Mus mu   223    6e-59 U G
ref|XP_998314.1|  PREDICTED: similar to Hemoglobin beta-H1 sub...   203    4e-53 G
ref|XP_978924.1|  PREDICTED: similar to Hemoglobin epsilon-Y2 ...   187    2e-48 G
ref|XP_912634.1|  PREDICTED: similar to Hemoglobin beta-2 subu...   161    2e-40 G
ref|XP_488069.1|  PREDICTED: similar to Hemoglobin beta-2 subu...   154    3e-38 U G
ref|NP_032244.1|  hemoglobin alpha 1 chain [Mus musculus]           105    1e-23 U G
ref|XP_994669.1|  PREDICTED: similar to Hemoglobin alpha subun...   101    3e-22 G
ref|XP_356935.3|  PREDICTED: similar to Hemoglobin alpha subun...   100    4e-22 U G
ref|NP_034535.1|  hemoglobin X, alpha-like embryonic chain in ...  94.0    4e-20 U G
ref|NP_001029153.1|  similar to hemoglobin, theta 1 [Mus musculus  88.2    2e-18 U G
ref|NP_778165.1|  hemoglobin, theta 1 [Mus musculus]               73.9    5e-14 U G
ref|XP_978150.1|  PREDICTED: similar to hemoglobin, beta adult...  41.6    2e-04 G
ref|NP_795942.2|  5'-nucleotidase, cytosolic II-like 1 protein [M  28.9    1.5   U G
```

Figure 0.28 tabulated search results

Module 053: INTRODUCTION TO FASTA

For comparing two sequences we use pair wise sequencing and for the comparison of many sequences we use multiple sequence alignment. To handle the multiple alignments we perform alignment through smith-waterman algorithm for local one. And for global alignment we use Needleman-wunsch algorithm.

Both local and global alignments are the dynamic approaches. Many of the sequences are compared, which takes time and we use BLAST which is an approximate local alignment search tool BLAST compares a large number of sequences, quickly. FASTA took a similar approach.

Developed in 1988.it does Fast Alignment .Searches databases for query protein and nucleotide sequences. Was later improved upon in BLAST.

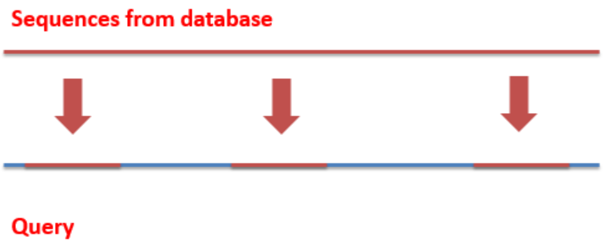
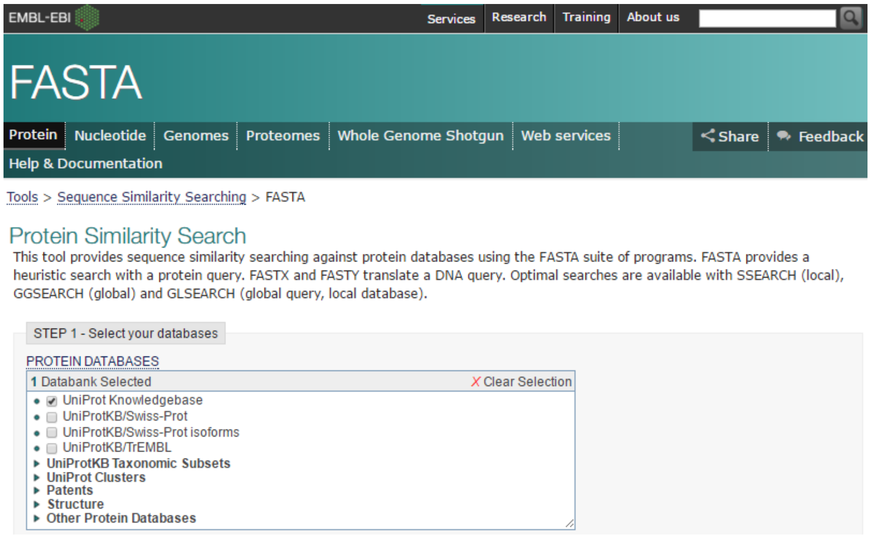


Figure 0.29 Regions of absolute identity



Protein Similarity Search

This tool provides sequence similarity searching against protein databases using the FASTA suite of programs. FASTA provides a heuristic search with a protein query. FASTX and FASTY translate a DNA query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

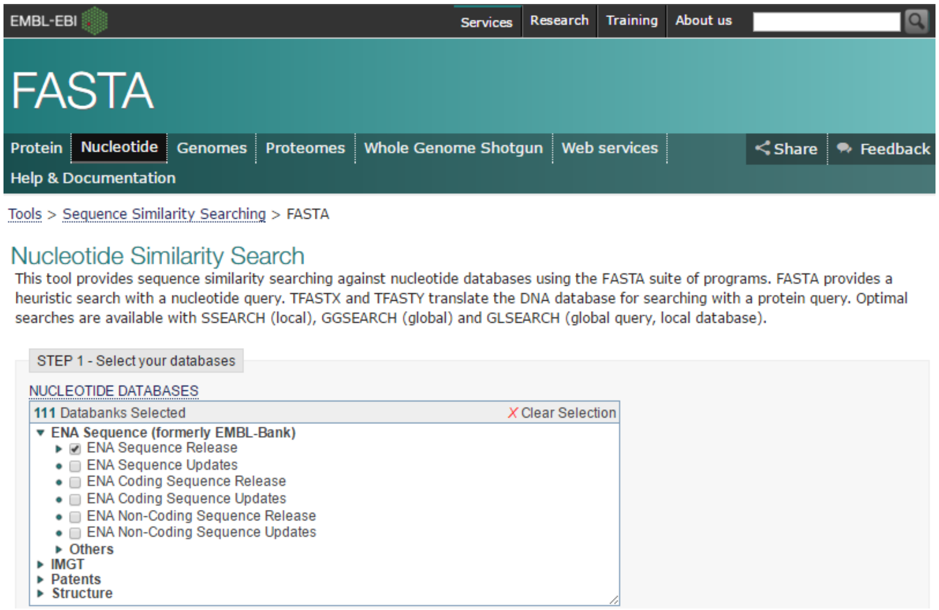
STEP 1 - Select your databases

PROTEIN DATABASES

1 Databank Selected   X Clear Selection

- ☑ UniProt Knowledgebase
- ☐ UniProtKB/Swiss-Prot
- ☐ UniProtKB/Swiss-Prot isoforms
- ☐ UniProtKB/TrEMBL
- ▸ UniProtKB Taxonomic Subsets
- ▸ UniProt Clusters
- ▸ Patents
- ▸ Structure
- ▸ Other Protein Databases

http://www.ebi.ac.uk/Tools/sss/fasta/



Tools > Sequence Similarity Searching > FASTA

Nucleotide Similarity Search

This tool provides sequence similarity searching against nucleotide databases using the FASTA suite of programs. FASTA provides a heuristic search with a nucleotide query. TFASTX and TFASTY translate the DNA database for searching with a protein query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

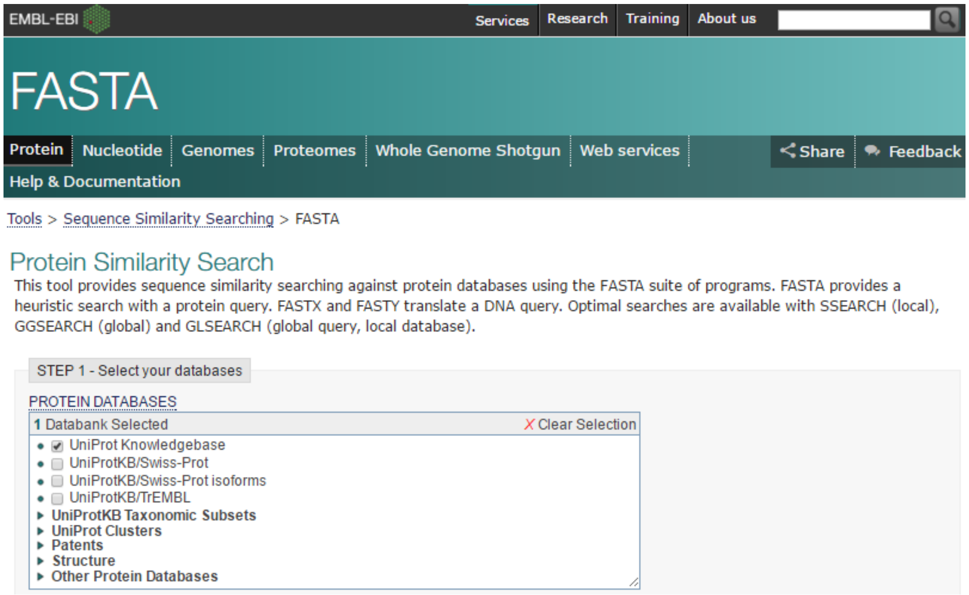
STEP 1 - Select your databases

NUCLEOTIDE DATABASES

111 Databanks Selected ✗ Clear Selection

- ENA Sequence (formerly EMBL-Bank)
  - ☑ ENA Sequence Release
  - □ ENA Sequence Updates
  - □ ENA Coding Sequence Release
  - □ ENA Coding Sequence Updates
  - □ ENA Non-Coding Sequence Release
  - □ ENA Non-Coding Sequence Updates
  - Others
- IMGT
- Patents
- Structure

Figure 0.30 Nucleotide FASTA



Tools > Sequence Similarity Searching > FASTA

Protein Similarity Search

This tool provides sequence similarity searching against protein databases using the FASTA suite of programs. FASTA provides a heuristic search with a protein query. FASTX and FASTY translate a DNA query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

STEP 1 - Select your databases

PROTEIN DATABASES

1 Databank Selected ✗ Clear Selection

- ☑ UniProt Knowledgebase
- □ UniProtKB/Swiss-Prot
- □ UniProtKB/Swiss-Prot isoforms
- □ UniProtKB/TrEMBL
- UniProtKB Taxonomic Subsets
- UniProt Clusters
- Patents
- Structure
- Other Protein Databases

Figure 0.31 Protein FASTA

Module 054: INTRODUCTION TO FASTA-II

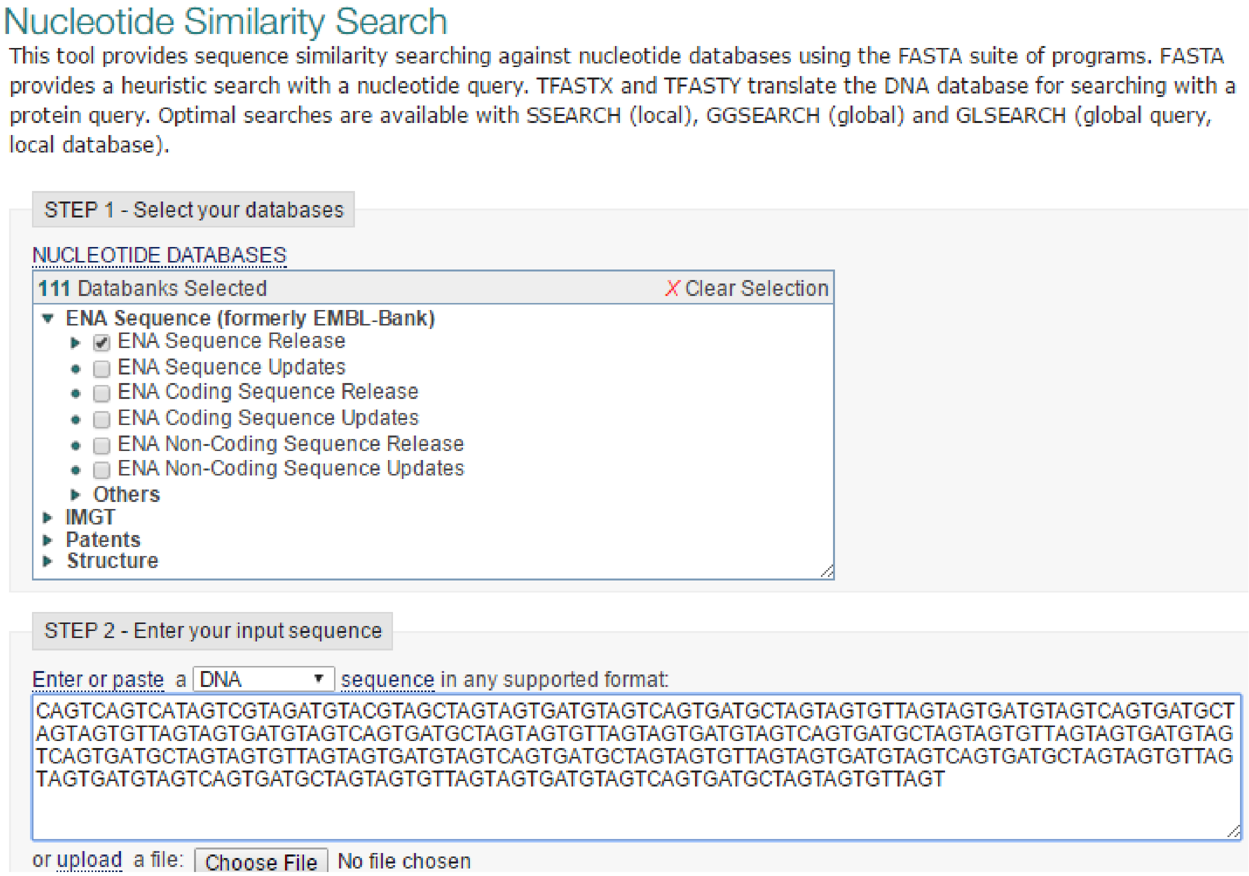
FASTA – Fast Alignment Algorithm. Classical global and local alignment algorithms are time consuming. FASTA achieves alignment by using short lengths of exact matches.

- **USES OF FASTA**

FASTA relies on aligning subsequences of absolute identity. Input to FASTA search can be in FASTA, EMBL, GenBank, PIR, NBRF, PHYLIP or UniProt formats

- **OUTPUT OF BLAST**

Results are output in visual format along with functional prediction. Makes table lists the sequence hits found along with scores. Users can click on each reported match to look at the details.



Nucleotide Similarity Search

This tool provides sequence similarity searching against nucleotide databases using the FASTA suite of programs. FASTA provides a heuristic search with a nucleotide query. TFASTX and TFASTY translate the DNA database for searching with a protein query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

STEP 1 - Select your databases

NUCLEOTIDE DATABASES

111 Databanks Selected X Clear Selection

- ▼ ENA Sequence (formerly EMBL-Bank)
  - ▶ ☑ ENA Sequence Release
  - ☐ ENA Sequence Updates
  - ☐ ENA Coding Sequence Release
  - ☐ ENA Coding Sequence Updates
  - ☐ ENA Non-Coding Sequence Release
  - ☐ ENA Non-Coding Sequence Updates
  - ▶ Others
- ▶ IMGT
- ▶ Patents
- ▶ Structure

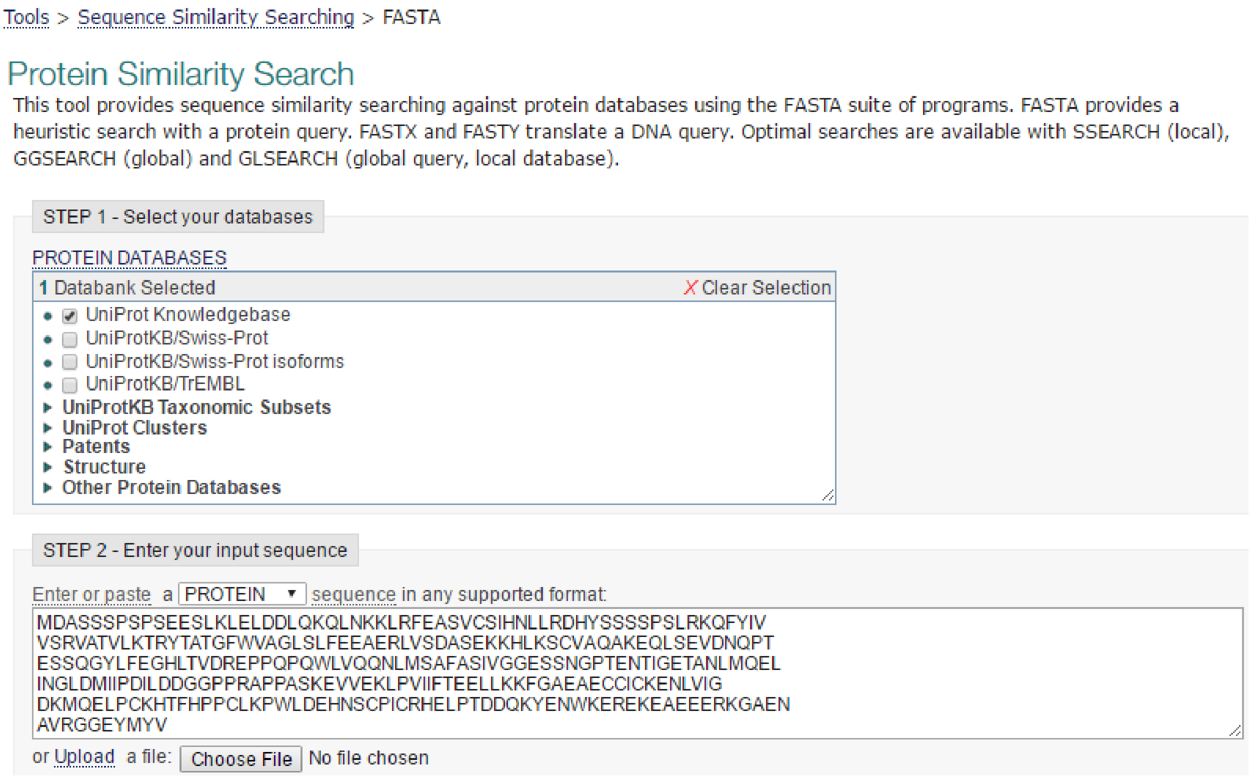
STEP 2 - Enter your input sequence

Enter or paste a DNA ▾ sequence in any supported format:

```
CAGTCAGTCATAGTCGTAGATGTACGTAGCTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCT
AGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAG
TCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAG
TAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGT
```

or upload a file: Choose File No file chosen

Figure 0.32 Input to FASTA: Gene IDs



Tools > Sequence Similarity Searching > FASTA

Protein Similarity Search

This tool provides sequence similarity searching against protein databases using the FASTA suite of programs. FASTA provides a heuristic search with a protein query. FASTX and FASTY translate a DNA query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

STEP 1 - Select your databases

PROTEIN DATABASES

1 Databank Selected X Clear Selection

- ☑ UniProt Knowledgebase
- □ UniProtKB/Swiss-Prot
- □ UniProtKB/Swiss-Prot isoforms
- □ UniProtKB/TrEMBL
- ▸ **UniProtKB Taxonomic Subsets**
- ▸ **UniProt Clusters**
- ▸ **Patents**
- ▸ **Structure**
- ▸ **Other Protein Databases**

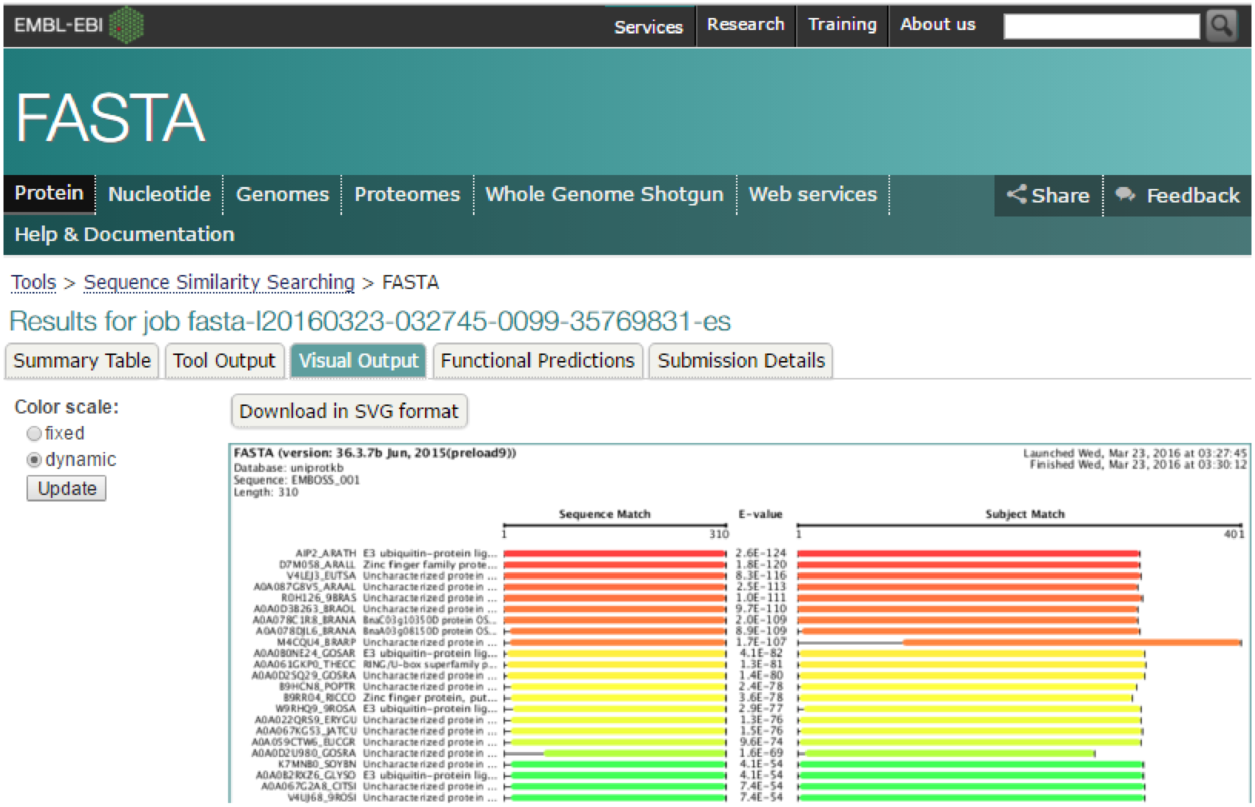
STEP 2 - Enter your input sequence

Enter or paste a PROTEIN ▾ sequence in any supported format:

```
MDASSSPSPSEESLKLELDDLQKQLNKKLRFEASVCSIHNLLRDHYSSSSPSLRKQFYIV
VSRVATVLKTRYTATGFWVAGLSLFEEAERLVSDASEKKHLKSCVAQAKEQLSEVDNQPT
ESSQGYLFEGHLTVDREPPQPQWLVQQNLMSAFASIVGGESSNGPTENTIGETANLMQEL
INGLDMIIPDILDDGGPPRAPPASKEVVEKLPVIIFTEELLKKFGAEAECCICKENLVIG
DKMQELPCKHTFHPPCLKPWLDEHNSCPICRHELPTDDQKYENWKEREKEAEEERKGAEN
AVRGGEYMYV
```

or Upload a file: Choose File No file chosen

Figure 0.33 Input to FASTA: Protein Sequence



Tools > Sequence Similarity Searching > FASTA

Results for job fasta-I20160323-032745-0099-35769831-es

Color scale:

○ fixed

● dynamic

Update

Download in SVG format

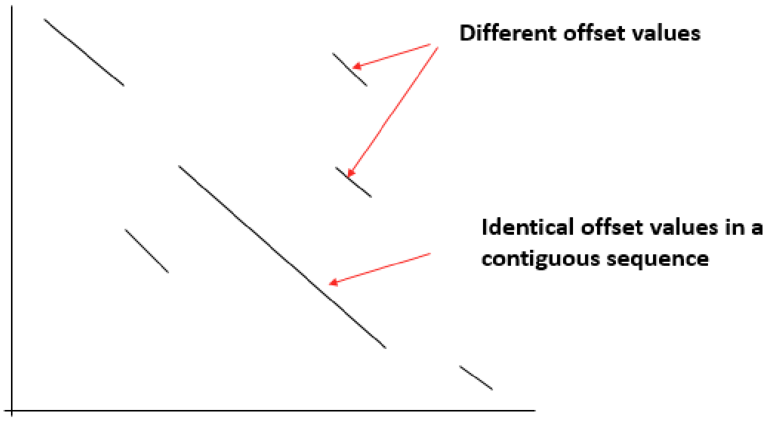
```
FASTA (version: 36.3.7b Jun, 2015(preload9))          Launched Wed, Mar 23, 2016 at 03:27:45
Database: uniprotkb                                   Finished Wed, Mar 23, 2016 at 03:30:12
Sequence: EMBOSS_001
Length: 310

                                        Sequence Match   E-value   Subject Match
                                        1          310             1          401
      AIP2_ARATH  E3 ubiquitin-protein lig...             2.6E-124
     D7M058_ARALL  Zinc finger family prote...            1.8E-120
     V4LEJ3_EUTSA  Uncharacterized protein ...            8.3E-116
 A0A087G8V5_ARAAL  Uncharacterized protein ...            2.5E-113
     R0H126_9BRAS  Uncharacterized protein ...            1.0E-111
 A0A0D3B263_BRAOL  Uncharacterized protein ...            9.7E-110
 A0A078C1R8_BRANA  BnaC03g10350D protein OS...            2.0E-109
 A0A078DJL6_BRANA  BnaA03g08150D protein OS...            8.9E-109
     M4CQU4_BRARP  Uncharacterized protein ...            1.7E-107
A0A0B0NE24_GOSAR  E3 ubiquitin-protein lig...             4.1E-82
A0A061GKP0_THECC  RING/U-box superfamily p...             1.3E-81
A0A0D2SQ29_GOSRA  Uncharacterized protein ...             1.4E-80
     B9HCN8_POPTR  Uncharacterized protein ...            2.4E-78
     B9RR04_RICCO  Zinc finger protein, put...            3.6E-78
     W9RHQ9_9ROSA  E3 ubiquitin-protein lig...            2.9E-77
A0A022QRS9_ERYGU  Uncharacterized protein ...             1.3E-76
A0A067KG53_JATCU  Uncharacterized protein ...             1.5E-76
A0A059CTW6_EUCGR  Uncharacterized protein ...             9.6E-74
A0A0D2U980_GOSRA  Uncharacterized protein ...             1.6E-69
     K7MNB0_SOYBN  Uncharacterized protein ...            4.1E-54
A0A0B2RXZ6_GLYSO  E3 ubiquitin-protein lig...             4.1E-54
A0A067G2A8_CITSI  Uncharacterized protein ...             7.4E-54
     V4UJ68_9ROSI  Uncharacterized protein ...            7.4E-54
```

Figure 0.34 Results from FASTA

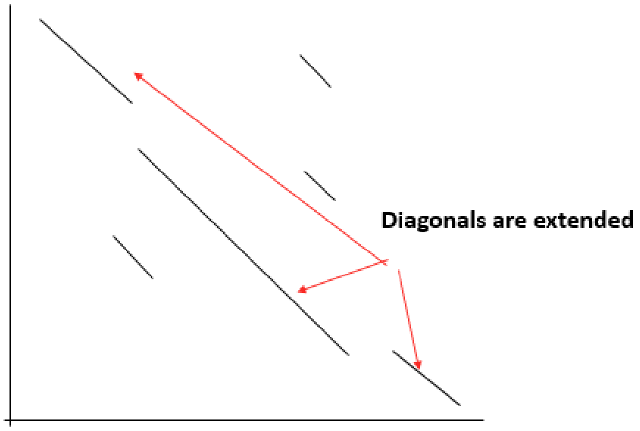
Module 055: FASTA ALGORITHM

FASTA can search sequence databases and identify unknown sequences by comparing them to the known sequence databases. This can help obtain information on the parent organism, function and evolutionary history.

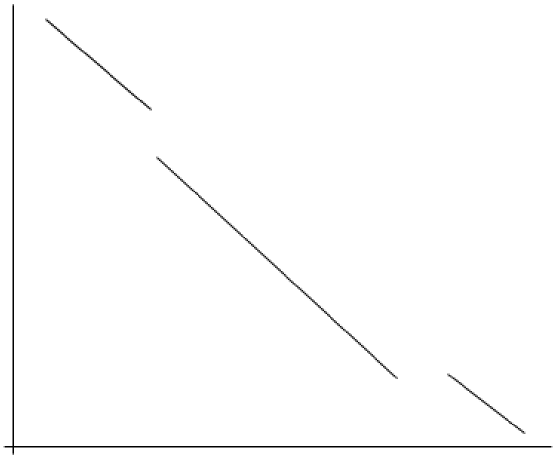
STEP1: Local regions of identity are found



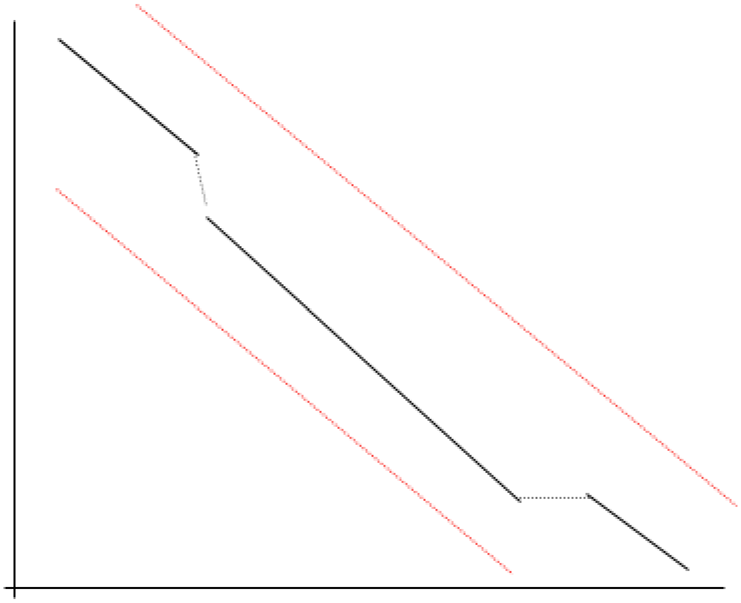
STEP2: Rescore the local regions using PAM or BLOSUM matrix



STEP3: Eliminate short diagonals below a cutoff score



STEP4: Create a gapped alignment in a narrow segment and then perform Smith Watermann alignment



Module 056: TYPES OF FASTA

There are six types of FASTS:

- **fasts35**

Compare unordered peptides to a protein sequence database

- **fastm35**

Compare ordered peptides (or short DNA sequences) to a protein (DNA) sequence database

- **Fasta35**

Scan a protein or DNA sequence library for similar sequences

- **Fastx35**

Compare a translated DNA sequence (6 ORFs) to a protein sequence database

- **tfastx35**

Compare a protein sequence to a DNA sequence database (6 ORFs)

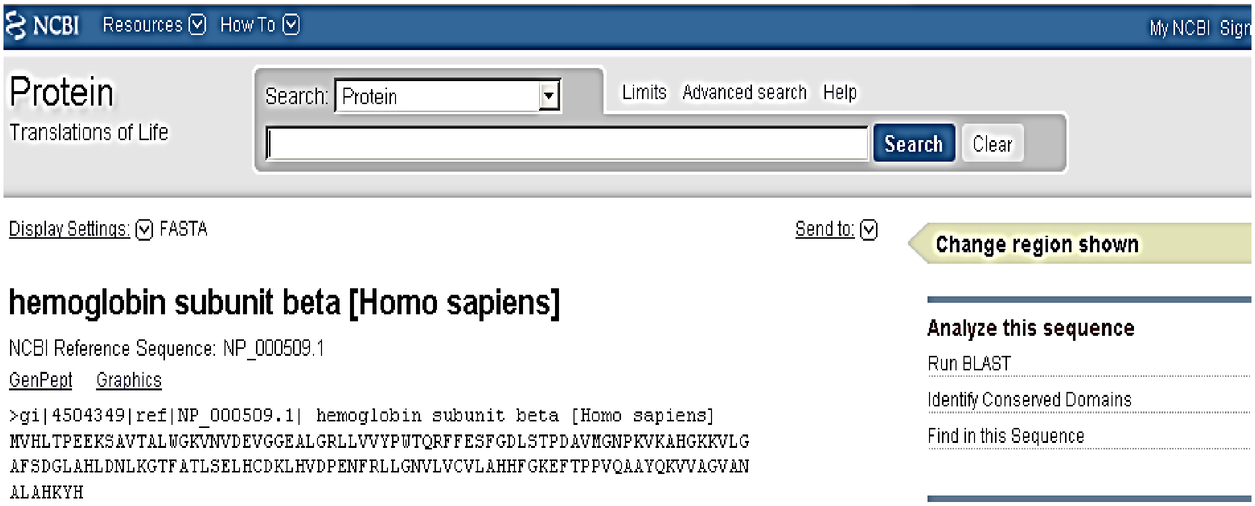
- **fasty35**

Compare a DNA sequence (6ORFs) to a protein sequence database

FASTA performs quick alignments on biological sequences. Several types of FASTA exist which can assist in comparing DNA/RNA/Protein sequences with each other

Module 057: SUMMERY OF FASTA

FASTA can briskly perform sequence search databases if given a query sequence. Multiple types of FASTA exist which assist in aligning DNA/RNA/Protein sequences



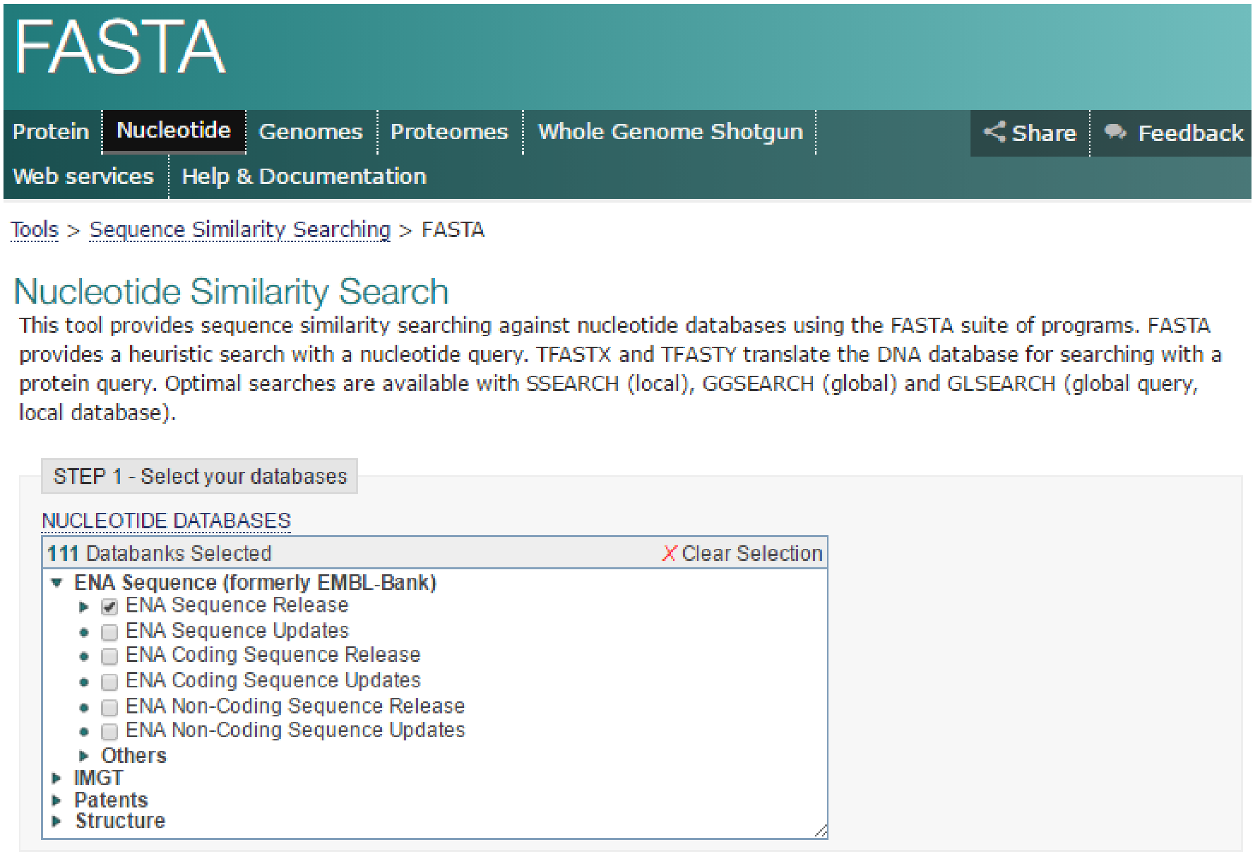
hemoglobin subunit beta [Homo sapiens]

NCBI Reference Sequence: NP_000509.1

GenPept Graphics

```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

Figure 0.35 Step 1: Obtain a query sequence



Nucleotide Similarity Search

This tool provides sequence similarity searching against nucleotide databases using the FASTA suite of programs. FASTA provides a heuristic search with a nucleotide query. TFASTX and TFASTY translate the DNA database for searching with a protein query. Optimal searches are available with SSEARCH (local), GGSEARCH (global) and GLSEARCH (global query, local database).

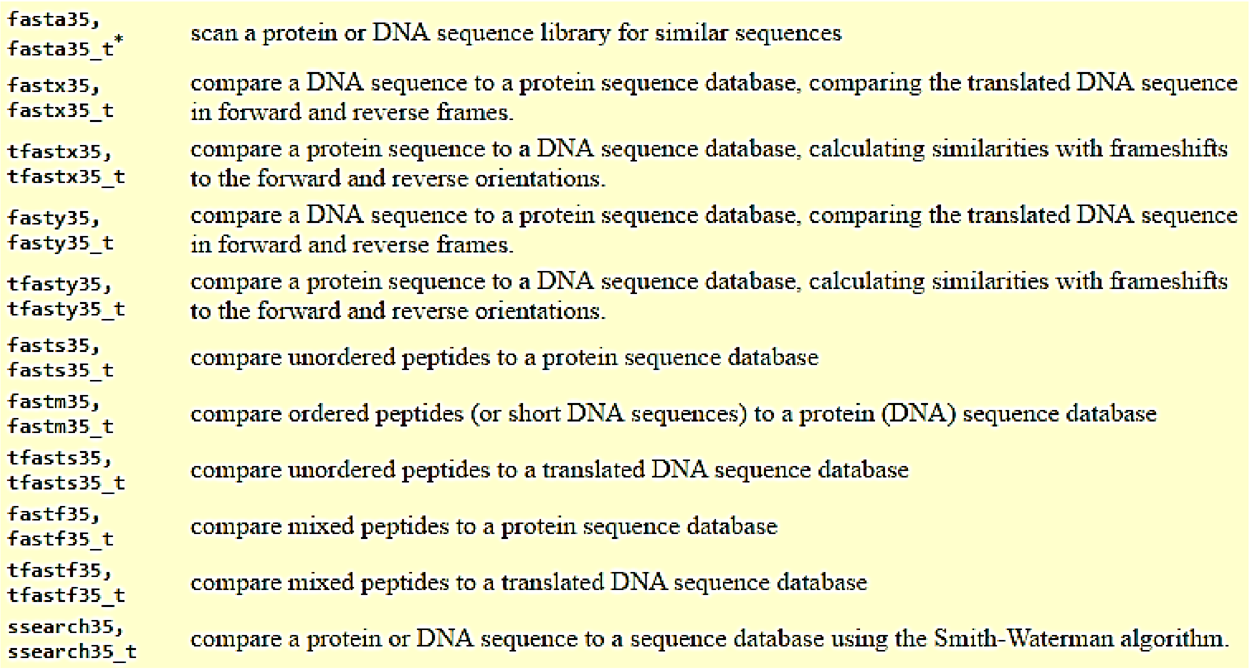
STEP 1 - Select your databases

NUCLEOTIDE DATABASES

111 Databanks Selected ✗ Clear Selection

- ENA Sequence (formerly EMBL-Bank)
  - ☑ ENA Sequence Release
  - □ ENA Sequence Updates
  - □ ENA Coding Sequence Release
  - □ ENA Coding Sequence Updates
  - □ ENA Non-Coding Sequence Release
  - □ ENA Non-Coding Sequence Updates
  - Others
- IMGT
- Patents
- Structure

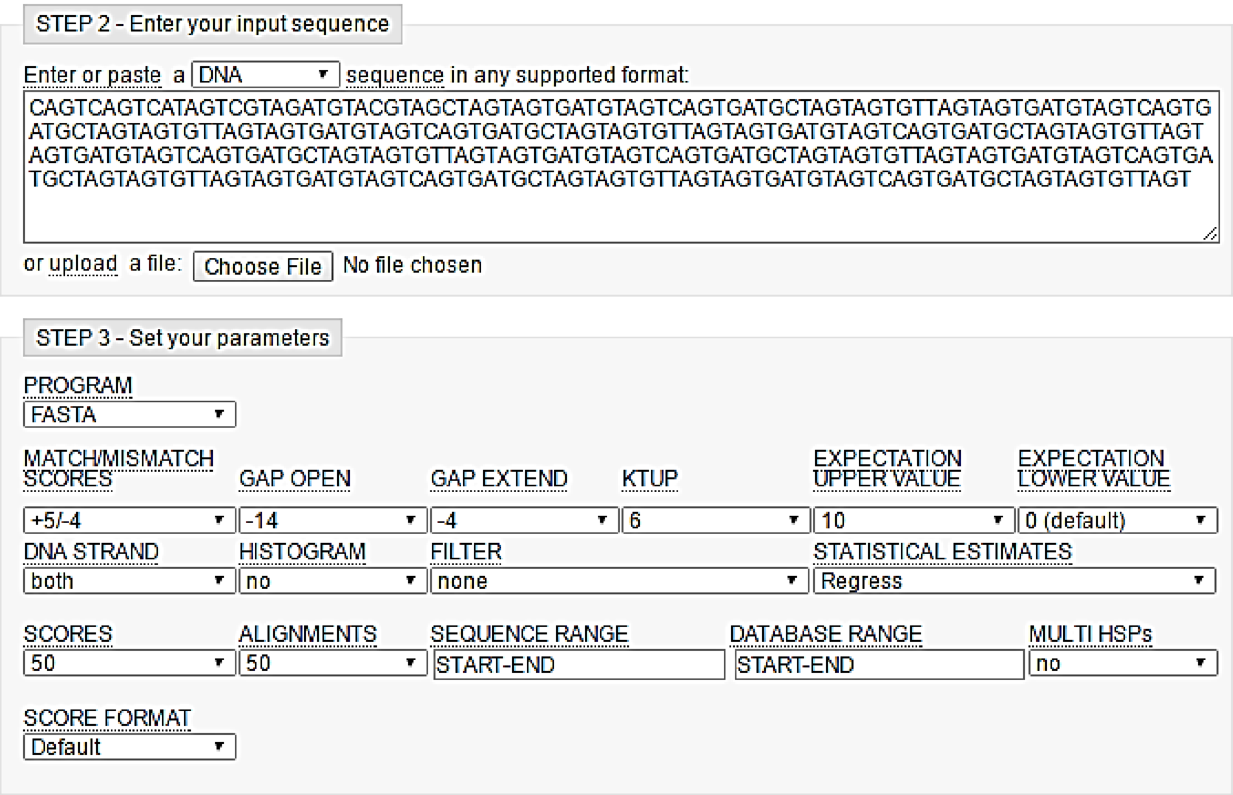
Figure 0.36 Step 2: Choose a type of FASTA



fasta35, fasta35_t*	scan a protein or DNA sequence library for similar sequences
fastx35, fastx35_t	compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.
tfastx35, tfastx35_t	compare a protein sequence to a DNA sequence database, calculating similarities with frameshifts to the forward and reverse orientations.
fasty35, fasty35_t	compare a DNA sequence to a protein sequence database, comparing the translated DNA sequence in forward and reverse frames.
tfasty35, tfasty35_t	compare a protein sequence to a DNA sequence database, calculating similarities with frameshifts to the forward and reverse orientations.
fasts35, fasts35_t	compare unordered peptides to a protein sequence database
fastm35, fastm35_t	compare ordered peptides (or short DNA sequences) to a protein (DNA) sequence database
tfasts35, tfasts35_t	compare unordered peptides to a translated DNA sequence database
fastf35, fastf35_t	compare mixed peptides to a protein sequence database
tfastf35, tfastf35_t	compare mixed peptides to a translated DNA sequence database
ssearch35, ssearch35_t	compare a protein or DNA sequence to a sequence database using the Smith-Waterman algorithm.

Figure 0.37 Type of FASTA

http://fasta.bioch.virginia.edu/fasta_docs/fasta35.shtml



STEP 2 - Enter your input sequence

Enter or paste a DNA ▾ sequence in any supported format:

```
CAGTCAGTCATAGTCGTAGATGTACGTAGCTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTG
ATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGT
AGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGA
TGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGTAGTGATGTAGTCAGTGATGCTAGTAGTGTTAGT
```

or upload a file: Choose File No file chosen

STEP 3 - Set your parameters

PROGRAM: FASTA ▾

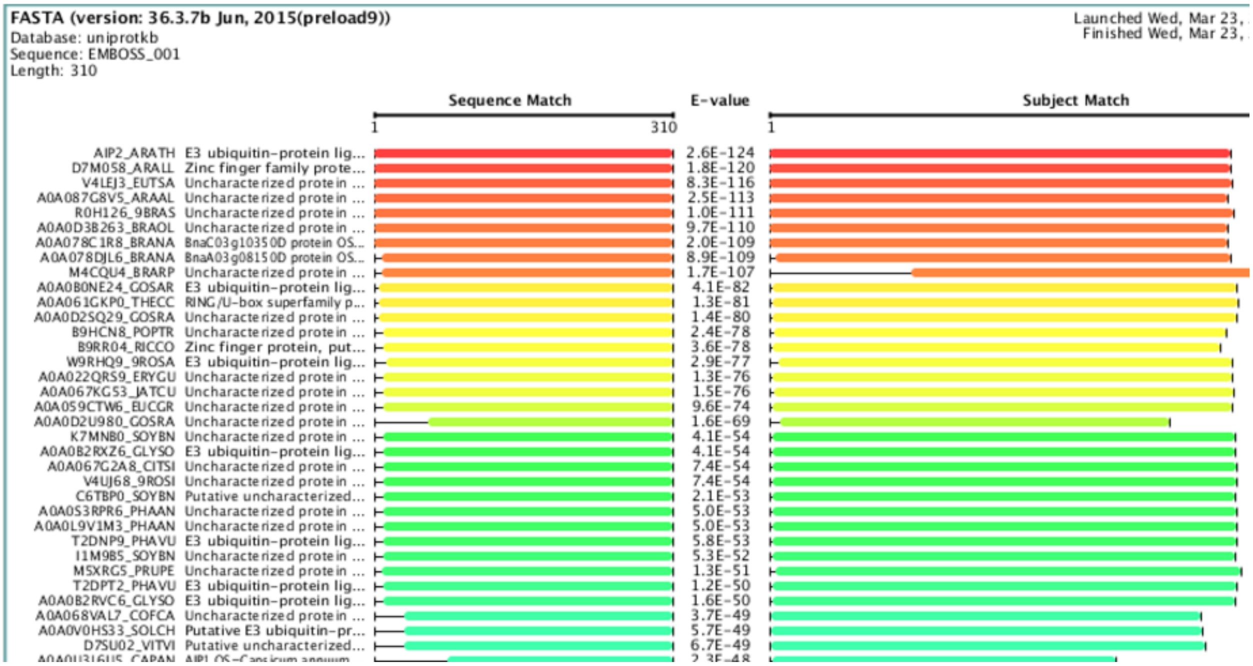
MATCH/MISMATCH SCORES	GAP OPEN	GAP EXTEND	KTUP	EXPECTATION UPPER VALUE	EXPECTATION LOWER VALUE
+5/-4 ▾	-14 ▾	-4 ▾	6 ▾	10 ▾	0 (default) ▾

DNA STRAND	HISTOGRAM	FILTER	STATISTICAL ESTIMATES
both ▾	no ▾	none ▾	Regress ▾

SCORES	ALIGNMENTS	SEQUENCE RANGE	DATABASE RANGE	MULTI HSPs
50 ▾	50 ▾	START-END	START-END	no ▾

SCORE FORMAT: Default ▾

Figure 0.38 Step 3: Setup Search Parameters

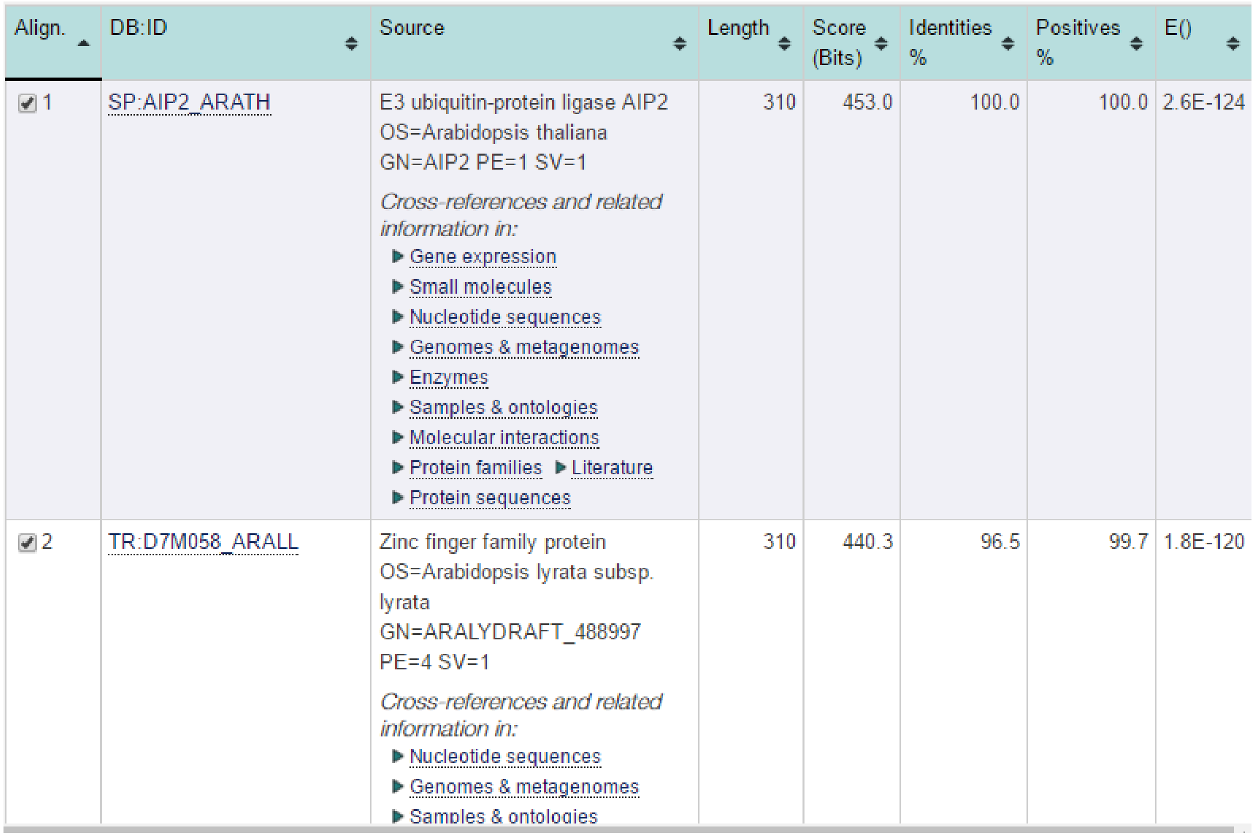


FASTA (version: 36.3.7b Jun, 2015(preload9))
Database: uniprotkb
Sequence: EMBOSS_001
Length: 310

Launched Wed, Mar 23, .
Finished Wed, Mar 23, .

		Sequence Match 1 310	E-value	Subject Match 1
AIP2_ARATH	E3 ubiquitin-protein lig...		2.6E-124	
D7M058_ARALL	Zinc finger family prote...		1.8E-120	
V4LEJ3_EUTSA	Uncharacterized protein ...		8.3E-116	
A0A087G8V5_ARAAL	Uncharacterized protein ...		2.5E-113	
R0H126_9BRAS	Uncharacterized protein ...		1.0E-111	
A0A0D3B263_BRAOL	Uncharacterized protein ...		9.7E-110	
A0A078C1R8_BRANA	BnaC03g10350D protein OS...		2.0E-109	
A0A078DJL6_BRANA	BnaA03g08150D protein OS...		8.9E-109	
M4CQU4_BRARP	Uncharacterized protein ...		1.7E-107	
A0A0B0NE24_GOSAR	E3 ubiquitin-protein lig...		4.1E-82	
A0A061GKP0_THECC	RING/U-box superfamily p...		1.3E-81	
A0A0D2SQ29_GOSRA	Uncharacterized protein ...		1.4E-80	
B9HCN8_POPTR	Uncharacterized protein ...		2.4E-78	
B9RR04_RICCO	Zinc finger protein, put...		3.6E-78	
W9RHQ9_9ROSA	E3 ubiquitin-protein lig...		2.9E-77	
A0A022QRS9_ERYGU	Uncharacterized protein ...		1.3E-76	
A0A067KG53_JATCU	Uncharacterized protein ...		1.5E-76	
A0A059CTW6_EUCGR	Uncharacterized protein ...		9.6E-74	
A0A0D2U980_GOSRA	Uncharacterized protein ...		1.6E-69	
K7MNB0_SOYBN	Uncharacterized protein ...		4.1E-54	
A0A0B2RXZ6_GLYSO	E3 ubiquitin-protein lig...		4.1E-54	
A0A067G2A8_CITSI	Uncharacterized protein ...		7.4E-54	
V4UJ68_9ROSI	Uncharacterized protein ...		7.4E-54	
C6TBP0_SOYBN	Putative uncharacterized...		2.1E-53	
A0A0S3RPR6_PHAAN	Uncharacterized protein ...		5.0E-53	
A0A0L9V1M3_PHAAN	Uncharacterized protein ...		5.0E-53	
T2DNP9_PHAVU	E3 ubiquitin-protein lig...		5.8E-53	
I1M9B5_SOYBN	Uncharacterized protein ...		5.3E-52	
M5XRG5_PRUPE	Uncharacterized protein ...		1.3E-51	
T2DPT2_PHAVU	E3 ubiquitin-protein lig...		1.2E-50	
A0A0B2RVC6_GLYSO	E3 ubiquitin-protein lig...		1.6E-50	
A0A068VAL7_COFCA	Uncharacterized protein ...		3.7E-49	
A0A0V0HS33_SOLCH	Putative E3 ubiquitin-pr...		5.7E-49	
D7SU02_VITVI	Putative uncharacterized...		6.7E-49	
A0A0U3I6U5_CAPAN	AIP1 OS=Capsicum annuum ...		2.3E-48	

Figure 0.39 Step 4: Tabulated Search Results



Align.	DB:ID	Source	Length	Score (Bits)	Identities %	Positives %	E()
☑ 1	SP:AIP2_ARATH	E3 ubiquitin-protein ligase AIP2 OS=Arabidopsis thaliana GN=AIP2 PE=1 SV=1 *Cross-references and related information in:* ▶ Gene expression ▶ Small molecules ▶ Nucleotide sequences ▶ Genomes & metagenomes ▶ Enzymes ▶ Samples & ontologies ▶ Molecular interactions ▶ Protein families ▶ Literature ▶ Protein sequences	310	453.0	100.0	100.0	2.6E-124
☑ 2	TR:D7M058_ARALL	Zinc finger family protein OS=Arabidopsis lyrata subsp. lyrata GN=ARALYDRAFT_488997 PE=4 SV=1 *Cross-references and related information in:* ▶ Nucleotide sequences ▶ Genomes & metagenomes ▶ Samples & ontologies	310	440.3	96.5	99.7	1.8E-120

Figure 0.40 Tabulated data

Module 058: BIOLOGICAL DATABASE AND ONLINE TOOLS

All molecular information of RNA, DNA, Proteins have need to be stored and retrieved. Sequences are obtained from genome sequencing and mass spectrometry

Structures are obtained from X-Ray Crystallography, Atomic Force Microscopy & Nuclear Magnetic Resonance Spectroscopy.

Vast amounts of such data exists. Moreover, this data is rapidly accumulating. Online Databases are formed to store and share this data.

- **OBJECTIVE**

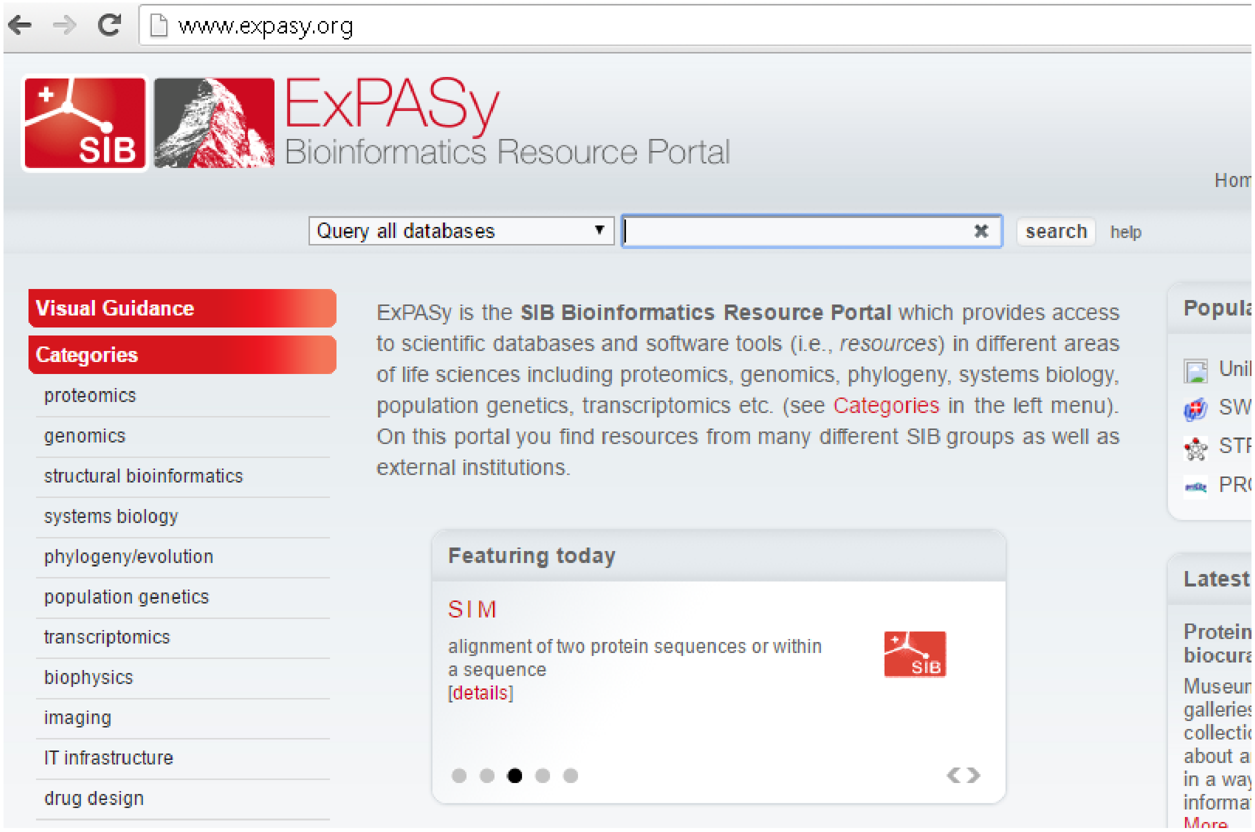
- Make biological data available to scientists in computer-readable form
- For handling, sharing and analysis of the data
- The best way to share is to keep this data on the web

Several sequence, structure and molecular interaction databases exist. These are available online on the web. Users can freely access and download such data

Module 059: EXPASY

It is developed by Swiss Bioinformatics Institute (SIB). Website provides access to databases and tools Proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. can be searched.

http://www.expasy.org/



ExPASy is the **SIB Bioinformatics Resource Portal** which provides access to scientific databases and software tools (i.e., *resources*) in different areas of life sciences including proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc. (see Categories in the left menu). On this portal you find resources from many different SIB groups as well as external institutions.

SIM

alignment of two protein sequences or within a sequence
[details]

Protein biocura

Museur galleries collectio about a in a way informa More

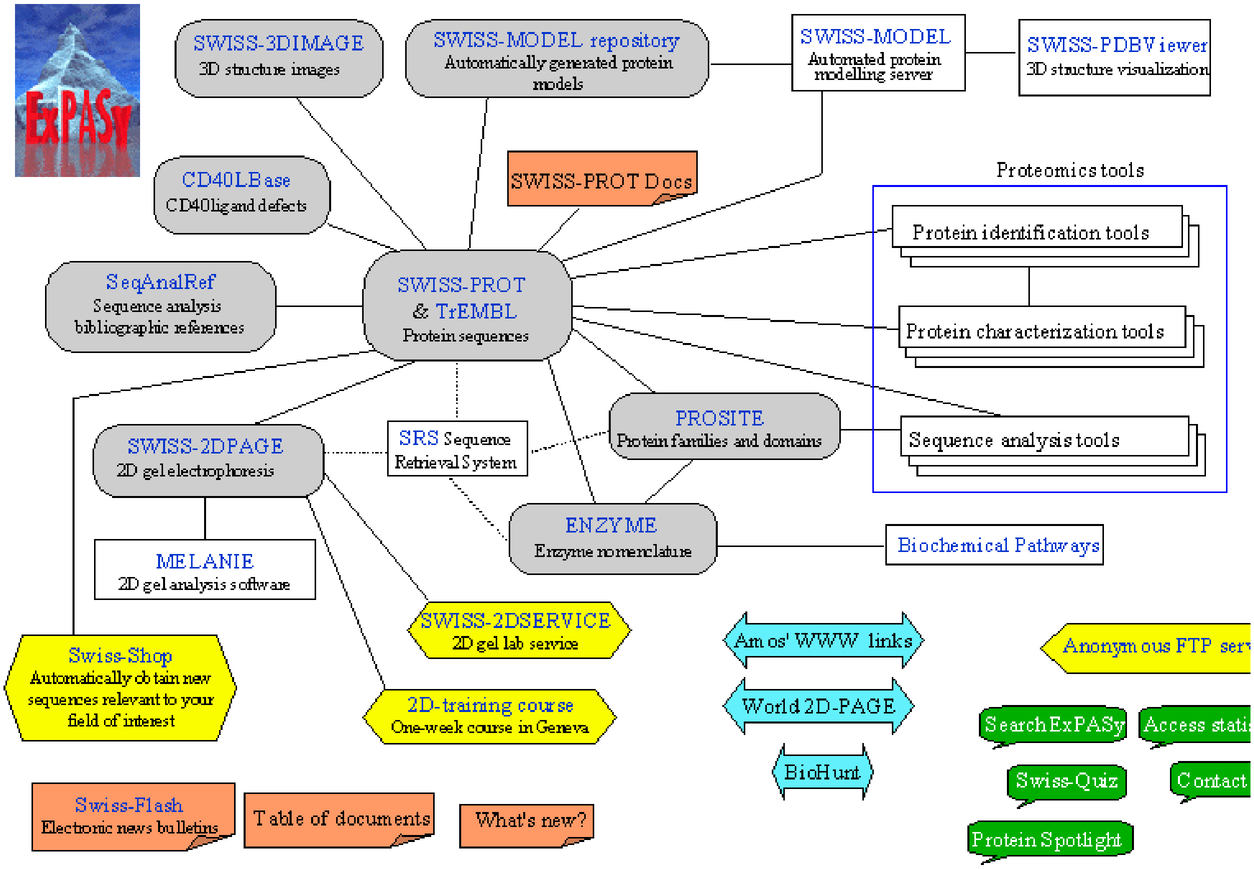
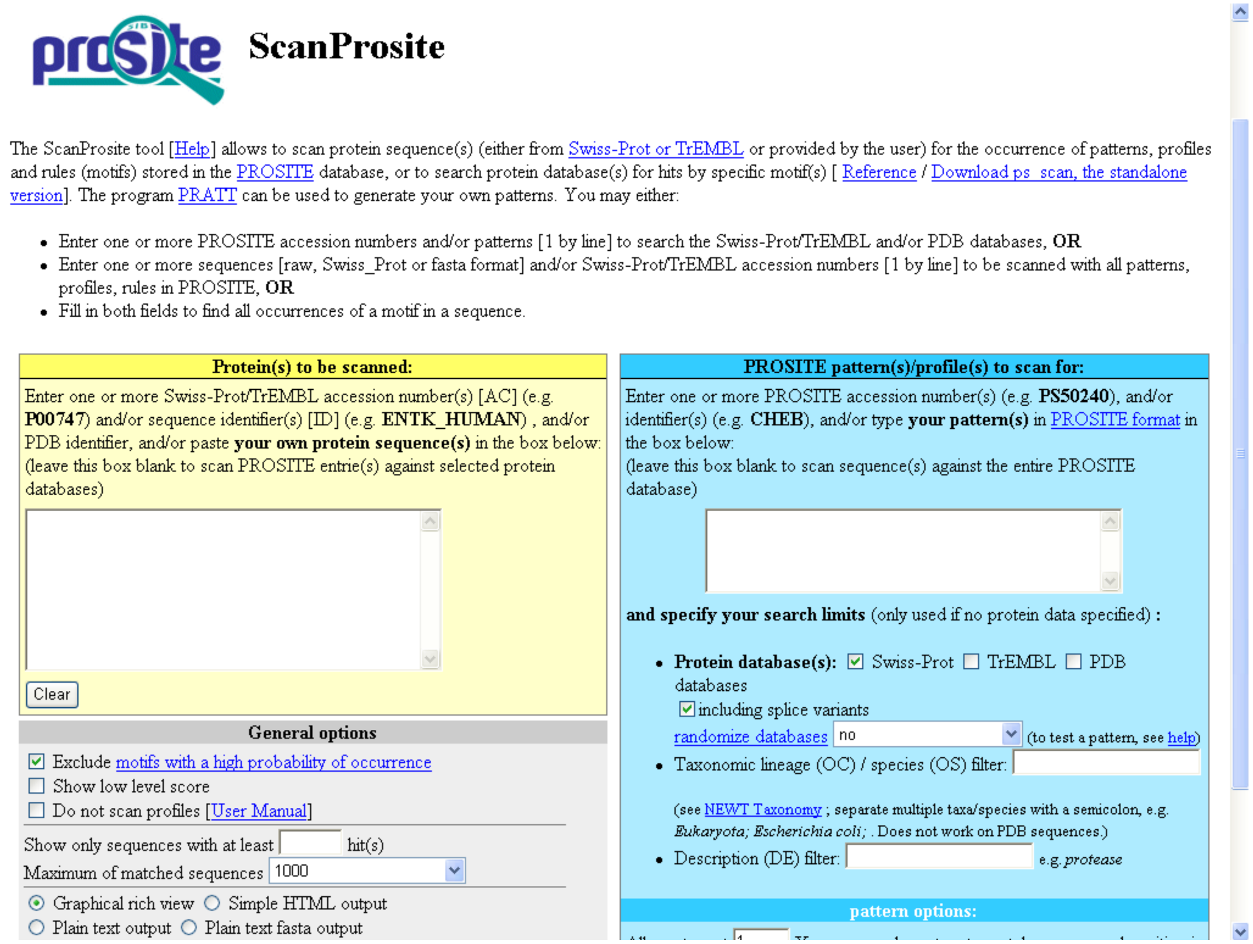


Figure 0.41 flowchart



ScanProsite

The ScanProsite tool [Help] allows to scan protein sequence(s) (either from Swiss-Prot or TrEMBL or provided by the user) for the occurrence of patterns, profiles and rules (motifs) stored in the PROSITE database, or to search protein database(s) for hits by specific motif(s) [Reference / Download ps_scan, the standalone version]. The program PRATT can be used to generate your own patterns. You may either:

- Enter one or more PROSITE accession numbers and/or patterns [1 by line] to search the Swiss-Prot/TrEMBL and/or PDB databases, **OR**
- Enter one or more sequences [raw, Swiss_Prot or fasta format] and/or Swiss-Prot/TrEMBL accession numbers [1 by line] to be scanned with all patterns, profiles, rules in PROSITE, **OR**
- Fill in both fields to find all occurrences of a motif in a sequence.

Protein(s) to be scanned:

Enter one or more Swiss-Prot/TrEMBL accession number(s) [AC] (e.g. **P00747**) and/or sequence identifier(s) [ID] (e.g. **ENTK_HUMAN**) , and/or PDB identifier, and/or paste **your own protein sequence(s)** in the box below: (leave this box blank to scan PROSITE entrie(s) against selected protein databases)

Clear

General options

☑ Exclude motifs with a high probability of occurrence

☐ Show low level score

☐ Do not scan profiles [User Manual]

Show only sequences with at least ____ hit(s)

Maximum of matched sequences 1000

● Graphical rich view ○ Simple HTML output

○ Plain text output ○ Plain text fasta output

PROSITE pattern(s)/profile(s) to scan for:

Enter one or more PROSITE accession number(s) (e.g. **PS50240**), and/or identifier(s) (e.g. **CHEB**), and/or type **your pattern(s)** in PROSITE format in the box below:
(leave this box blank to scan sequence(s) against the entire PROSITE database)

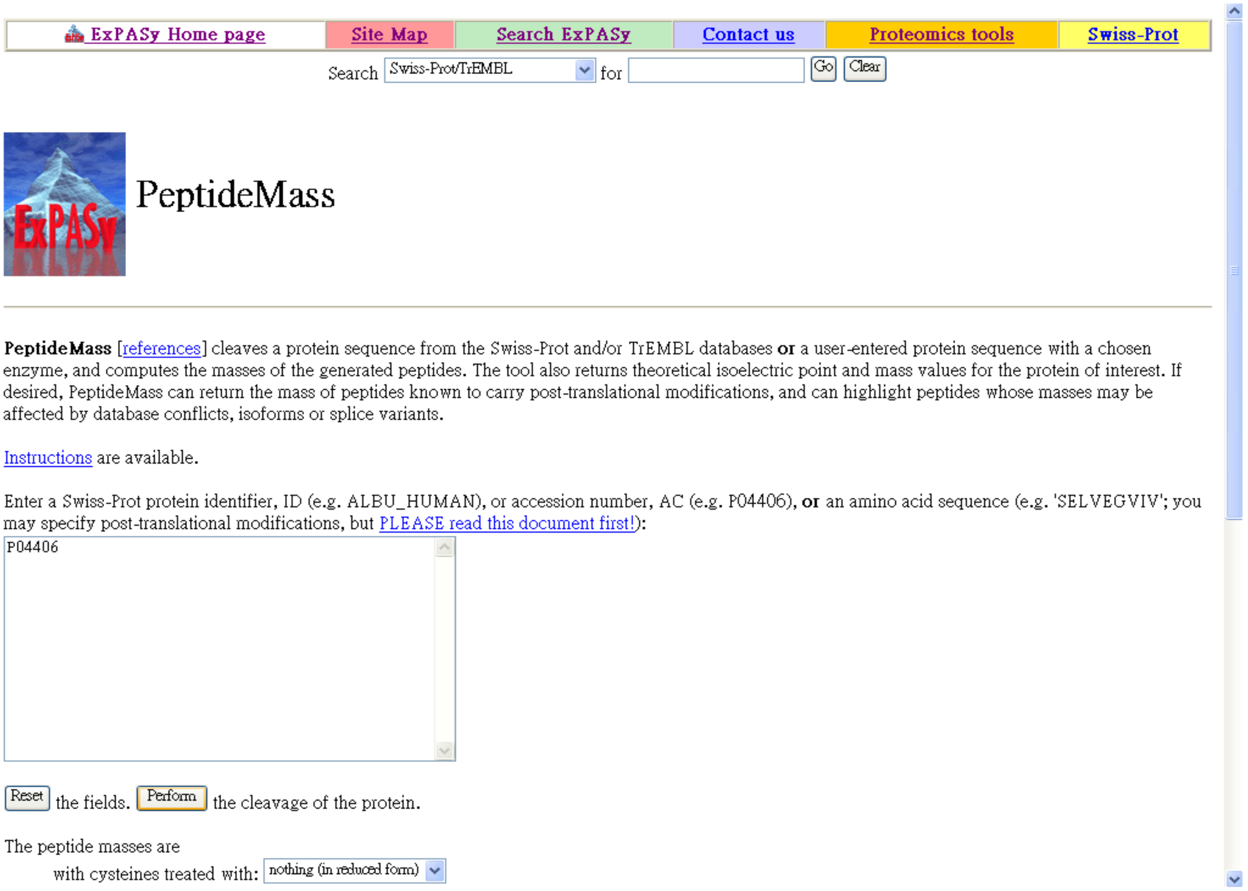
and specify your search limits (only used if no protein data specified) :

- **Protein database(s):** ☑ Swiss-Prot ☐ TrEMBL ☐ PDB databases
  ☑ including splice variants
  randomize databases no (to test a pattern, see help)
- Taxonomic lineage (OC) / species (OS) filter:

  (see NEWT Taxonomy ; separate multiple taxa/species with a semicolon, e.g. *Eukaryota; Escherichia coli;* . Does not work on PDB sequences.)
- Description (DE) filter: e.g. *protease*

pattern options:

Figure 0.42 prosite scanning section



PeptideMass

PeptideMass [references] cleaves a protein sequence from the Swiss-Prot and/or TrEMBL databases **or** a user-entered protein sequence with a chosen enzyme, and computes the masses of the generated peptides. The tool also returns theoretical isoelectric point and mass values for the protein of interest. If desired, PeptideMass can return the mass of peptides known to carry post-translational modifications, and can highlight peptides whose masses may be affected by database conflicts, isoforms or splice variants.

Instructions are available.

Enter a Swiss-Prot protein identifier, ID (e.g. ALBU_HUMAN), or accession number, AC (e.g. P04406), **or** an amino acid sequence (e.g. 'SELVEGVIV'; you may specify post-translational modifications, but PLEASE read this document first!):

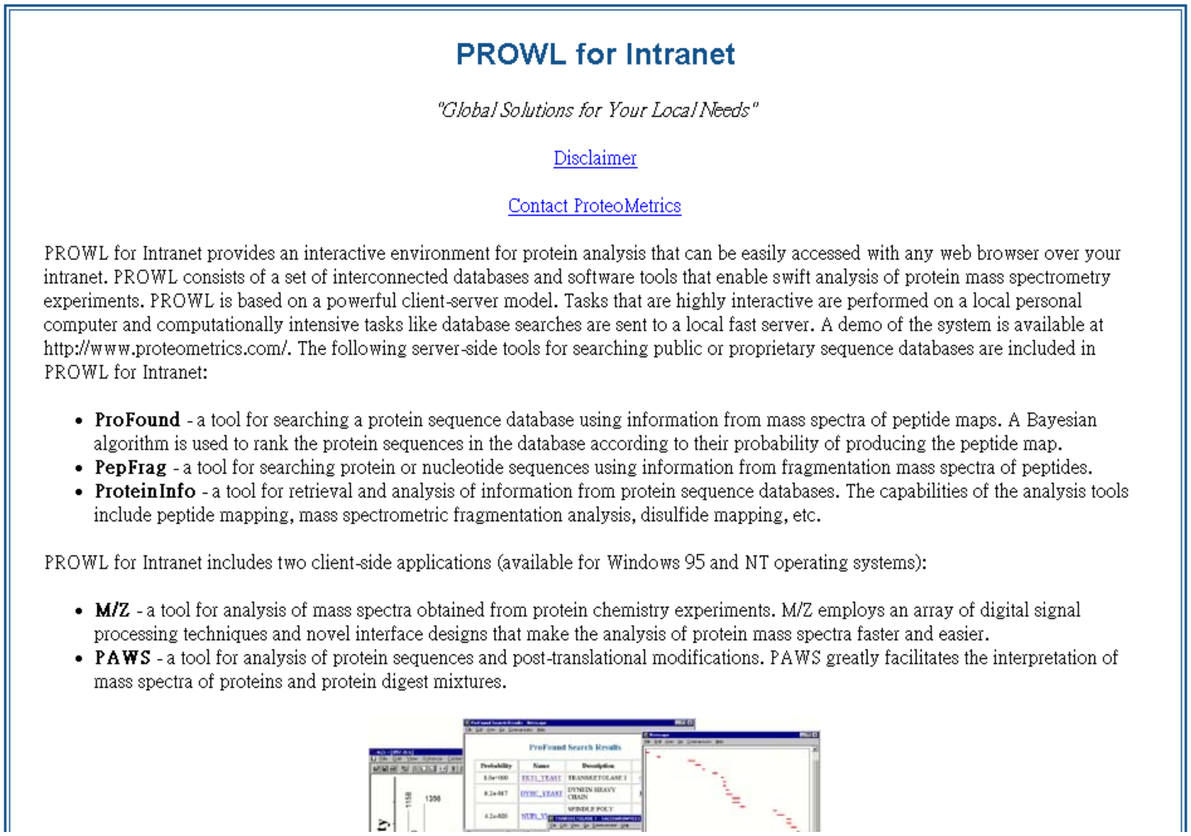
P04406

Reset the fields. Perform the cleavage of the protein.

The peptide masses are

with cysteines treated with: nothing (in reduced form)

Figure 0.43 peptide mass finding



PROWL for Intranet

"Global Solutions for Your Local Needs"

Disclaimer

Contact ProteoMetrics

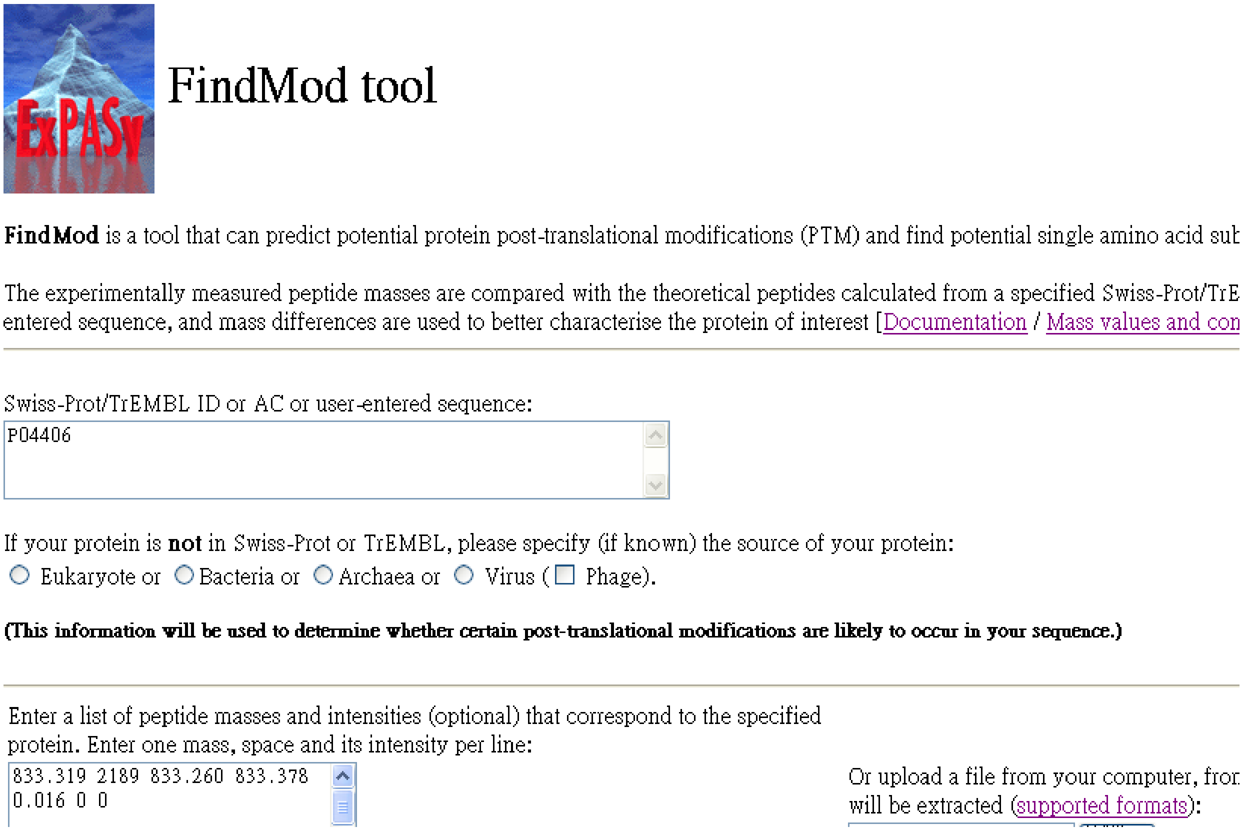
PROWL for Intranet provides an interactive environment for protein analysis that can be easily accessed with any web browser over your intranet. PROWL consists of a set of interconnected databases and software tools that enable swift analysis of protein mass spectrometry experiments. PROWL is based on a powerful client-server model. Tasks that are highly interactive are performed on a local personal computer and computationally intensive tasks like database searches are sent to a local fast server. A demo of the system is available at http://www.proteometrics.com/. The following server-side tools for searching public or proprietary sequence databases are included in PROWL for Intranet:

- **ProFound** - a tool for searching a protein sequence database using information from mass spectra of peptide maps. A Bayesian algorithm is used to rank the protein sequences in the database according to their probability of producing the peptide map.
- **PepFrag** - a tool for searching protein or nucleotide sequences using information from fragmentation mass spectra of peptides.
- **ProteinInfo** - a tool for retrieval and analysis of information from protein sequence databases. The capabilities of the analysis tools include peptide mapping, mass spectrometric fragmentation analysis, disulfide mapping, etc.

PROWL for Intranet includes two client-side applications (available for Windows 95 and NT operating systems):

- **M/Z** - a tool for analysis of mass spectra obtained from protein chemistry experiments. M/Z employs an array of digital signal processing techniques and novel interface designs that make the analysis of protein mass spectra faster and easier.
- **PAWS** - a tool for analysis of protein sequences and post-translational modifications. PAWS greatly facilitates the interpretation of mass spectra of proteins and protein digest mixtures.

Figure 0.44 for local use of protein sequencing



FindMod tool

FindMod is a tool that can predict potential protein post-translational modifications (PTM) and find potential single amino acid sub

The experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified Swiss-Prot/TrE entered sequence, and mass differences are used to better characterise the protein of interest [Documentation / Mass values and con

Swiss-Prot/TrEMBL ID or AC or user-entered sequence:

```
P04406
```

If your protein is **not** in Swiss-Prot or TrEMBL, please specify (if known) the source of your protein:

○ Eukaryote or ○ Bacteria or ○ Archaea or ○ Virus (□ Phage).

(This information will be used to determine whether certain post-translational modifications are likely to occur in your sequence.)

Enter a list of peptide masses and intensities (optional) that correspond to the specified protein. Enter one mass, space and its intensity per line:

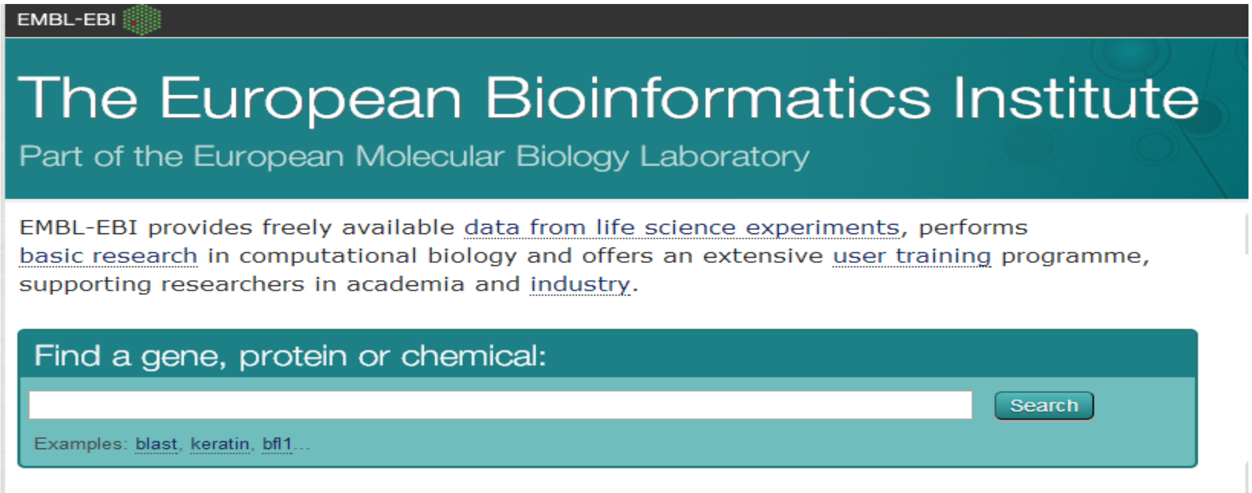
```
833.319 2189 833.260 833.378
0.016 0 0
```

Or upload a file from your computer, fror will be extracted (supported formats):

Figure 0.45 potential protein finding tool

Module 060: UNIPROT AND SWISSPROT

Both UniProt and SwissProt are the online database for proteins.



The European Bioinformatics Institute

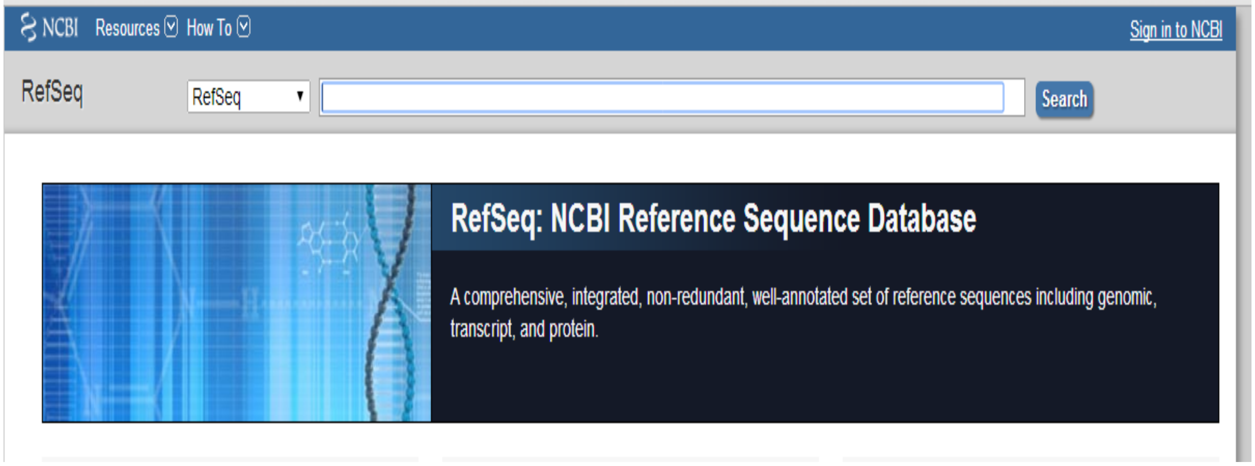
Part of the European Molecular Biology Laboratory

EMBL-EBI provides freely available data from life science experiments, performs basic research in computational biology and offers an extensive user training programme, supporting researchers in academia and industry.

Find a gene, protein or chemical:

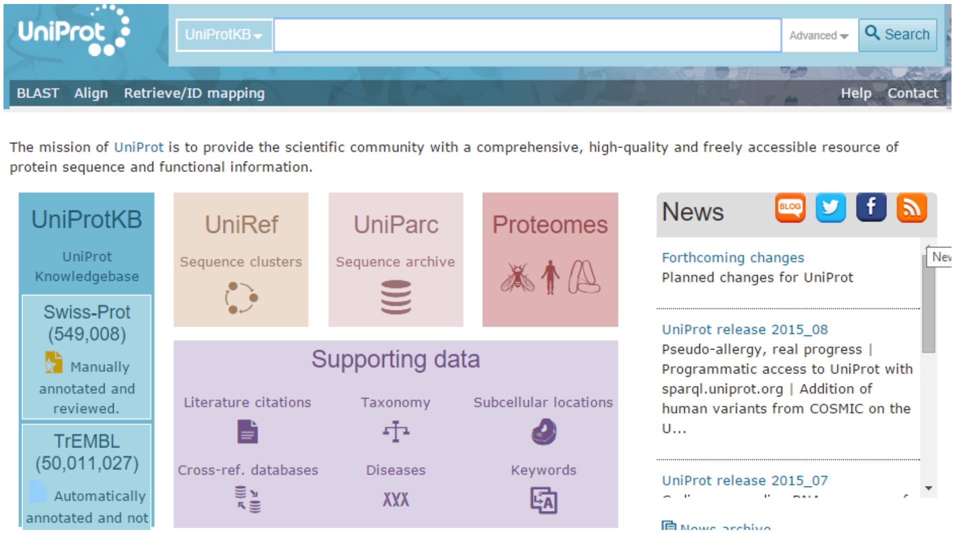
Examples: blast, keratin, bfl1 ...

Figure 0.46 gene, protein or chemical can be find



RefSeq: NCBI Reference Sequence Database

A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.



The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB
UniProt Knowledgebase

Swiss-Prot (549,008)
Manually annotated and reviewed.

TrEMBL (50,011,027)
Automatically annotated and not

UniRef
Sequence clusters

UniParc
Sequence archive

Proteomes

Supporting data
Literature citations
Taxonomy
Subcellular locations
Cross-ref. databases
Diseases
Keywords

News

Forthcoming changes
Planned changes for UniProt

UniProt release 2015_08
Pseudo-allergy, real progress | Programmatic access to UniProt with sparql.uniprot.org | Addition of human variants from COSMIC on the U...

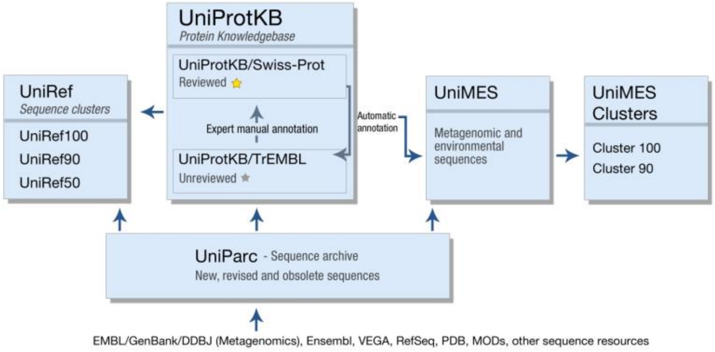
UniProt release 2015_07

News archive

Figure 0.47 online database for proteins

Swiss-Prot contains human curated protein information

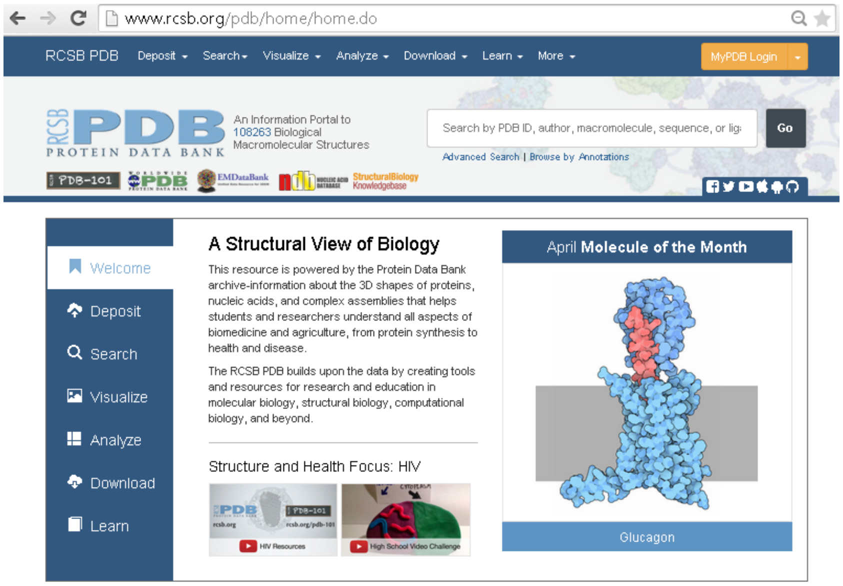
- Accession number, unique identifier
- The sequence
- Molecular mass
- Observed and predicted modifications



Protein sequences from various species and organisms can be found in uniprot. SwissProt is the manually annotated version of the UniProt Database.

Module 061: PROTEIN DATA BANK

Protein Data Bank is the premier resource of protein structures. These structures have been determined using experimental techniques. It’s Open & Free


www.rcsb.org/pdb/home/home.do
RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB Login
RCSB PDB PROTEIN DATA BANK
Search by PDB ID, author, macromolecule, sequence, or lig
Go
Advanced Search | Browse by Annotations
PDB-101
WORLDWIDE PDB PROTEIN DATA BANK
EMDataBank
ndb NUCLEIC ACID DATABASE
StructuralBiology Knowledgebase
Welcome
Deposit
Search
Visualize
Analyze
Download
Learn
HIV Resources
High School Video Challenge
April Molecule of the Month
Glucagon

An Information Portal to 108263 Biological Macromolecular Structures

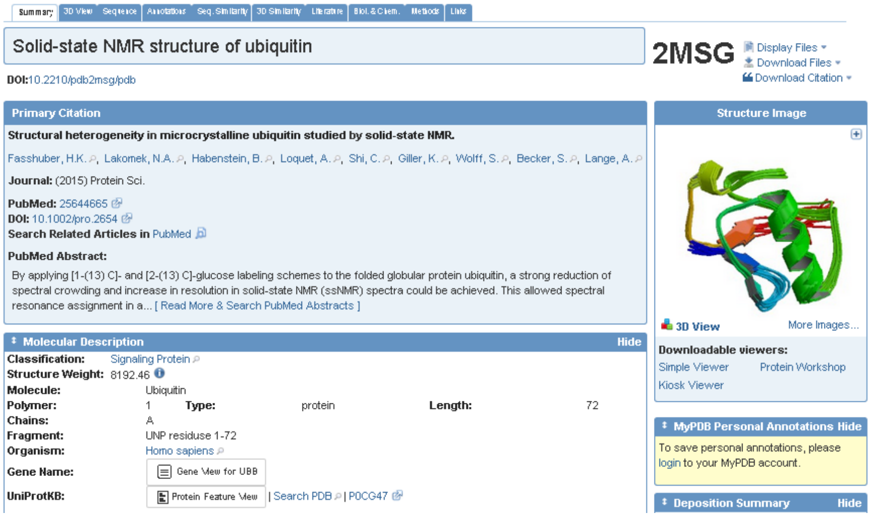
A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

Structure and Health Focus: HIV

Figure 0.48 protein data bank



Solid-state NMR structure of ubiquitin

2MSG

DOI:10.2210/pdb2msg/pdb

Primary Citation

Structural heterogeneity in microcrystalline ubiquitin studied by solid-state NMR.

Fasshuber, H.K., Lakomek, N.A., Habenstein, B., Loquet, A., Shi, C., Giller, K., Wolff, S., Becker, S., Lange, A.

Journal: (2015) Protein Sci.

PubMed: 25644665
DOI: 10.1002/pro.2654
Search Related Articles in PubMed

PubMed Abstract:

By applying [1-(13) C]- and [2-(13) C]-glucose labeling schemes to the folded globular protein ubiquitin, a strong reduction of spectral crowding and increase in resolution in solid-state NMR (ssNMR) spectra could be achieved. This allowed spectral resonance assignment in a... [Read More & Search PubMed Abstracts]

Molecular Description

Classification: Signaling Protein
Structure Weight: 8192.46
Molecule: Ubiquitin
Polymer: 1 **Type:** protein **Length:** 72
Chains: A
Fragment: UNP residuse 1-72
Organism: Homo sapiens
Gene Name: Gene View for UBB
UniProtKB: Protein Feature View | Search PDB | P0CG47

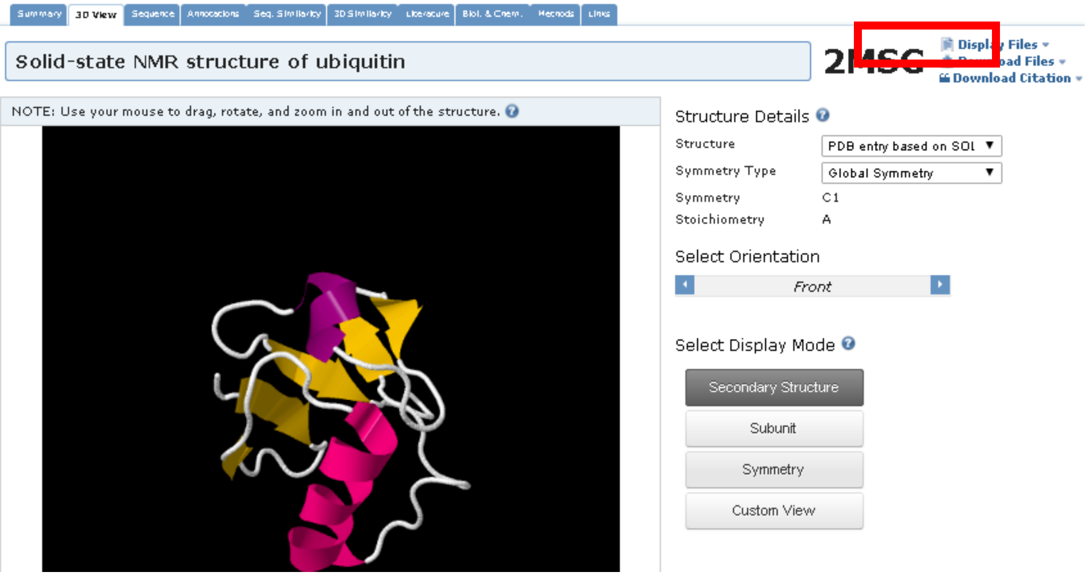
Downloadable viewers:

MyPDB Personal Annotations Hide

To save personal annotations, please login to your MyPDB account.

Deposition Summary Hide

Figure 0.49 P0CG47 - UBB_HUMAN



Solid-state NMR structure of ubiquitin

NOTE: Use your mouse to drag, rotate, and zoom in and out of the structure.

Structure Details

Structure: PDB entry based on SOL

Symmetry Type: Global Symmetry

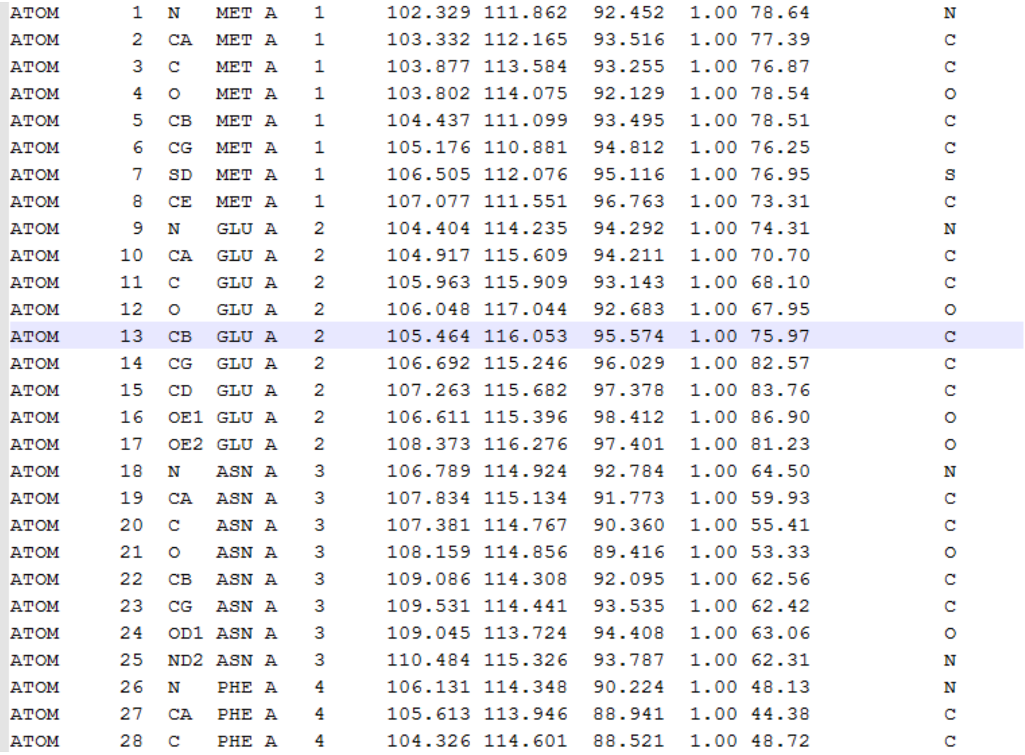
Symmetry: C1

Stoichiometry: A

Select Orientation

Select Display Mode

Figure 0.50 P0CG47 - UBB_HUMAN



```
ATOM      1  N   MET A   1     102.329 111.862  92.452  1.00 78.64           N
ATOM      2  CA  MET A   1     103.332 112.165  93.516  1.00 77.39           C
ATOM      3  C   MET A   1     103.877 113.584  93.255  1.00 76.87           C
ATOM      4  O   MET A   1     103.802 114.075  92.129  1.00 78.54           O
ATOM      5  CB  MET A   1     104.437 111.099  93.495  1.00 78.51           C
ATOM      6  CG  MET A   1     105.176 110.881  94.812  1.00 76.25           C
ATOM      7  SD  MET A   1     106.505 112.076  95.116  1.00 76.95           S
ATOM      8  CE  MET A   1     107.077 111.551  96.763  1.00 73.31           C
ATOM      9  N   GLU A   2     104.404 114.235  94.292  1.00 74.31           N
ATOM     10  CA  GLU A   2     104.917 115.609  94.211  1.00 70.70           C
ATOM     11  C   GLU A   2     105.963 115.909  93.143  1.00 68.10           C
ATOM     12  O   GLU A   2     106.048 117.044  92.683  1.00 67.95           O
ATOM     13  CB  GLU A   2     105.464 116.053  95.574  1.00 75.97           C
ATOM     14  CG  GLU A   2     106.692 115.246  96.029  1.00 82.57           C
ATOM     15  CD  GLU A   2     107.263 115.682  97.378  1.00 83.76           C
ATOM     16  OE1 GLU A   2     106.611 115.396  98.412  1.00 86.90           O
ATOM     17  OE2 GLU A   2     108.373 116.276  97.401  1.00 81.23           O
ATOM     18  N   ASN A   3     106.789 114.924  92.784  1.00 64.50           N
ATOM     19  CA  ASN A   3     107.834 115.134  91.773  1.00 59.93           C
ATOM     20  C   ASN A   3     107.381 114.767  90.360  1.00 55.41           C
ATOM     21  O   ASN A   3     108.159 114.856  89.416  1.00 53.33           O
ATOM     22  CB  ASN A   3     109.086 114.308  92.095  1.00 62.56           C
ATOM     23  CG  ASN A   3     109.531 114.441  93.535  1.00 62.42           C
ATOM     24  OD1 ASN A   3     109.045 113.724  94.408  1.00 63.06           O
ATOM     25  ND2 ASN A   3     110.484 115.326  93.787  1.00 62.31           N
ATOM     26  N   PHE A   4     106.131 114.348  90.224  1.00 48.13           N
ATOM     27  CA  PHE A   4     105.613 113.946  88.941  1.00 44.38           C
ATOM     28  C   PHE A   4     104.326 114.601  88.521  1.00 48.72           C
```

Figure 0.51 searched results

Protein Data Bank provides Cartesian coordinates of each atom in the protein structure. Over 50,000 protein structures are reported and present in this database

Module 062: REVIEW OF SEQUENCE ALIGNMENT

We use next generation sequencing and whole genome sequencing to obtain the genetic information. For protein sequencing we use Mass Spectrometry and Edman Degradation

STORAGE:

- Sequence information is stored digitally
- Databases are designed to store sequence data
- Several databases exist depending on the type of sequence data

SHARING AND ACCESS:

- Sequence databases are shared via online websites
- Access to several such websites is free
- Data can be downloaded or searched on these website

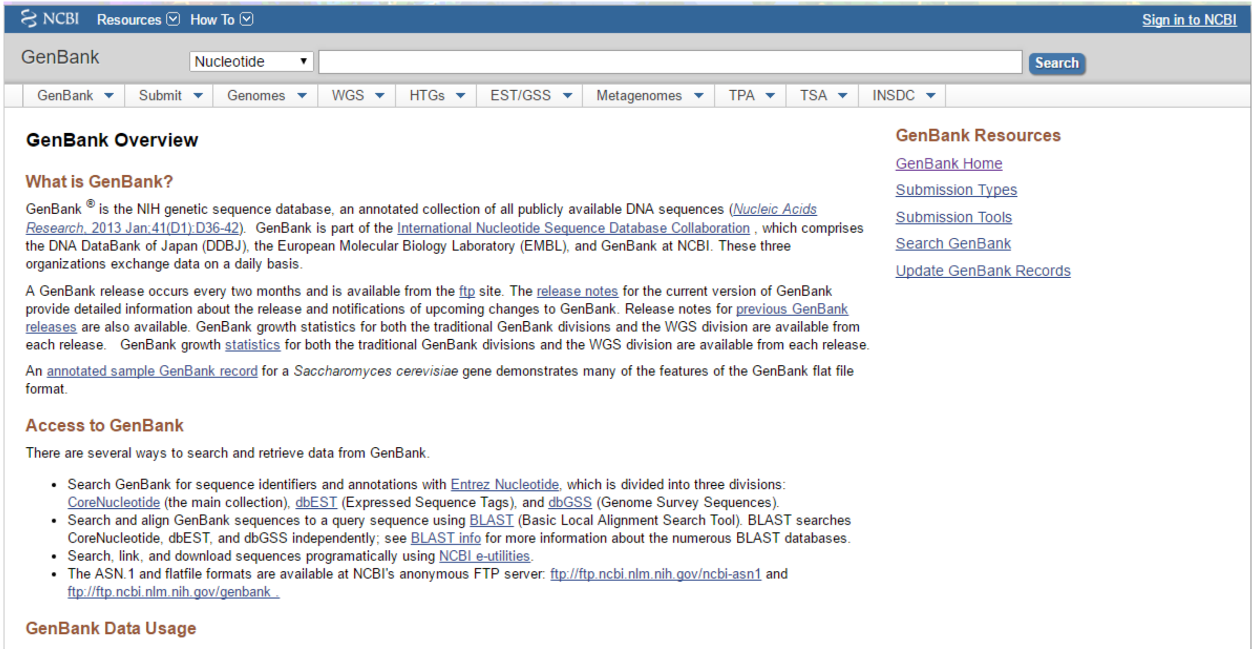
USAGE OF DATA:

Sequence data can be used to obtain:

- Similarity of sequences
- Evolutionary History
- Predict the function of molecules

Module 063: GENBANK

- Developed by Swiss Bioinformatics Institute (SIB)
- Website provides access to databases and tools
- Proteomics, genomics, phylogeny, systems biology, population genetics, transcriptomics etc.



GenBank Overview

What is GenBank?

GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration , which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the ftp site. The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for previous GenBank releases are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release.

An annotated sample GenBank record for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with Entrez Nucleotide, which is divided into three divisions: CoreNucleotide (the main collection), dbEST (Expressed Sequence Tags), and dbGSS (Genome Survey Sequences).
- Search and align GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently; see BLAST info for more information about the numerous BLAST databases.
- Search, link, and download sequences programatically using NCBI e-utilities.
- The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1 and ftp://ftp.ncbi.nlm.nih.gov/genbank .

GenBank Data Usage

GenBank Resources

GenBank Home
Submission Types
Submission Tools
Search GenBank
Update GenBank Records

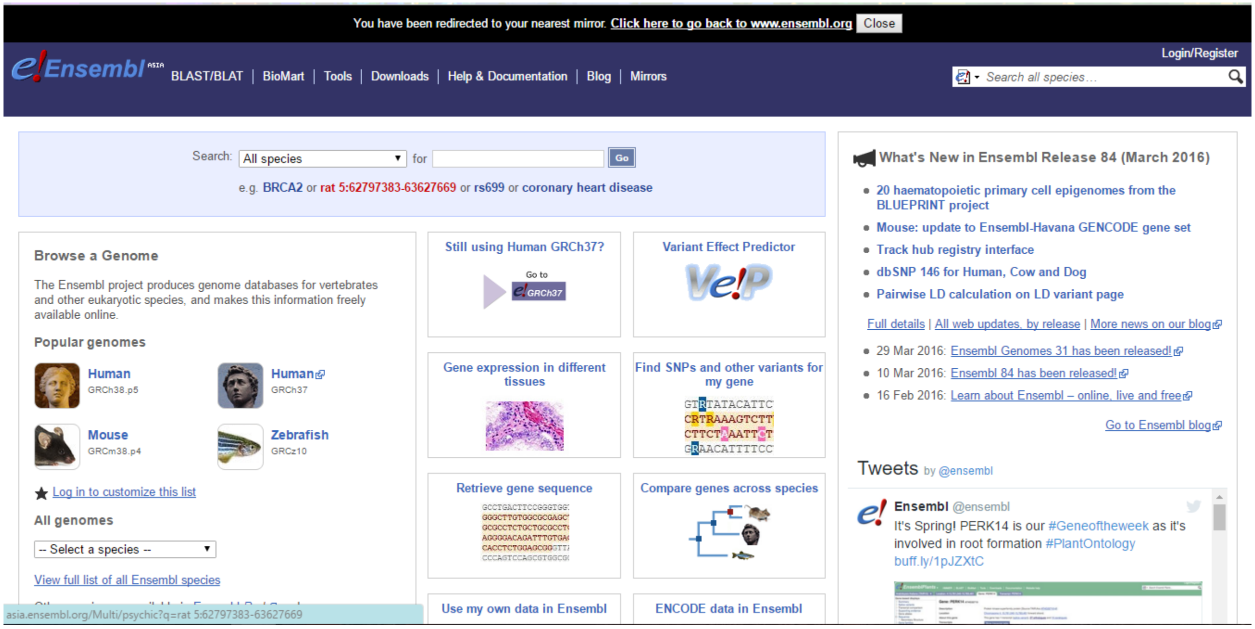
Figure 0.52 *http://www.ncbi.nlm.nih.gov/genbank/*

Several sequence, structure and molecular interaction databases exist. These are available online on the web. Users can freely access and download such data

Module 064: ENSEMBLE

As human brain is limited to remember and store the information for long time that's why we use online database for the storage of Molecular information.

ESEMBLE is genome search engine which is used to search the genome of every recorded species.


Close
Ensembl ASIA
BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors
Login/Register
Search all species…
Search: All species for Go
Still using Human GRCh37?
Go to GRCh37
Variant Effect Predictor
VeP
Gene expression in different tissues
Find SNPs and other variants for my gene
Retrieve gene sequence
Compare genes across species
Use my own data in Ensembl
ENCODE data in Ensembl
Human GRCh38.p5
Human GRCh37
Mouse GRCm38.p4
Zebrafish GRCz10
-- Select a species --
Tweets by @ensembl

You have been redirected to your nearest mirror. Click here to go back to www.ensembl.org

e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

Log in to customize this list

All genomes

View full list of all Ensembl species

asia.ensembl.org/Multi/psychic?q=rat 5:62797383-63627669

What's New in Ensembl Release 84 (March 2016)

- 20 haematopoietic primary cell epigenomes from the BLUEPRINT project
- Mouse: update to Ensembl-Havana GENCODE gene set
- Track hub registry interface
- dbSNP 146 for Human, Cow and Dog
- Pairwise LD calculation on LD variant page

Full details | All web updates, by release | More news on our blog

- 29 Mar 2016: Ensembl Genomes 31 has been released!
- 10 Mar 2016: Ensembl 84 has been released!
- 16 Feb 2016: Learn about Ensembl – online, live and free

Go to Ensembl blog

Ensembl @ensembl
It's Spring! PERK14 is our #Geneoftheweek as it's involved in root formation #PlantOntology buff.ly/1pJZXtC

http://asia.ensembl.org/index.html

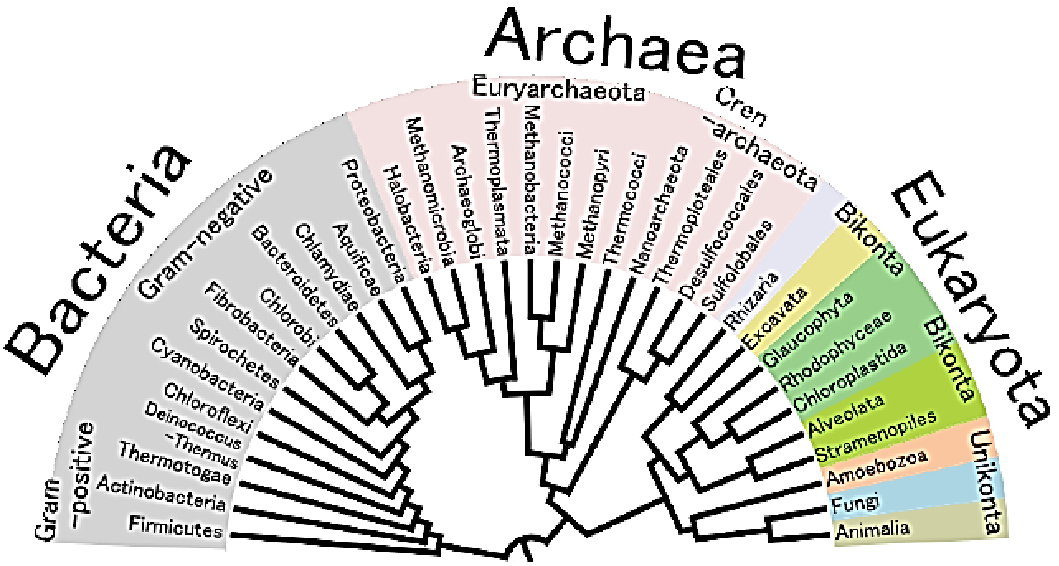
Chapter 3 - Molecular Evolution

Module 001: Molecular Evolution & Phylogeny

Molecular evolution is the process of change in the sequence composition of cellular molecules such as DNA, RNA, and proteins across generations. The field of molecular evolution uses principles of evolutionary biology and population genetics to explain patterns in these changes. Genes and Proteins are modified in this process.

All molecules have an evolutionary history. Phylogenetics is the science of studying evolutionary relationships. Phylogenetics has led to the creation of relationship trees between various species of Bacteria, Archaea, and Eukaryota.

(Page and Holmes 2009)



http://bacterialphylogeny.info/overview.html

Figure 0.1 Phylogenetic tree

Types of Phylogenetic Trees

Scaled Trees

- Branch lengths are equal to the magnitude of change in the nodes

Unscaled Trees

- Only representing the relationship between sequences

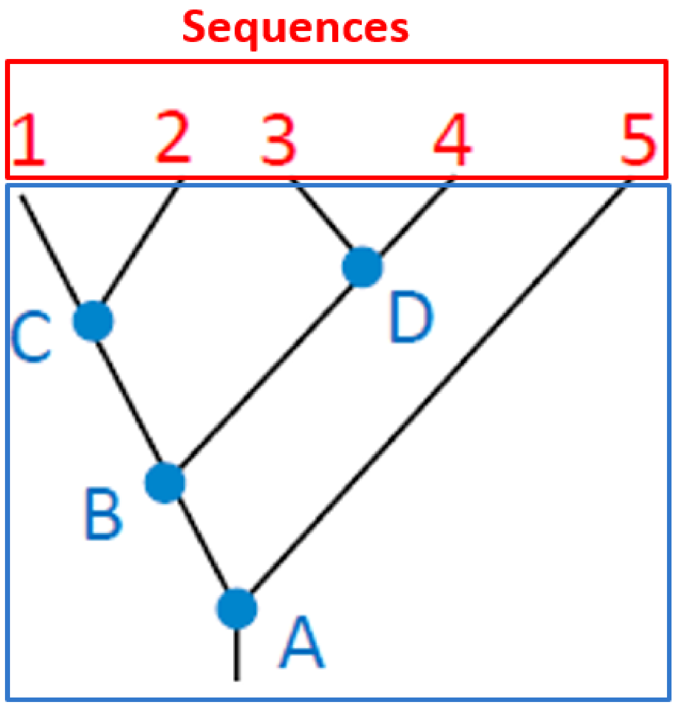


Figure 0.2 phylogenetic tree interference

Conclusion

Phylogenetics is the study of extracting evolutionary relationships between species. Sequence information from each species is used to measure the difference between the species.

Page, R. D. and E. C. Holmes (2009). Molecular evolution: a phylogenetic approach, John Wiley & Sons.

Module 002: Evolution of Sequences

DNA acts as cellular memory unit and protein are the translated product of DNA coded information. And evaluation is very important to survive in different type of environments. There are some methods which brings change or evolution in any organism. (Kluger 2015)

Method of Change

DNA gets modified by:

- Mutation & Substitution
- Insertion
- Deletion

Discussion

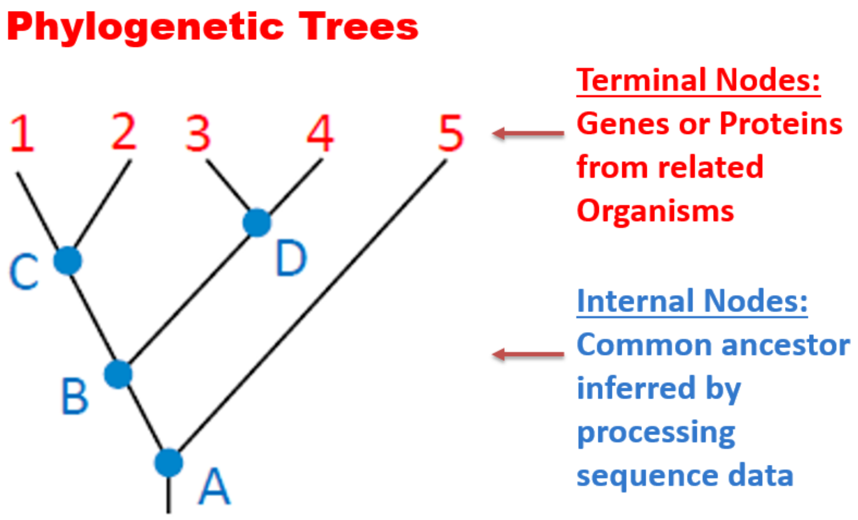
Over time, species evolve to adapt to their circumstances. Since the environment and circumstances may be different for each species, they evolve uniquely. Unique evolutionary pressures may be encountered by each cell for struggle of life. However, in which sequence they are presented to the cells is also unique. Combinations of evolutionary factors are involve in evolution. The evolutionary events and their combination impart relationships between sequences. These relationships are explored in Phylogenetics .Several algorithms exist for finding such relationships

Kluger, M. J. (2015). Fever: its biology, evolution, and function, Princeton University Press.

Page, R. D. and E. C. Holmes (2009). Molecular evolution: a phylogenetic approach, John Wiley & Sons.

Module 003: Concepts and Terminologies - I

To understand the concept of evolution we follow some rules. Phylogenetics involves processing sequence information from different species to find evolutionary relationships. Output from such studies include Phylogenetic Trees



Phylogenetic Trees

Terminal Nodes: Genes or Proteins from related Organisms

Internal Nodes: Common ancestor inferred by processing sequence data

Figure 3 phylogenetic tree from ancestor to evolution

In above figure the point A stands for ancestor and with the passage of time the evolution occurred with and the genome sequence of organisms changed.

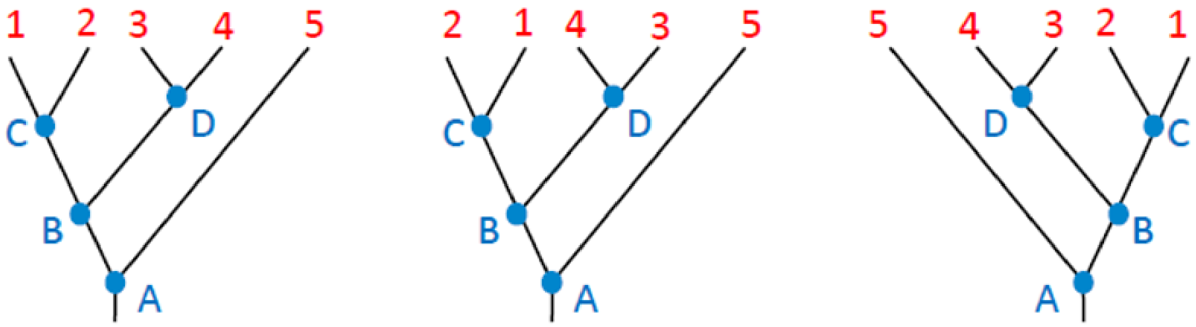


Figure 4 layout of trees

All trees have same meanings.

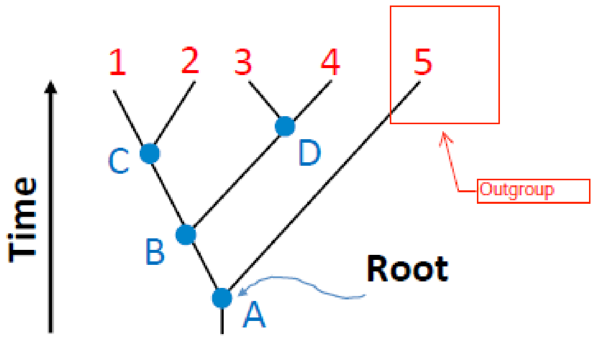


Figure 5 rooted tree

Root node is the ancestor of all other nodes. The direction of evolution is from ancestor to the terminal nodes.

Conclusion

Phylogenetics specifies evolutionary relationship with the help of trees. Trees can be rooted or unrooted. Rooted trees can show temporal evolutionary direction.

Module 004: Concepts and Terminologies - II

Rooted and Unrooted trees can be used to show phylogenetic relationships between sequences. Let’s examine the properties of these trees further.

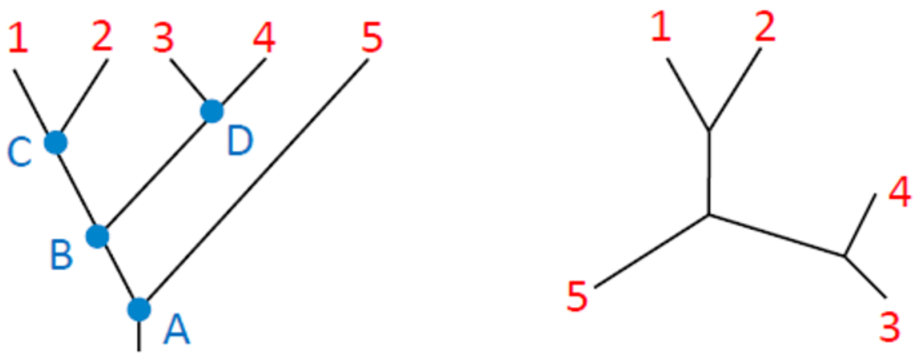
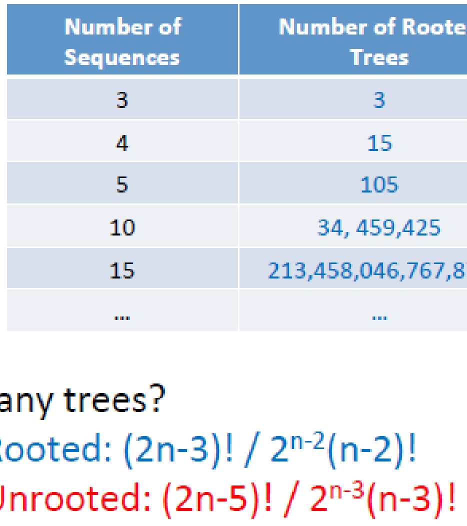


Figure 6 rooted tree vs unrooted

Rooted trees are computationally expensive.
http://everything.explained.today/Computational_phylogenetics/

https://github.com/joey711/phyloseq/issues/597



Number of Sequences	Number of Rooted Trees	Number of Unrooted Trees
3	3	1
4	15	3
5	105	15
10	34, 459,425	2,027,025
15	213,458,046,767,875	7,905,853,580,625
…	…	…

How many trees?

Rooted: $(2n-3)! / 2^{n-2}(n-2)!$

Unrooted: $(2n-5)! / 2^{n-3}(n-3)!$

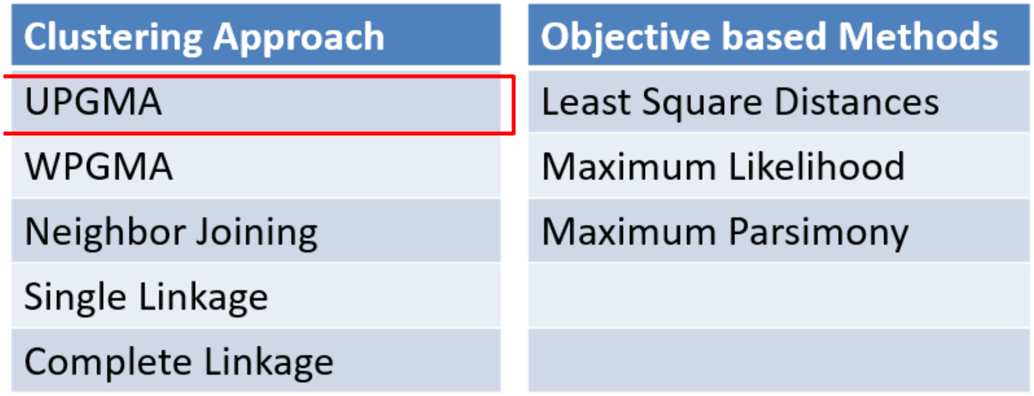
Figure 7 computation comparison

Conclusion

Rooted and Unrooted trees have their own advantages and disadvantages. Depending on our requirement, we can choose between them.

Module 005: Algorithms and Techniques

Rooted and Unrooted trees can be used to show phylogenetic relationships between sequences. Several types of algorithms exist which are divided into two classes. There are many methods for constructing evolutionary trees.

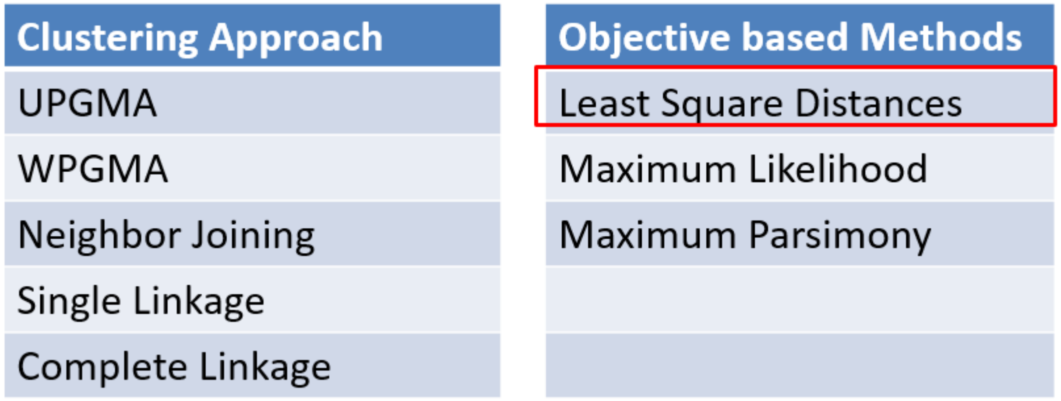


Clustering Approach	Objective based Methods
UPGMA	Least Square Distances
WPGMA	Maximum Likelihood
Neighbor Joining	Maximum Parsimony
Single Linkage	
Complete Linkage	

Figure 9 construction methods

UPGMA (Unweighted Pair Group Method with Arithmetic Mean) is a simple agglomerative (bottom-up) hierarchical clustering method. The method is generally attributed to Sokal and Michener.

In this method two sequences with with the shortest evolutionary distance between them are considered and these sequences will be the last to diverge, and represented by the most recent internal node.



Clustering Approach	Objective based Methods
UPGMA	Least Square Distances
WPGMA	Maximum Likelihood
Neighbor Joining	Maximum Parsimony
Single Linkage	
Complete Linkage	

Least Squares Distance Method. Branch lengths, represent the “observed” distances between sequences (i *& j)*.

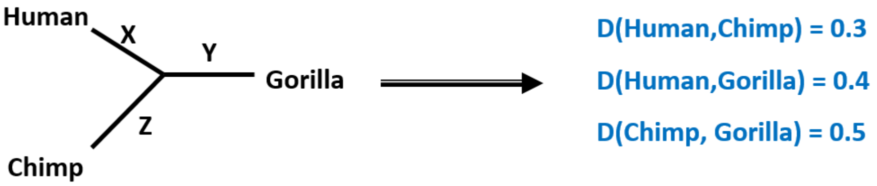
Observation

D(Human,Chimp) = 0.3

D(Human,Gorilla) = 0.4

D(Chimp, Gorilla) = 0.5

Find X, Y and Z such that *D (i, j)* are conserved?



D(Human,Chimp) = 0.3

D(Human,Gorilla) = 0.4

D(Chimp, Gorilla) = 0.5

Conclusion

- Several methods exist for constructing phylogenetic trees.
- Broadly, they belong to objective methods or clustering methods.
- We will study UPGMA and Distance Methods.

Module 006: Introduction to UPGMA

Phylogenetic trees can be used to show phylogenetic relationships between sequences. To construct these trees, several types of algorithms exist which are divided into two classes.

UPGMA: Unweighted Pair – Group Method using arithmetic Averages

- **Calculating distance between two clusters:**

Cluster X + Cluster Y = Cluster Z

Calculate the distance of a cluster (e.g. W) to the new cluster Z

$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$

N_x is the number of sequences in cluster x

- **Calculating distance between two trees:**

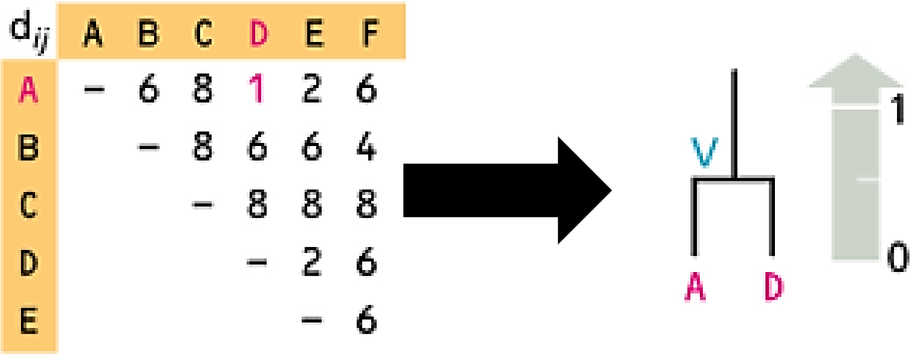
Assume we have N sequences

Cluster X has N_X sequences, cluster Y has N_Y sequences

d_{XY} : the evlotionary distance between X and Y

$$d_{XY} = \frac{1}{N_X N_Y} \sum_{i \in X, j \in Y} d_{ij}$$

- **Methods for constructing trees**



d_{ij}	A	B	C	D	E	F
A	–	6	8	1	2	6
B		–	8	6	6	4
C			–	8	8	8
D				–	2	6
E					–	6

The distance matrix is obtained using pairwise sequence alignment.

- **Calculating distance between two clusters:**

Cluster X + Cluster Y = Cluster Z

Calculate the distance of a cluster (e.g. W) to the new cluster Z

$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$

N_x is the number of sequences in cluster x

- **Calculating distance between two trees:**

Assume we have N sequences

Cluster X has N_X sequences, cluster Y has N_Y sequences

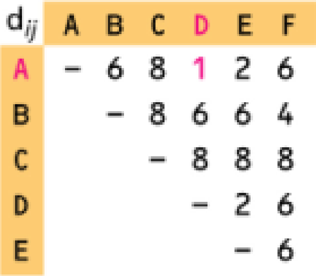
d_{XY} : the evlotionary distance between X and Y

$$d_{XY} = \frac{1}{N_X N_Y} \sum_{i \in X, j \in Y} d_{ij}$$

- **Methods for constructing trees**

A – D becomes a new cluster lets say V. We have to modify the distance matrix. What are the distances between:

- V and B (Calculate),
- V and C,
- V and E,
- V and F.



d_{ij}	A	B	C	D	E	F
A	–	6	8	1	2	6
B		–	8	6	6	4
C			–	8	8	8
D				–	2	6
E					–	6

$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$

$$d_{VB} = \frac{N_A d_{AB} + N_D d_{DB}}{N_A + N_D} = \frac{1*6+1*6}{1+1} = 6$$

Conclusion

UPGMA is a clustering algorithm which can help us compute phylogenetic trees. We will see the detailed working of this approach in later modules.

Module 007: UPGMA-I

UPGMA has two components to it. These include distance calculations between two clusters and between two trees.

- **Building trees using UPGMA**

Combining Clusters: Cluster X + Cluster Y = Cluster Z

Calculate the distance of each cluster (e.g. W) to the new cluster Z

$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$

N_x is the number of sequences in cluster x

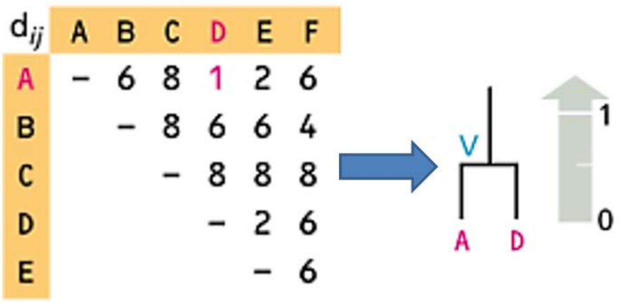
- **Calculating the distance between two trees**

Assume we have N sequences

Cluster X has N_X sequences, cluster Y has N_Y sequences

d_{XY} : the evlotionary distance between X and Y

$$d_{XY} = \frac{1}{N_X N_Y} \sum_{i \in X, j \in Y} d_{ij}$$



d_{ij}	A	B	C	D	E	F
A	–	6	8	1	2	6
B		–	8	6	6	4
C			–	8	8	8
D				–	2	6
E					–	6

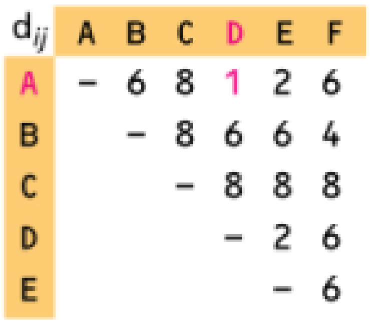
Figure 10 the distance matrix is obtained using pairwise sequence alignment

- **Methods for constructing trees**

A – D becomes a new cluster lets say V.

We have to modify the distance matrix!

What are the distances between:



d_{ij}	A	B	C	D	E	F
A	–	6	8	1	2	6
B		–	8	6	6	4
C			–	8	8	8
D				–	2	6
E					–	6

- V and B (Calculate),

- V and C,
- V and E,
- V and F.

$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$

$$d_{VB} = \frac{N_A d_{AB} + N_D d_{DB}}{N_A + N_D} = \frac{1*6+1*6}{1+1} = 6$$

Conclusion

UPGMA starts with creating clusters of sequences which are the closest. Next, distance is computed between the new cluster and the remaining sequences. The process is repeated for all sequences.

Module 008: UPGMA-II

UPGMA steps include distance calculations between two clusters and between two trees. We formed clusters from sequences which had the shortest distance.

Building trees using UPGMA

Combining Clusters: Cluster X + Cluster Y = Cluster Z

Calculate the distance of each cluster (e.g. W) to the new cluster Z

$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$

N_x is the number of sequences in cluster x

Calculating the distance between two trees

Assume we have N sequences

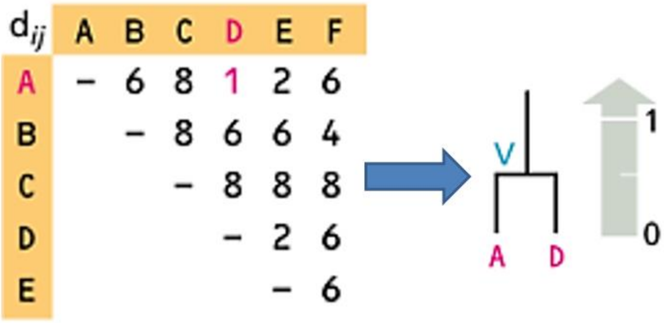
Cluster X has N_X sequences, cluster Y has N_Y sequences

d_{XY} : the evlotionary distance between X and Y

$$d_{XY} = \frac{1}{N_X N_Y} \sum_{i \in X, j \in Y} d_{ij}$$

Methods for constructing trees

The distance matrix is obtained using pairwise sequence alignment.



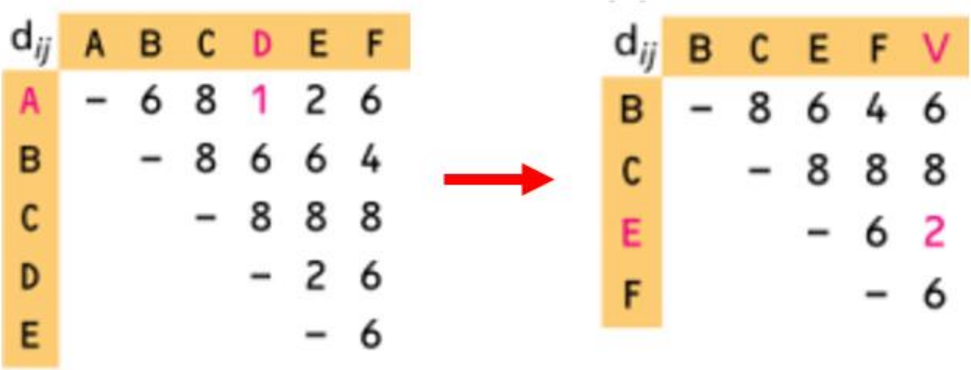
d_{ij}	A	B	C	D	E	F
A	–	6	8	1	2	6
B		–	8	6	6	4
C			–	8	8	8
D				–	2	6
E					–	6

$$d_{VB} = \frac{N_A d_{AB} + N_D d_{DB}}{N_A + N_D} = \frac{1*6+1*6}{1+1} = 6$$

$$d_{VC} = \frac{N_A d_{AC} + N_D d_{DC}}{N_A + N_D} = \frac{1*8+1*8}{1+1} = 8$$

$$d_{VE} = \frac{N_A d_{AE} + N_D d_{DE}}{N_A + N_D} = \frac{1*2+1*2}{1+1} = 2$$

$$d_{VF} = \frac{N_A d_{AF} + N_D d_{DF}}{N_A + N_D} = \frac{1*6+1*6}{1+1} = 6$$



d_{ij}	A	B	C	D	E	F
A	–	6	8	1	2	6
B		–	8	6	6	4
C			–	8	8	8
D				–	2	6
E					–	6

d_{ij}	B	C	E	F	V
B	–	8	6	4	6
C		–	8	8	8
E			–	6	2
F				–	6

V – E becomes a new cluster lets say W

Now we have to modify the distance matrix again.

What are the distances between:

W and B,

W and C,

W and F.

Conclusion

Once a cluster is selected and its distance is computed with all other sequences, we update the distance matrix. Next, we select the shortest distance from the new matrix and repeat the process.

Module 009: UPGMA-III

UPGMA has two components to it. These include progressive distance calculations between two clusters or between two trees.

Building trees using UPGMA

Combining Clusters: Cluster X + Cluster Y = Cluster Z. Calculate the distance of each cluster (e.g. W) to the new cluster Z.

$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$

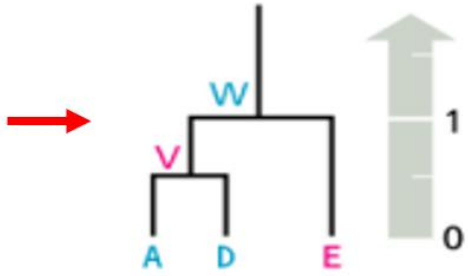
N_x is the number of sequences in cluster x

Calculating the distance between two trees

- Assume we have N sequences
- Cluster X has N_X sequences, cluster Y has N_Y sequences
- d_{XY} : the evlotionary distance between X and Y

$$d_{XY} = \frac{1}{N_X N_Y} \sum_{i \in X, j \in Y} d_{ij}$$

d_{ij}	B	C	E	F	V
B	–	8	6	4	6
C		–	8	8	8
E			–	6	2
F				–	6



V – E becomes a new cluster lets say W. Now we have to modify the distance matrix again.

What are the distances between:

W and B,

W and C,

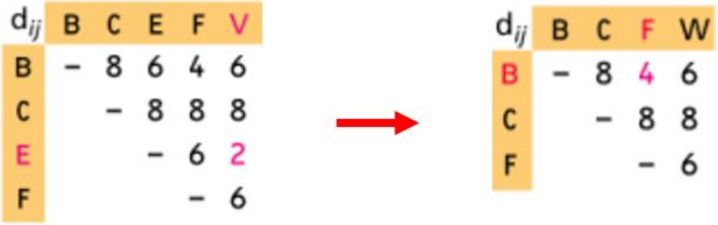
W and F.

$$d_{WB} = \frac{N_V d_{VB} + N_E d_{EB}}{N_V + N_E} = \frac{2*6+1*6}{2+1} = 6$$

$$d_{WC} = \frac{N_V d_{VC} + N_E d_{EC}}{N_V + N_E} = \frac{2*8+1*8}{2+1} = 8$$

$$d_{WF} = \frac{N_V d_{VF} + N_E d_{EF}}{N_V + N_E} = \frac{2*6+1*6}{2+1} = 6$$

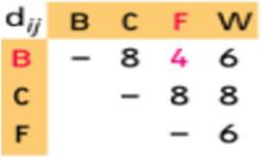
New matrix



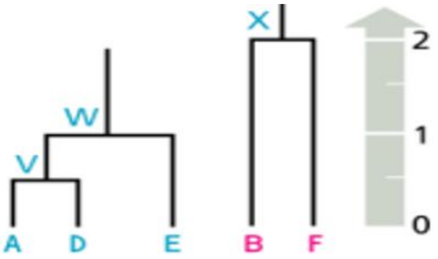
d_{ij}	B	C	E	F	V
B	–	8	6	4	6
C		–	8	8	8
E			–	6	2
F				–	6

d_{ij}	B	C	F	W
B	–	8	4	6
C		–	8	8
F			–	6

Cluster according to min distance



d_{ij}	B	C	F	W
B	–	8	4	6
C		–	8	8
F			–	6

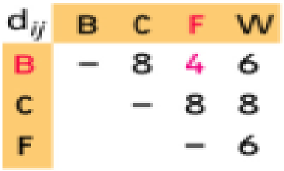


Conclusion

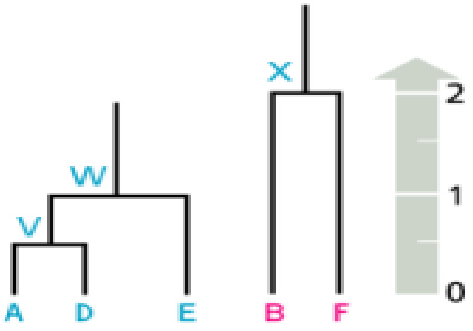
Now we have formed three clusters. Also, two separate trees have been formed. Next, we need to join these trees to create a complete tree.

Module 010: UPGMA-IV

Application of UPGMA resulted in formation of two sub-trees. The need now was to join them into a single tree. Let's see how that is done.



d_{ij}	B	C	F	W
B	–	8	4	6
C		–	8	8
F			–	6

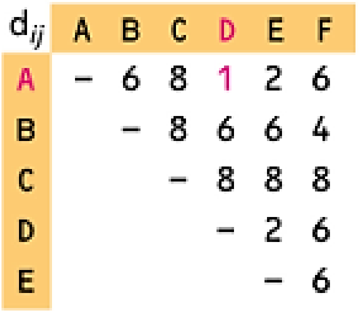


F – B becomes a new cluster lets say X. We have to modify the distance matrix yet again. What is the distance between trees: W and X.

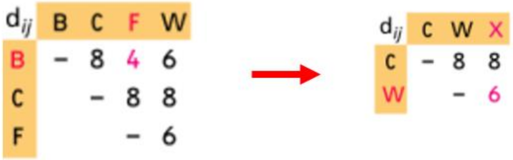
$$d_{WX} = \frac{1}{N_W N_X} \sum_{i\in W, j\in X} d_{ij} =$$

$$\frac{1}{N_W N_X}(d_{AB} + d_{AF} + d_{DB} + d_{DF} + d_{EB} + d_{EF}) =$$

$$\frac{1}{3*2}*(6+6+6+6+6+6) = 6$$

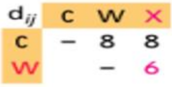


d_{ij}	A	B	C	D	E	F
A	–	6	8	1	2	6
B		–	8	6	6	4
C			–	8	8	8
D				–	2	6
E					–	6

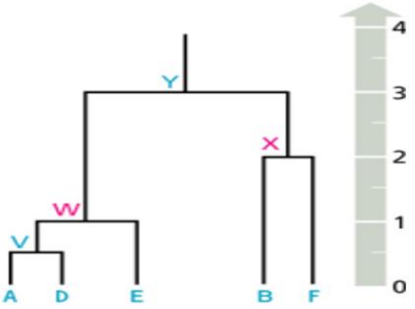


d_{ij}	B	C	F	W
B	–	8	4	6
C		–	8	8
F			–	6

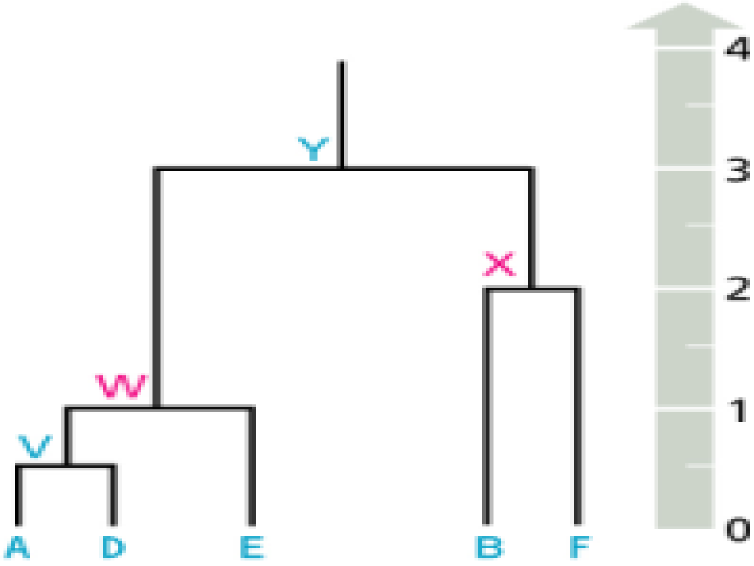
d_{ij}	C	W	X
C	–	8	8
W		–	6



d_{ij}	C	W	X
C	–	8	8
W		–	6



d_{ij}	C	W	X
C	–	8	8
W		–	6



X – W becomes a new cluster lets say Y. We have to modify the distance matrix

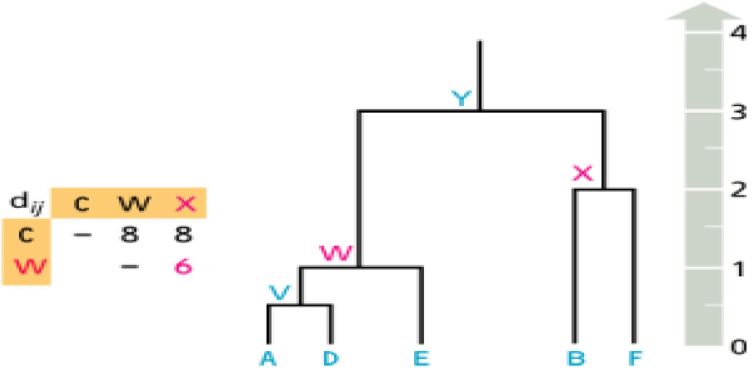
What is the distance between: Y and C.

Conclusion

We have now seen how trees are generated and connected. Next, we need to finalize the tree by adding the last two clusters.

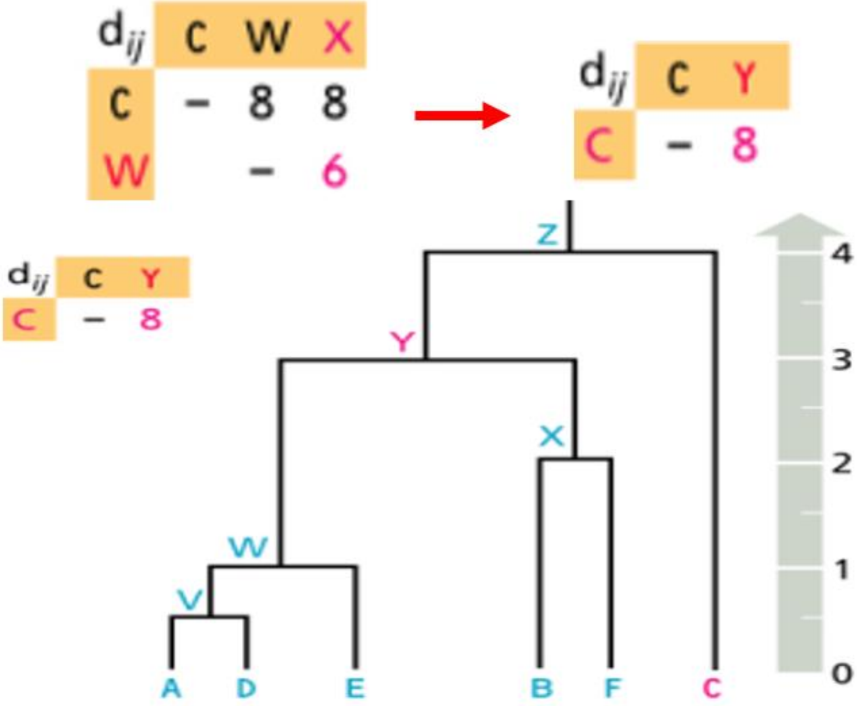
Module 011: UPGMA-V

Application of UPGMA resulted in formation of two sub-trees. The need now was to join them into a single tree. Let’s see how that is done.



d_{ij}	C	W	X
C	–	8	8
W		–	6

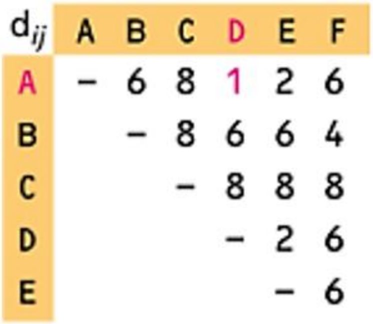
X – W becomes a new cluster lets say Y. We have to modify the distance matrix. What is the distance between: Y and C.



d_{ij}	C	W	X
C	–	8	8
W		–	6

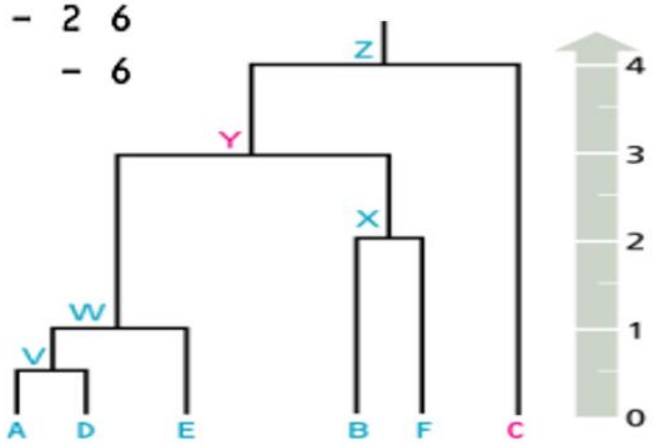
d_{ij}	C	Y
C	–	8

d_{ij}	C	Y
C	–	8



d_{ij}	A	B	C	D	E	F
A	–	6	8	1	2	6
B		–	8	6	6	4
C			–	8	8	8
D				–	2	6
E					–	6

$$d_{ZW} = \frac{N_X d_{XW} + N_Y d_{YW}}{N_X + N_Y}$$



Conclusion

Un-weighted Pair Group Method using Arithmetic Averages is a clustering method to construct phylogenetic trees. Non-clustering methods such as Maximum Parsimony may be used for making trees as well.

Chapter 4 - RNA Secondary Structure Prediction

Module 001: DNA TO RNA SEQUENCES

- **MOTIVATION**

In early days RNA was a considered as a structure which was involve between DNA and protein, means takes information from DNA and converts that information into protein synthesis. Now we know that it has multiple types like mRNA, tRNA, miRNA and siRNA. And they perform most of the work in gene expression and proteins. Not All RNA molecules are same, they differ in nucleotide sequences and functions also.

Many viruses assemble their genomes from RNAs. They are therefore called RNA viruses. Examples include Human Immunodeficiency Virus and Hepatitis C Virus.

There is little difference between RNA and DNA:

- Thymine is replaced by Uracil in RNA molecule.
- RNA molecule is single stand.
- RNA contain ribose sugar.

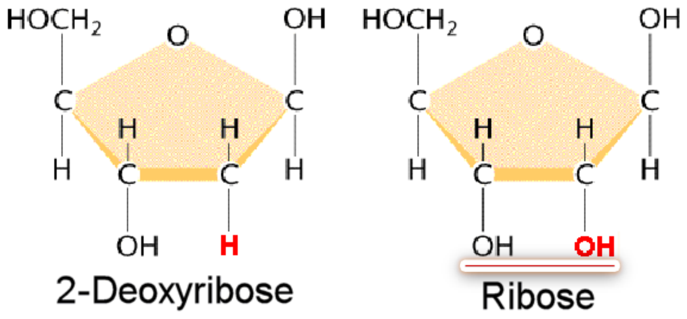


Figure 0.1 Ribose sugar has (OH) and Deoxyribose (H)

Because RNA has two (OH) groups that’s why it has shot life spam because of both (OH) repulsion.

Module 002: TYPES OF RNA &THEIR FUNCTIONS

There are two categories of RNA:

- Coding RNA
- Non-Coding RNA

Coding RNA perform their coded function in protein synthesis. And Non-coding RNA helps in translation process.

- **TYPES OF RNA**

There are many types of RNA according to their funtions like:

- Messenger RNA (mRNA)
- Transfer RNA (tRNA)
- Ribosomal RNA (rRNA)
- Micro RNAs (miRNA)
- Small Interfering RNA (siRNA)
-

- **MESSENGER RNA**

Only 5-10% of this RNA type is present in cell. Which has variable sequence, variable size and it carries the genetic information form DNA to Ribosomes where proteins to be assembled. Messenger RNA 5’ end is capped with (**7-Methyl Guanosine Triphosphate)** which helps the Ribosomes to identify the mRNA. And 3’ end of the mRNA is poly A tail (around 30-200 adenylate residues) which help shield against 3’ exonucleases)

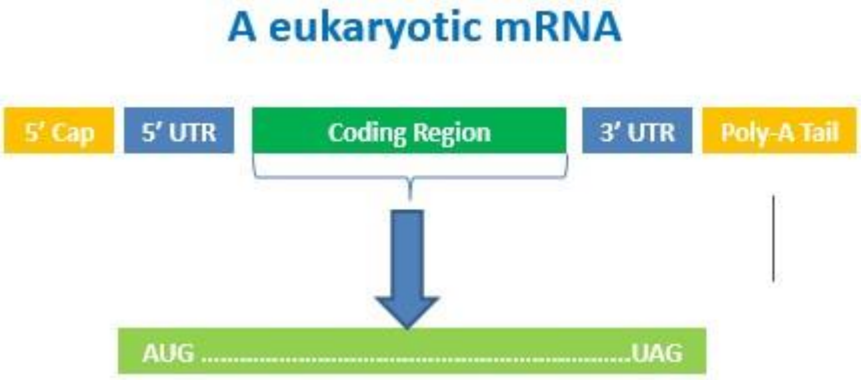

A eukaryotic mRNA
5’ Cap
5’ UTR
Coding Region
3’ UTR
Poly-A Tail
AUG ....UAG

Figure 2 RNA sequence is complementary to the DNA sequence and is translated as codons of three nucleotides

As RNA has differ in nucleotides sequences therefore differ in functions.

Module 003: SIGNIFICANCES OF RNA STRUCTURE

RNA can form 3D structures {Sarver, 2008 #5}, such structural properties helps the RNA molecule to perform different functions.

As RNA is composed of sugars, phosphate and nucleotides and these nucleotides have ability to form hydrogen bonds.

- A' can make hydrogen bonds with 'U'
- 'G' makes hydrogen bonds with 'C'
- 'G' can also make hydrogen bonds with 'U' (Wobble Pair)

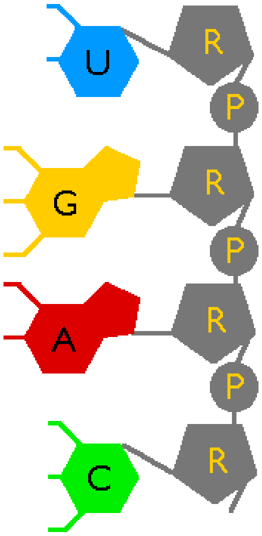


Figure 3 In RNA ribose is used in place of deoxyribose 3 In RNA uracil is used in place of thymine

Due to this ability of bonding RNA forms many structures and due to variety of structures RNA performs many functions in cell like:

- DNA information transfer

- Regulatory roles
- Catalytic roles
- Defense & immune response
- Structure-based special roles

Module 004: RNA FOLDING AND ENERGY FOLDING

RNA molecules form many structures for stability and different functions. “Gibs Free Energy” (LANGRIDGE and KOLLMAN 1987) is the free energy available for RNA molecule for reactions and RNA structure formation takes place at this lower energy. Incase if RNA has two structure we can select the one with lowest energy state.

http://chemwiki.ucdavis.edu/Core/Physical_Chemistry/Thermodynamics/State_Functions/Free_Energy/Gibbs_Free_Energy

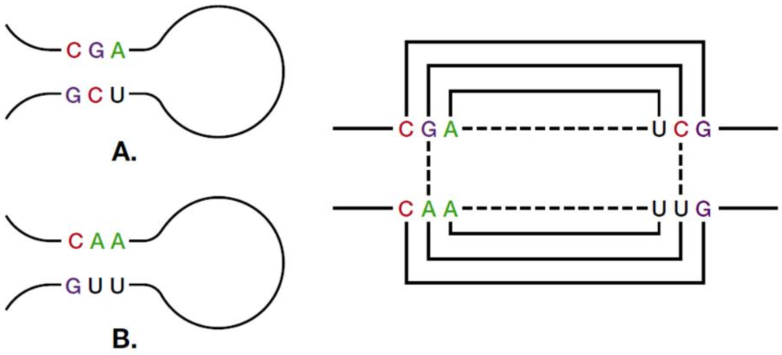
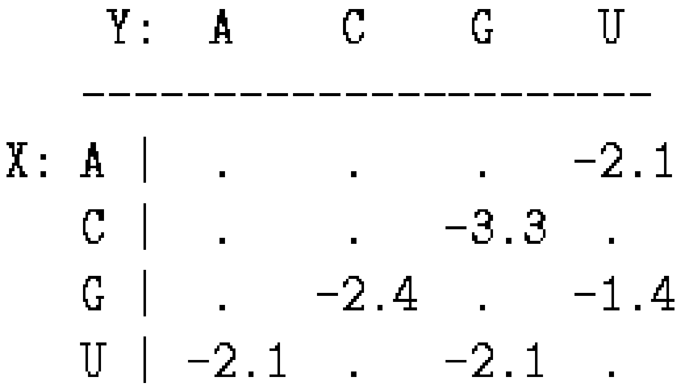


Figure 4 Energy is continuously given out as the RNA molecule folds by pairing complementary bases



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

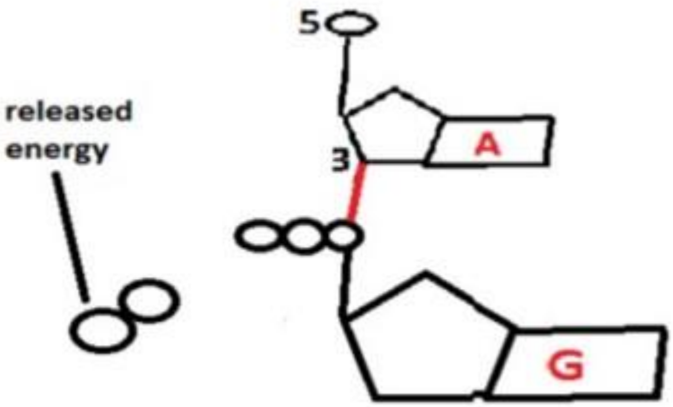
We can calculate the overall energy of RNA structures by summing up energies given out during the process of folding. For knowing the positive and negative values of calculations of stabilizing and destabilizing energies we may factor in ways in which RNA can be destabilized.

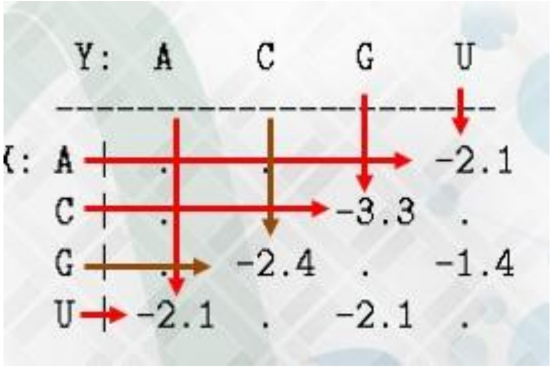
Figure 5 calculation of stabilizing and destabilizing values

Module 005: CALCULATING ENERGIES OF FOLDING-AN EXAMPLE

RNA is composed of four nucleotides (A, U, C and G) and these nucleotides are attached with ribose sugar in backbone. And these nucleotides have hydrogen bonding between them. G always bond with C and Always bonds with U through hydrogen bonding and energy is released.

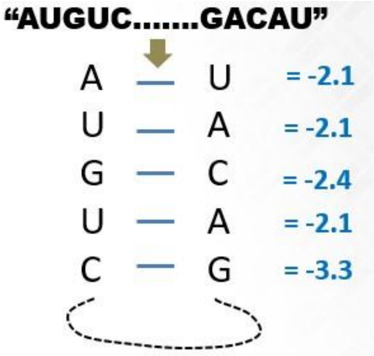
That’s why RNA molecule become more stable.





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

Figure 0.2nucleotides are held together by strong bonds which are created by the release of energy



“AUGUC…….GACAU”

A — U = -2.1
U — A = -2.1
G — C = -2.4
U — A = -2.1
C — G = -3.3

5 nucleotides formed H-Bonds. This bond formation released energy (-12.0 kcal/mol) RNA molecule took up a 2' structure. Hence became more stable.

Module 006: TYPES OF RNA SECONDARY STRUCTURES-I

All the complimentary bases of RNA combine together to form RNA secondary structures. A simple nucleotide sequences of RNA is called as Primary structure and denoted by 1' while when these nucleotides fold together and form a complex structure that is called secondary structure and denoted by 2'.

The preferred structure of RNA is 2' which has many structural patterns like **Helices, Loops, Bulges and Junctions**

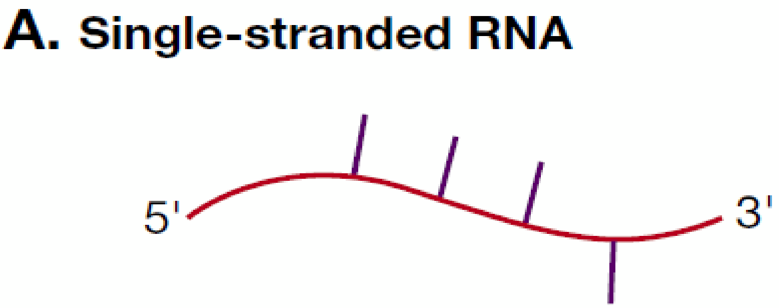
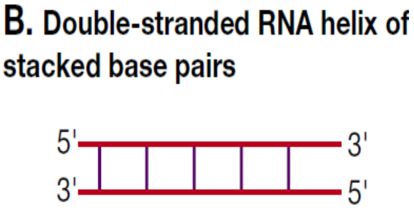
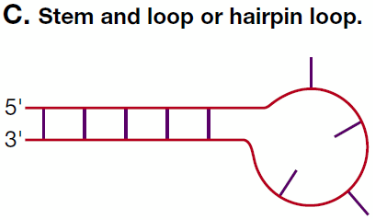


Figure 8 RNA sequence extends from its 5' end to 3' end. Upon folding, 3' end may fold on to the 5' end

The first 2' RNA structure is called helix. Unlike the DNA helix, the RNA helix is formed when the RNA folds onto itself.



The second 2' structure is the hairpin loop



The loop of the hairpin must at least four bases long to avoid steric hindrance with base-pairing in the stem part of the structure.

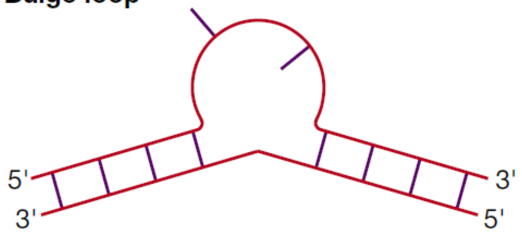
Note that hairpins reverses the chemical direction of the RNA molecule.

Module 007: TYPES OF RNA SECONDARY STRUCTURES-II

RNA 1’ structure fold the (5’-3’) ends and make RNA 2’ structure just like helix and hairpin structure.

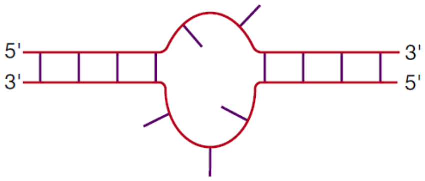
The third type of 2’ structure is bulge loop.

D. Bulge loop



Bulges, are formed when a double-stranded region cannot form base pairs perfectly. Bulges can be asymmetric with varying number of base pairs on one side of the loop. Bulge loops are commonly found in helical segments of cellular RNAs and used to measure the helical twist of RNA in solution. (Tang and Draper 1990)The forth type of 2’ RNA structure is interior loop.

E. Interior loop



Interior loops are formed by an asymmetric number of unpaired bases on each side of the loop.(Turner, Sugimoto et al. 1988)

Module 008: TYPES OF RNA SECONDARY STRUCTURES-III

Another 2’ RNA structure is the Junction or Intersection.

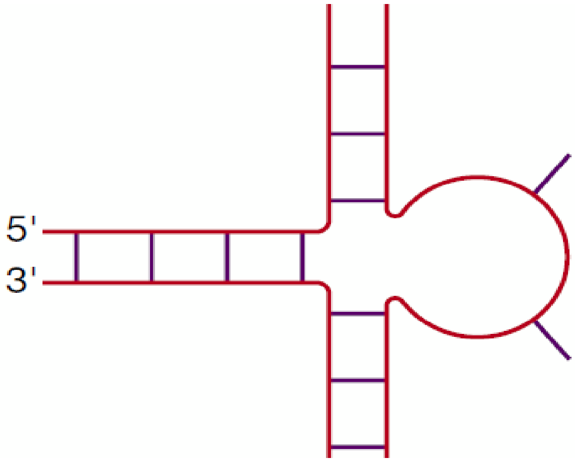


Figure 9 2' RNA structure called junction

Junctions include two or more double-stranded regions converging to form a closed structure. The unpaired bases appear as a bulge.(Zuker and Sankoff 1984)

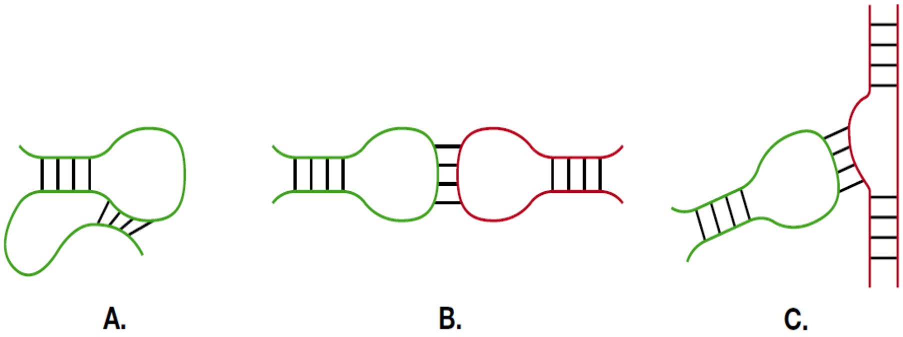


Figure 10 Unpaired bases in two 2’ structures form hydrogen bonds with each other

RNA tertiary structures are formed when RNA unpaired base bond in 2’ region bond.

Module 009: RNA TERTIARY STRUCTURES

2’ RNA structures is formed due to folding of nucleotide with in RNA molecule but after folding some nucleotides remain open for interaction. And they form hydrogen bonds together.

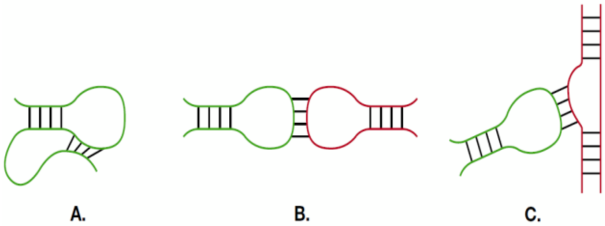
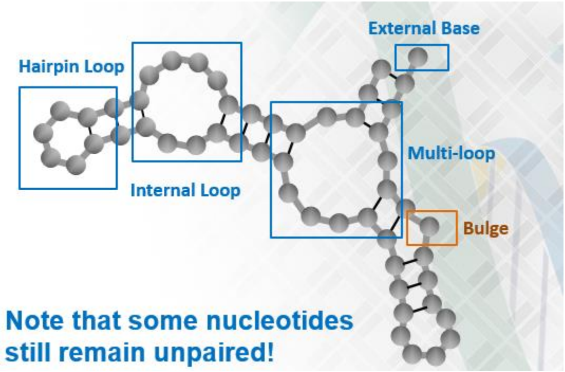


Figure 11 Hydrogen bonding formation in open nucleotides.

These unpaired nucleotides of 2’ structure interact with other unpaired nucleotides and form a third structure called tertiary 3’ structure. For example 4 nucleotides in hairpin loop structure does.



Note that some nucleotides still remain unpaired!

The above figure:

1. Indicate how these 2’ structures come together
2. Indicate the difference between internal loop and multi loop
3. Indicate the yet unpaired bases

The unpaired bases in 3’ structure remain paired by abnormal folding called (pseudoknots) but instead of pairing they remain available or pairing.

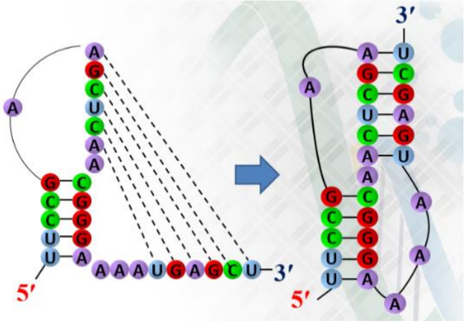
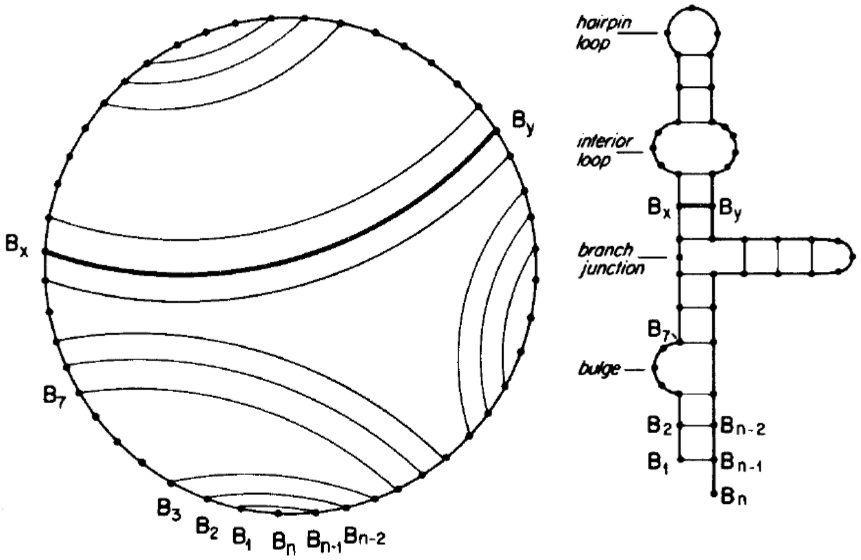


Figure 12 pseudoknots

Module 010: CIRCULAR REPRESENTATION OF STRUCTURES

Tertiary or 3’ structure of RNA may form pseudoknots to detect the pseudoknots in RNA structure we need “circular plot” which is a graphical approach.



Intersecting arcs in circular plot are the pseudoknot.

Module 011: EXPERIMENTAL METHODS TO DETERMINE RNA STRUCTURES

RNA has 1’, 2’ and 3’ structures. 1’ has simple nucleotide sequence and 2’ has nucleotides folding and 3’ has knots.

For measuring the RNA structure we use **X-ray crystallography** (Smyth and Martin 2000), which works according to the principle of diffraction. Crystallized RNA diffracts X-rays which helps estimate atomic positions

All isotopes that contain an odd number of protons and/or of neutrons (see Isotope) have an intrinsic magnetic moment and angular momentum, in other words a nonzero spin, while all nuclides with even numbers of both have a total spin of zero. The most commonly studied nuclei are 1H and 13C, al

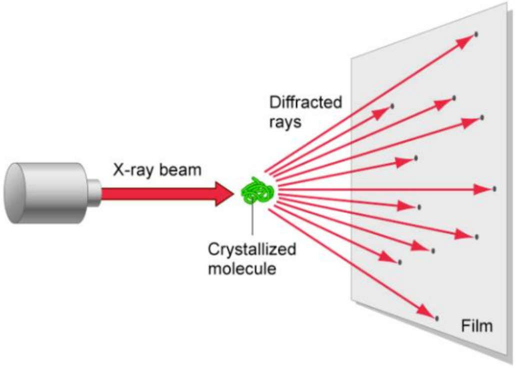


Figure 13 X-ray Crystallography
https://260h.pbworks.com/w/page/30814223/X%20Ray%20Crystallography

Another method to measure the RNA structure is called as **Atomic Force microscopy** in this technique a laser connected to a Si_3N_4 piezoelectric probe scans an RNA sample. It works well in air and liquid environment.

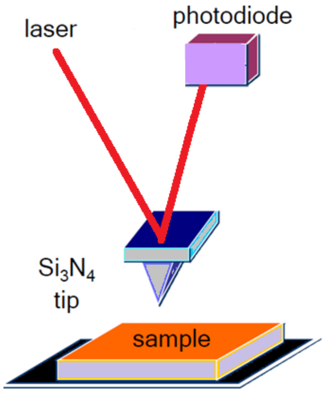
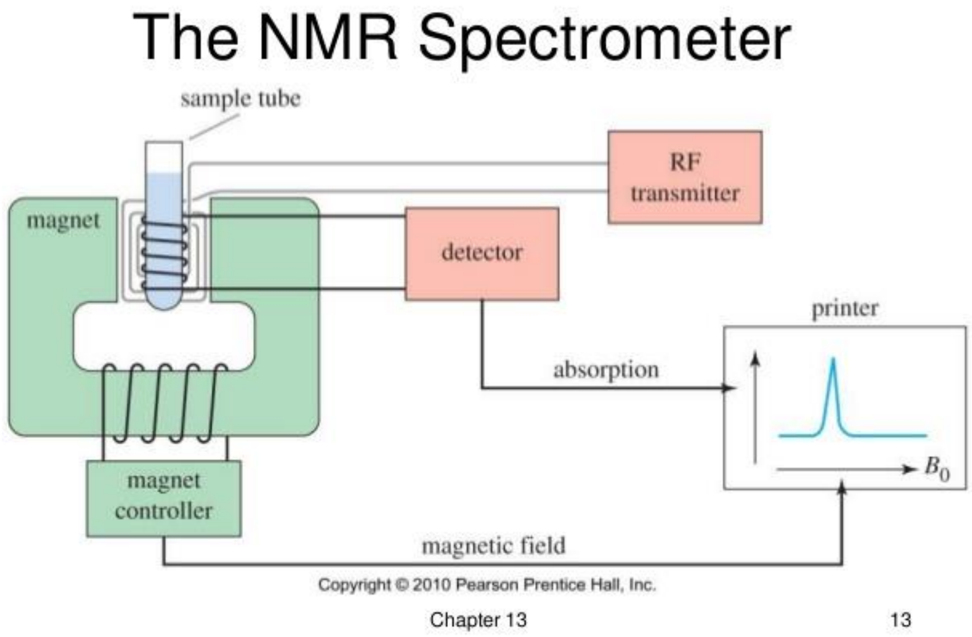


Figure 14 Atomic microscopy

The third method for measuring the RNA structure is **Nuclear Magnetic Resonance Imaging** in this method Hydrogen atoms in RNA resonate upon placement in a high magnetic field. It Works well without crystallizing RNA



The NMR Spectrometer

Copyright © 2010 Pearson Prentice Hall, Inc.

Chapter 13 13

http://www.slideshare.net/Oatsmith/13-nuclear-magnetic-resonance-spectroscopy-wade-7th

- **STORAGE OF STRUCTURES**

Reported structures are stored in online databases. Example includes RNA Bricks and RMDB etc. Bioinformaticians can refer to these databases for RNA structure studies

RNA Bricks is a database of RNA 3D structure motifs and their contacts, both with themselves and with proteins

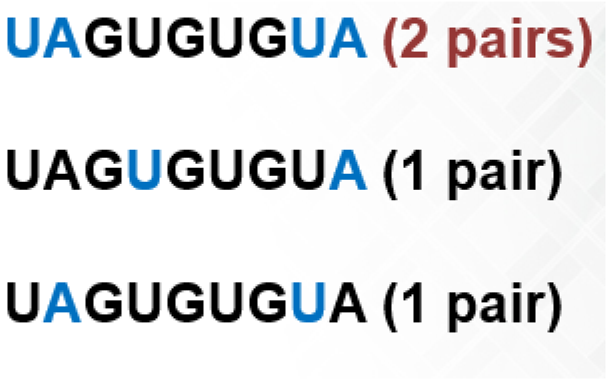
Stanford University's RNA Mapping Database is an archive that contains results of diverse structural mapping experiments performed on ribonucleic acids.

Module 012: Strategies for RNA Structure Prediction

RNA structure 2' and 3' can be measured experimentally, but RNA molecule readily degrade due to their short shelf life.

Give 1’ RNA structure creates the 2’ structure because the simple nucleotides folds and form 2’ structure. And on the base of folding we can predict the stability of the RNA molecule.

For example.



UAGUGUGUA (2 pairs)

UAGUGUGUA (1 pair)

UAGUGUGUA (1 pair)

Figure 15 pairs represent the stability of the RNA molecule

Maximizing the number of nucleotides can increase the structure and we have to select the structure according to the stability.

Module 013: Dot Plots for RNA 2' Structure Prediction

Structure measurement through experiments is slow and costly and there is maximum chances of more than one structure existence.

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:

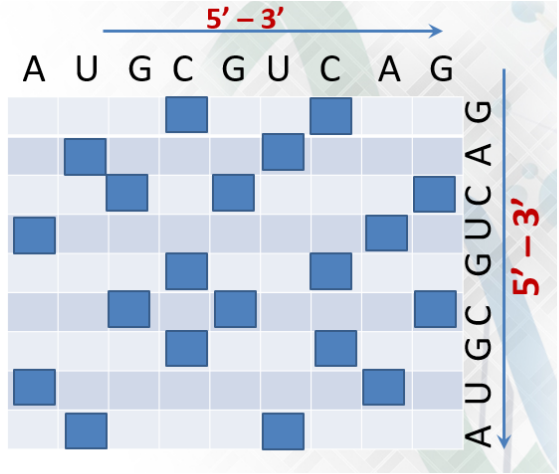


Figure 16 dot are placed at complimentary base pair place.

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

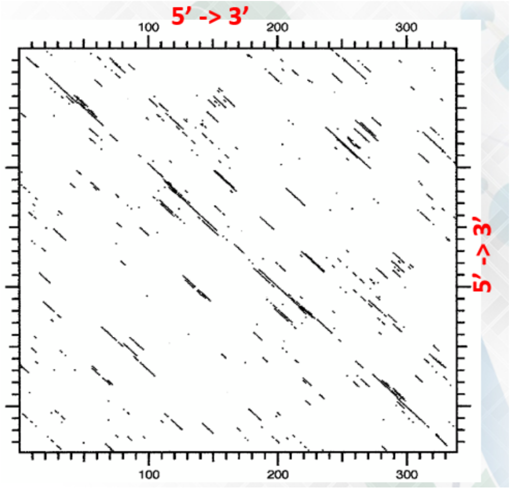


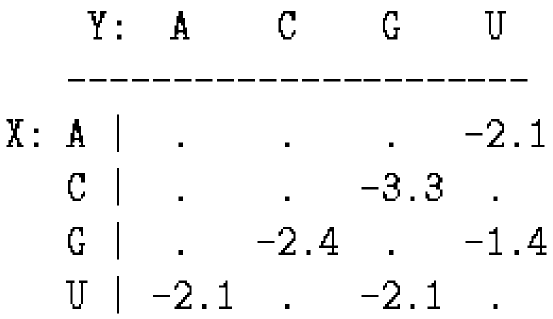
Figure 17 Potato Tuber Spindle Viroid

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.

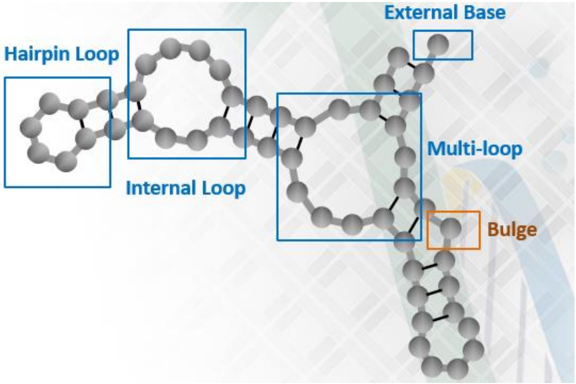


Figure 19 Hairpin+IntLoop+ExtLoop+Bulge+Hairpin

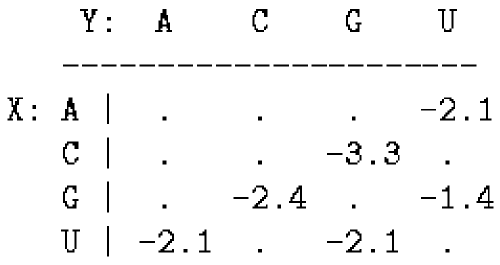
- **SUM OF ENERGIES**

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

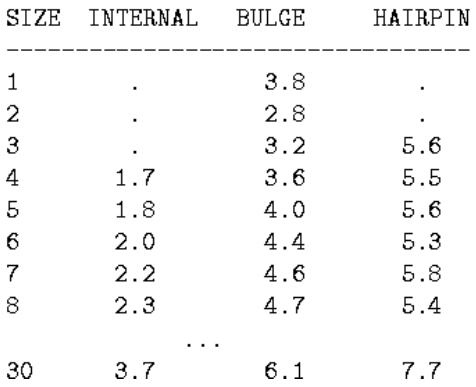
Energy based methods involve evaluating the free energy structures. To compute the RNA sequence for 1’ or 2’ optimal structure prediction we use Zuker’s Algorithm.

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

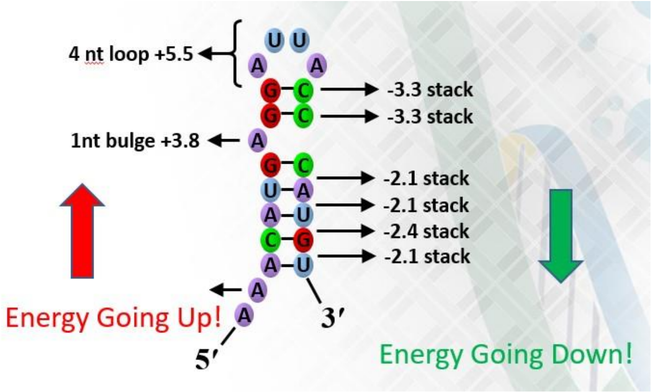
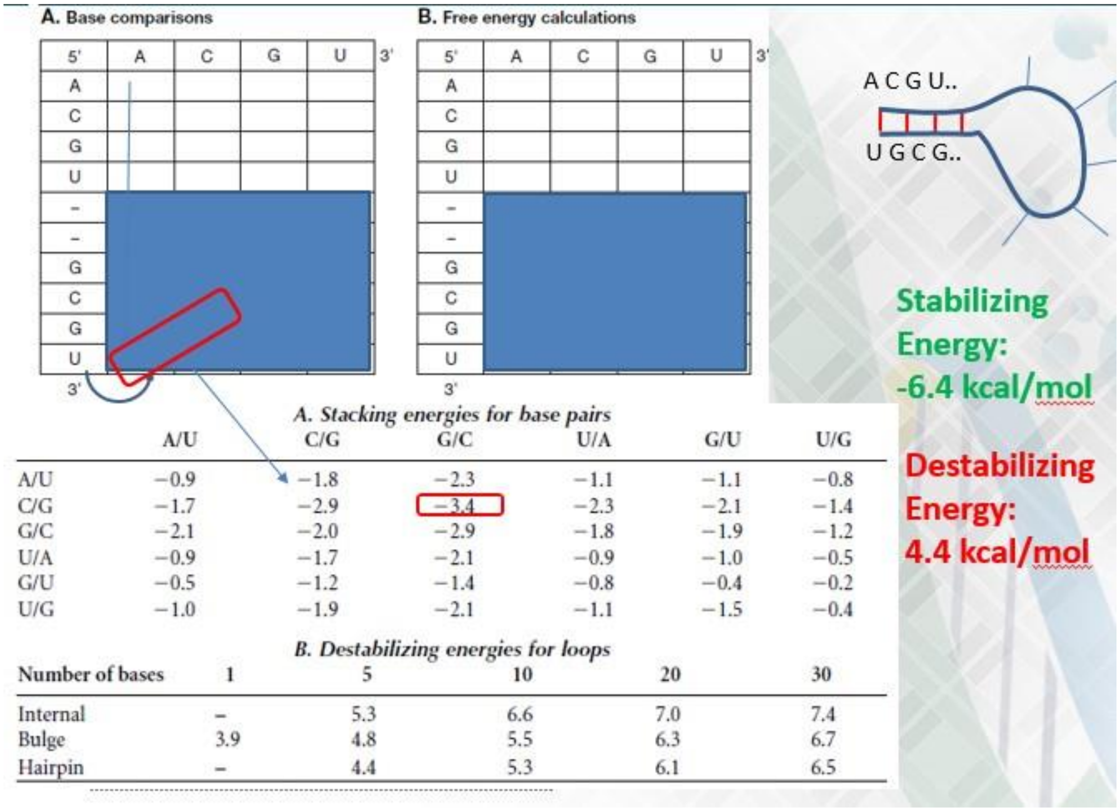


Figure 22 working principle of Zuker’s Algorithm (2003)

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!



A. Stacking energies for base pairs

	A/U	C/G	G/C	U/A	G/U	U/G
A/U	−0.9	−1.8	−2.3	−1.1	−1.1	−0.8
C/G	−1.7	−2.9	−3.4	−2.3	−2.1	−1.4
G/C	−2.1	−2.0	−2.9	−1.8	−1.9	−1.2
U/A	−0.9	−1.7	−2.1	−0.9	−1.0	−0.5
G/U	−0.5	−1.2	−1.4	−0.8	−0.4	−0.2
U/G	−1.0	−1.9	−2.1	−1.1	−1.5	−0.4

B. Destabilizing energies for loops

Number of bases	1	5	10	20	30
Internal	–	5.3	6.6	7.0	7.4
Bulge	3.9	4.8	5.5	6.3	6.7
Hairpin	–	4.4	5.3	6.1	6.5

Stabilizing Energy: -6.4 kcal/mol

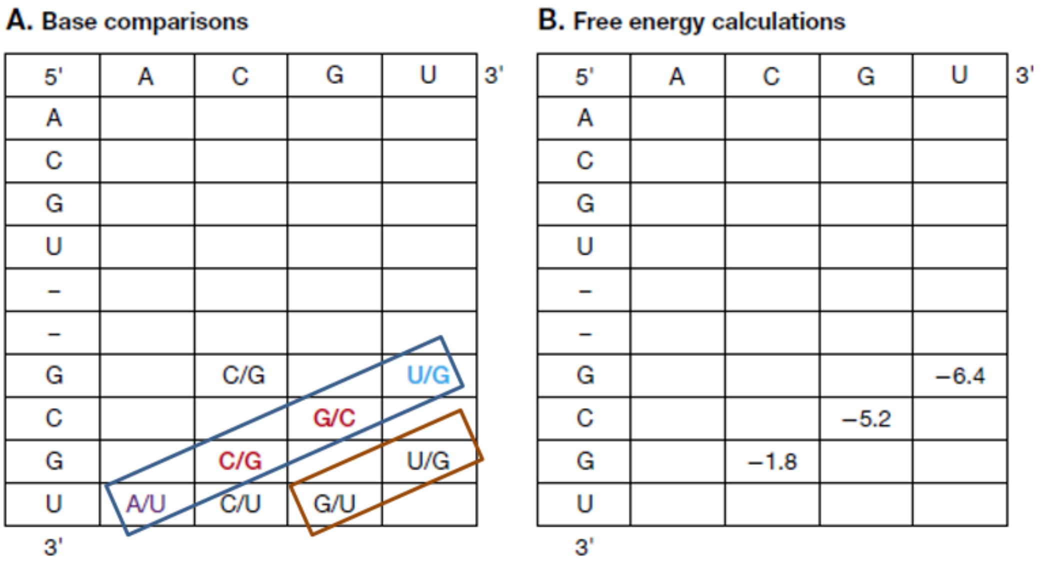
Destabilizing Energy: 4.4 kcal/mol

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
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
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:

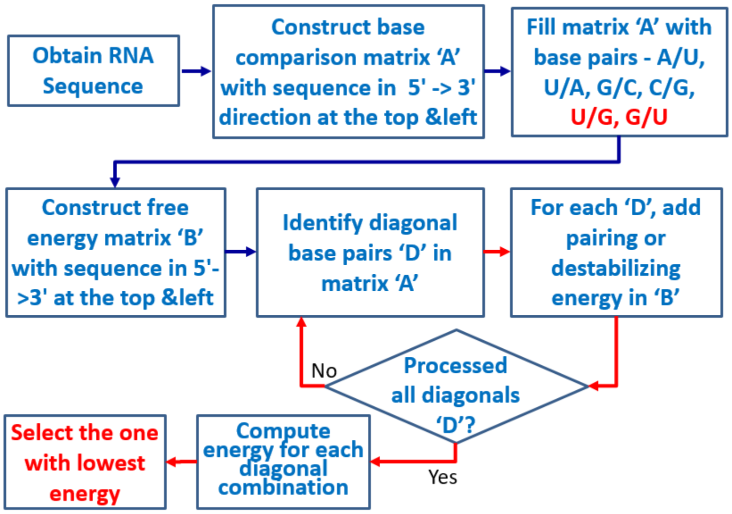

Obtain RNA Sequence
Construct base comparison matrix 'A' with sequence in 5' -> 3' direction at the top &left
Fill matrix 'A' with base pairs - A/U, U/A, G/C, C/G, U/G, G/U
Construct free energy matrix 'B' with sequence in 5'->3' at the top &left
Identify diagonal base pairs 'D' in matrix 'A'
For each 'D', add pairing or destabilizing energy in 'B'
Processed all diagonals 'D'?
No
Yes
Compute energy for each diagonal combination
Select the one with lowest energy

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.

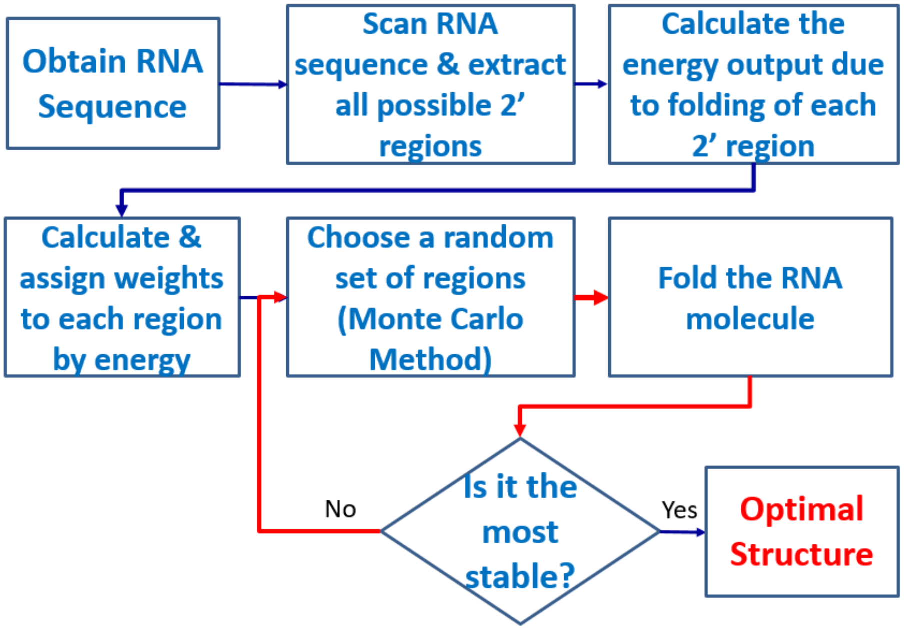

Obtain RNA Sequence
Scan RNA sequence & extract all possible 2' regions
Calculate the energy output due to folding of each 2' region
Calculate & assign weights to each region by energy
Choose a random set of regions (Monte Carlo Method)
Fold the RNA molecule
Is it the most stable?
No
Yes
Optimal Structure

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 us 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:

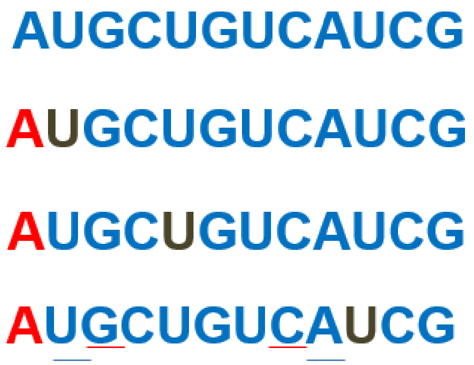


Figure 26 all possible complementary bases combinations.

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.

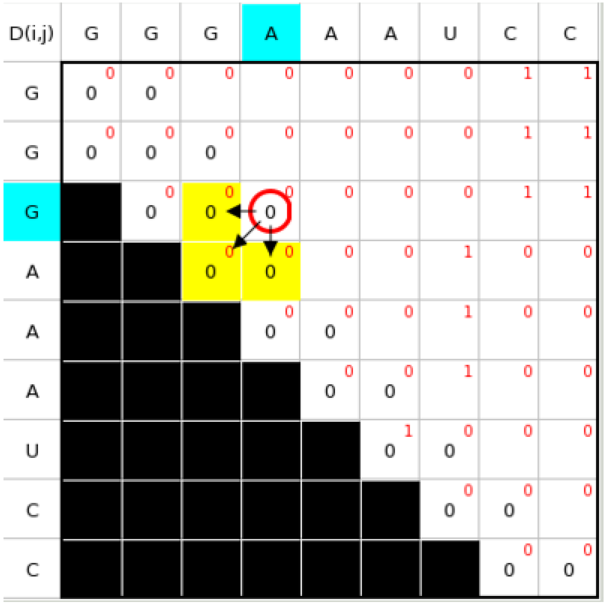
The score *S (i , j)* is the maximum of the following four possibilities

$$S(i,j) = \max \begin{cases} S(i+1,j) & \text{Lower Element} \quad r_i \ \textit{unpaired} \\ S(i,j-1) & \text{Left Element} \quad r_j \ \textit{unpaired} \\ S(i+1,j-1) + e(r_i,r_j) & \text{Energy of pairing} \quad i,j \ \textit{base pair} \\ \max\limits_{i<k<j} \{S(i,k) + S(k+1,j)\} & i,j \ \textit{paired, but not to each other} \end{cases}$$

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	0	0	1	1
G	0 0	0 0	0 0	0	0	0	0	1	1
G		0 0	0 0	0 0	0	0	0	1	1
A			0 0	0 0	0	0	1	0	0
A				0 0	0 0	0	1	0	0
A					0 0	0 0	1	0	0
U						1 0	0 0	0	0
C							0 0	0 0	0
C								0 0	0 0

Figure 0.3 for maximum score 4 positions are used in scoring

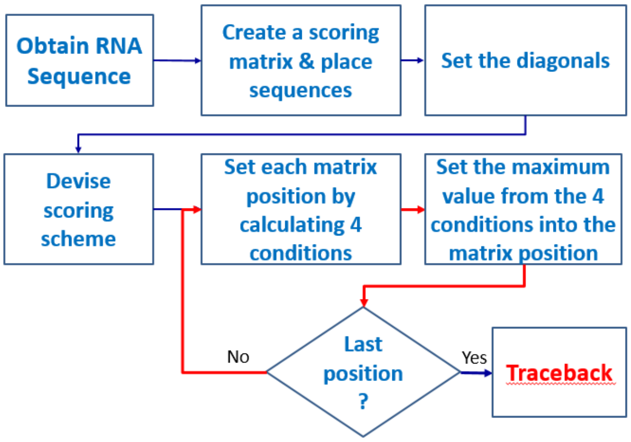

Obtain RNA Sequence
Create a scoring matrix & place sequences
Set the diagonals
Devise scoring scheme
Set each matrix position by calculating 4 conditions
Set the maximum value from the 4 conditions into the matrix position
Last position ?
No
Yes
Traceback

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

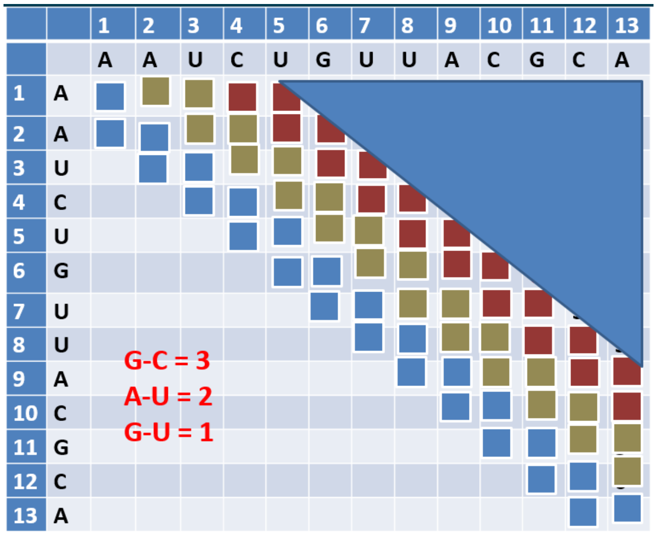
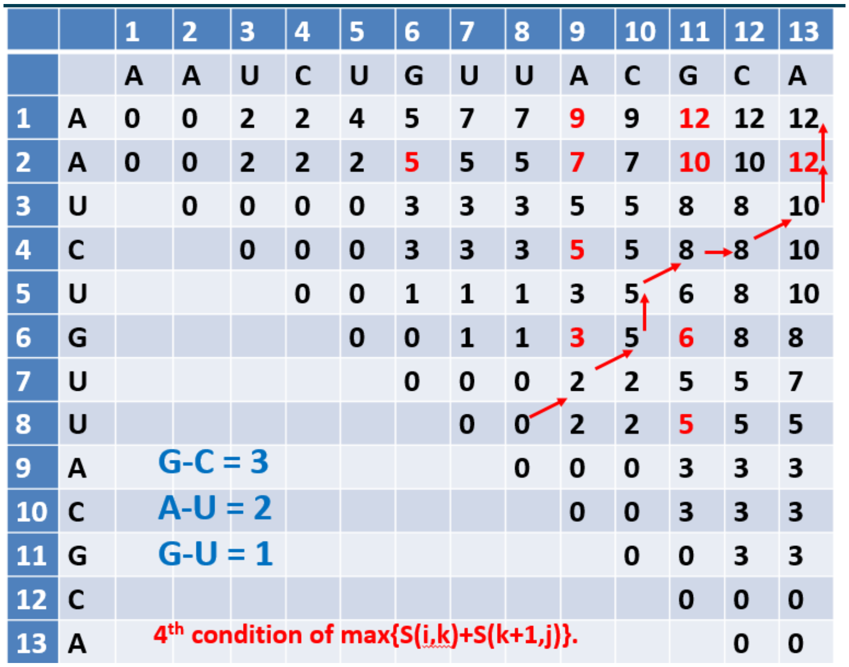


Figure 29 N-J Algorithm scoring

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:

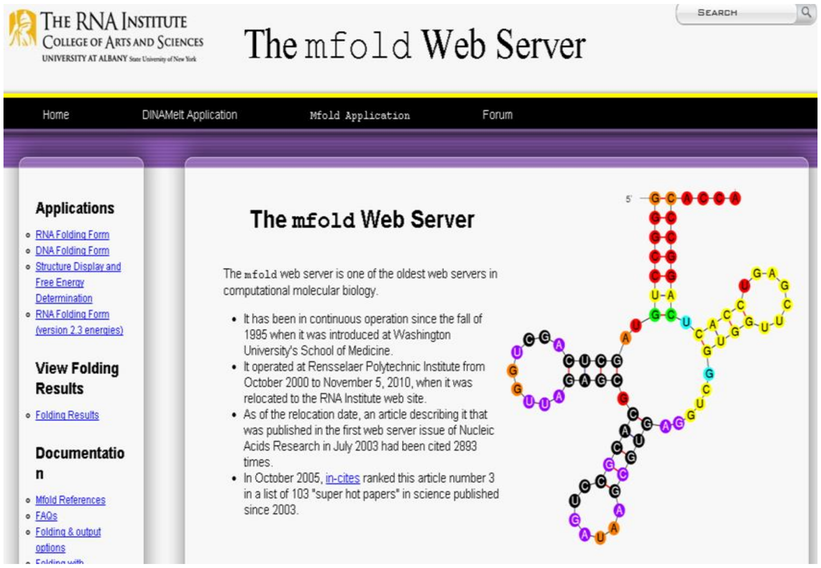
- ➢ Sequences
- ➢ Comparison
- ➢ Nucleotide
- ➢ Covariance analysis

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.

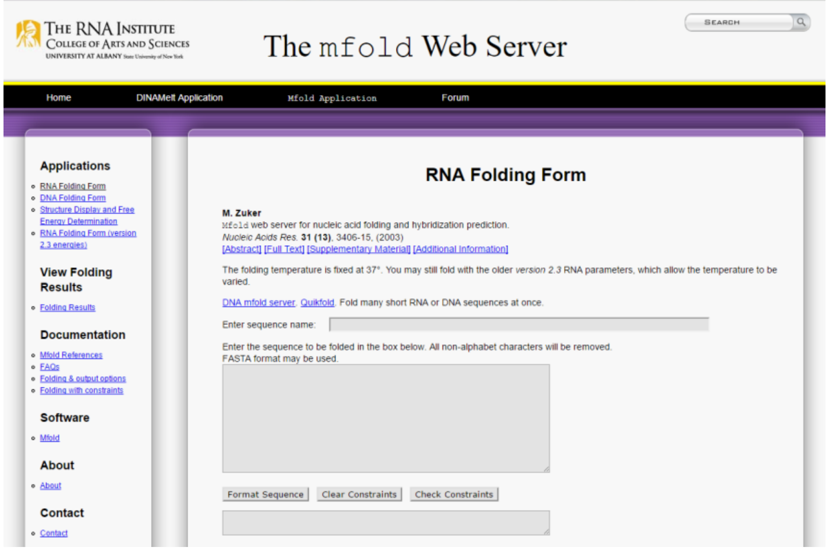


The mfold Web Server

The mfold web server is one of the oldest web servers in computational molecular biology.

- It has been in continuous operation since the fall of 1995 when it was introduced at Washington University's School of Medicine.
- It operated at Rensselaer Polytechnic Institute from October 2000 to November 5, 2010, when it was relocated to the RNA Institute web site.
- As of the relocation date, an article describing it that was published in the first web server issue of Nucleic Acids Research in July 2003 had been cited 2893 times.
- In October 2005, in-cites ranked this article number 3 in a list of 103 "super hot papers" in science published since 2003.

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



RNA Folding Form

M. Zuker
Mfold web server for nucleic acid folding and hybridization prediction.
Nucleic Acids Res. **31 (13)**, 3406-15, (2003)
[Abstract] [Full Text] [Supplementary Material] [Additional Information]

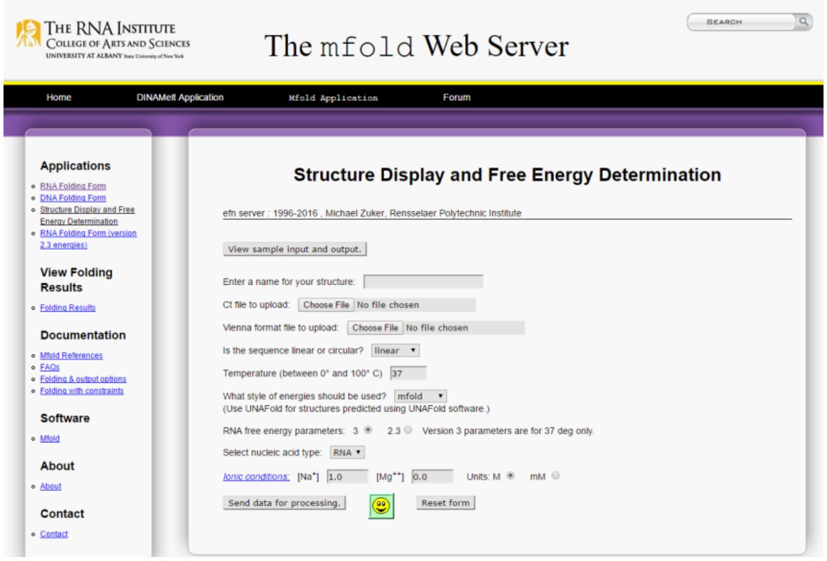
The folding temperature is fixed at 37°. You may still fold with the older *version 2.3* RNA parameters, which allow the temperature to be varied.

DNA mfold server. Quikfold. Fold many short RNA or DNA sequences at once.

Enter sequence name:

Enter the sequence to be folded in the box below. All non-alphabet characters will be removed.
FASTA format may be used.

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



The mfold Web Server

Structure Display and Free Energy Determination

efn server : 1996-2016 , Michael Zuker, Rensselaer Polytechnic Institute

Enter a name for your structure:

Ct file to upload: No file chosen

Vienna format file to upload: No file chosen

Is the sequence linear or circular? linear

Temperature (between 0° and 100° C) 37

What style of energies should be used? mfold
(Use UNAFold for structures predicted using UNAFold software.)

RNA free energy parameters: 3 ● 2.3 ○ Version 3 parameters are for 37 deg only.

Select nucleic acid type: RNA

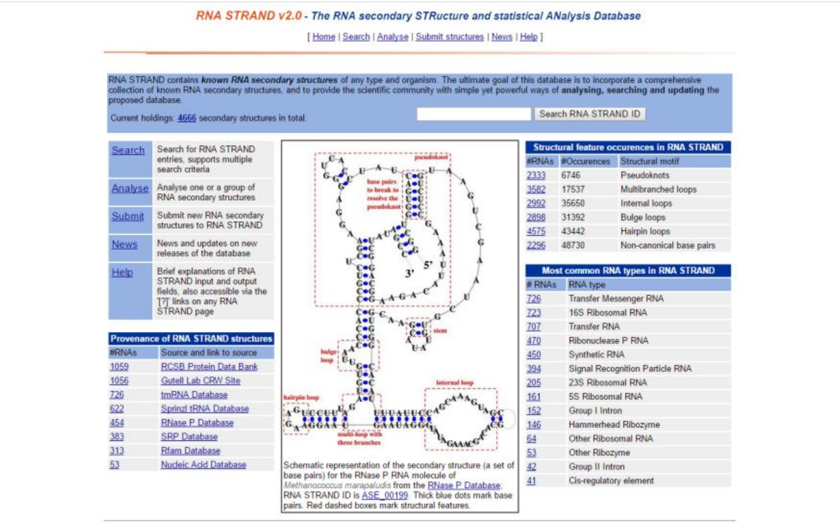
Ionic conditions: [Na$^+$] 1.0 [Mg^{++}] 0.0 Units: M ● mM ○

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’.



RNA STRAND v2.0 - *The RNA secondary STRucture and statistical ANalysis Database*

[Home | Search | Analyse | Submit structures | News | Help]

RNA STRAND contains ***known RNA secondary structures*** of any type and organism. The ultimate goal of this database is to incorporate a comprehensive collection of known RNA secondary structures, and to provide the scientific community with simple yet powerful ways of **analysing, searching and updating** the proposed database.

Current holdings: 4666 secondary structures in total.

Search	Search for RNA STRAND entries, supports multiple search criteria
Analyse	Analyse one or a group of RNA secondary structures
Submit	Submit new RNA secondary structures to RNA STRAND
News	News and updates on new releases of the database
Help	Brief explanations of RNA STRAND input and output fields, also accessible via the '[?]' links on any RNA STRAND page

Provenance of RNA STRAND structures

#RNAs	Source and link to source
1059	RCSB Protein Data Bank
1056	Gutell Lab CRW Site
726	tmRNA Database
622	Sprinzl tRNA Database
454	RNase P Database
383	SRP Database
313	Rfam Database
53	Nucleic Acid Database

pseudoknot
base pairs to break to resolve the pseudoknot
5’
3’
stem
bulge loop
internal loop
hairpin loop
multi-loop with three branches

Schematic representation of the secondary structure (a set of base pairs) for the RNase P RNA molecule of *Methanococcus marapaludis* from the RNase P Database; RNA STRAND ID is ASE_00199. Thick blue dots mark base pairs. Red dashed boxes mark structural features.

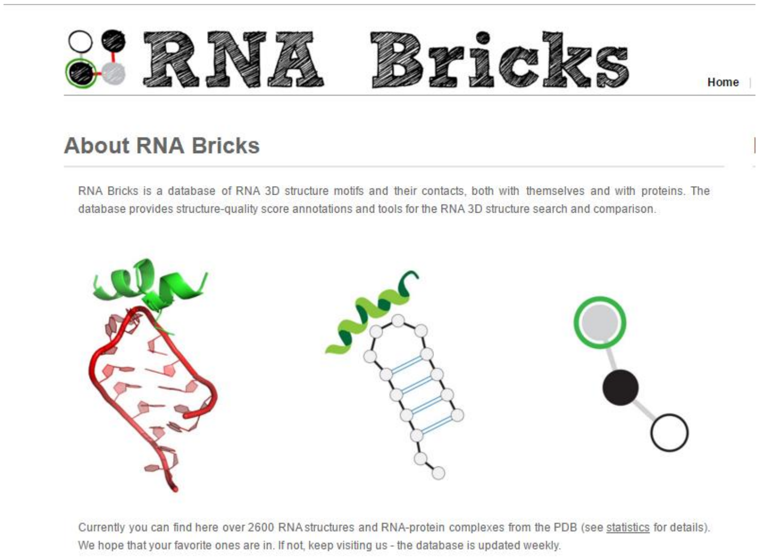
Structural feature occurences in RNA STRAND

#RNAs	#Occurences	Structural motif
2333	6746	Pseudoknots
3582	17537	Multibranched loops
2992	35650	Internal loops
2898	31392	Bulge loops
4575	43442	Hairpin loops
2296	48730	Non-canonical base pairs

Most common RNA types in RNA STRAND

# RNAs	RNA type
726	Transfer Messenger RNA
723	16S Ribosomal RNA
707	Transfer RNA
470	Ribonuclease P RNA
450	Synthetic RNA
394	Signal Recognition Particle RNA
205	23S Ribosomal RNA
161	5S Ribosomal RNA
152	Group I Intron
146	Hammerhead Ribozyme
64	Other Ribosomal RNA
53	Other Ribozyme
42	Group II Intron
41	Cis-regulatory element

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



About RNA Bricks

RNA Bricks is a database of RNA 3D structure motifs and their contacts, both with themselves and with proteins. The database provides structure-quality score annotations and tools for the RNA 3D structure search and comparison.

Currently you can find here over 2600 RNA structures and RNA-protein complexes from the PDB (see statistics for details). We hope that your favorite ones are in. If not, keep visiting us - the database is updated weekly.

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

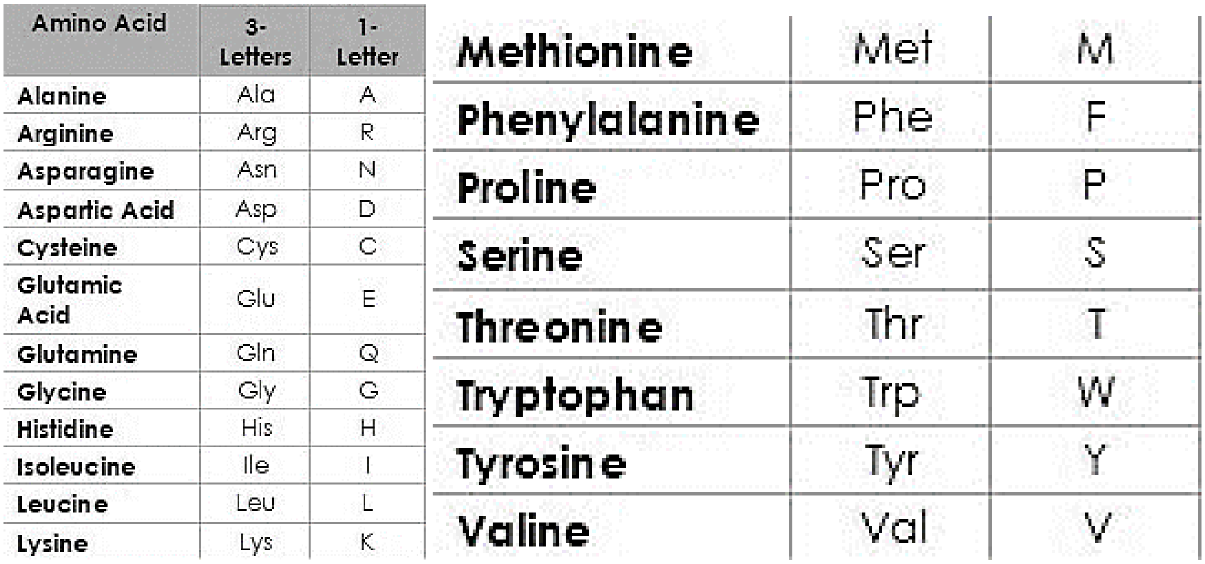
RNA 1’ folds and makes RNA 2’ structure and this online database is established for 2’ RNA structure and it act as dictionary for 2’ RNA structure.

Chapter 5 - Protein Sequences

Module 001: FROM DNA/RNA SEQUENCES TO PROTEIN

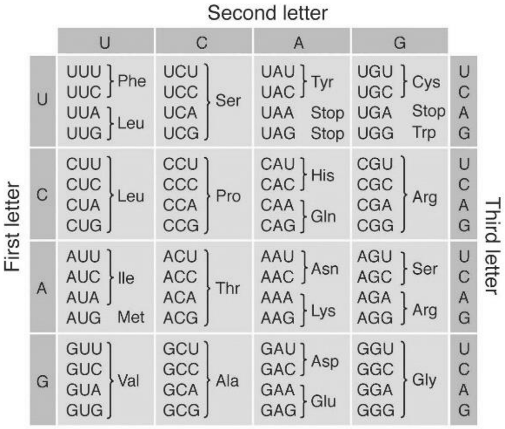
We are aware that DNA has four nucleotides bases **(A, C, T & G).** RNA contains **(A, C, U & G).** And protein contain 20 different amino acids. **DNA** to **RNA** then **Protein** is called as **central dogma.** Which includes translation, transcription and protein modifications.

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



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

Figure 0.1 amino acids letters information according to codons



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

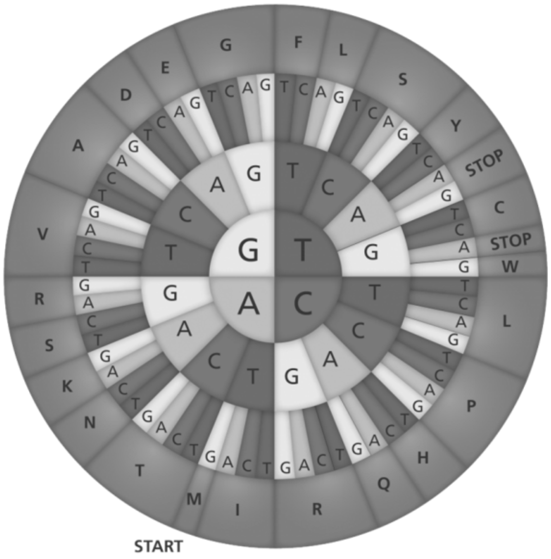
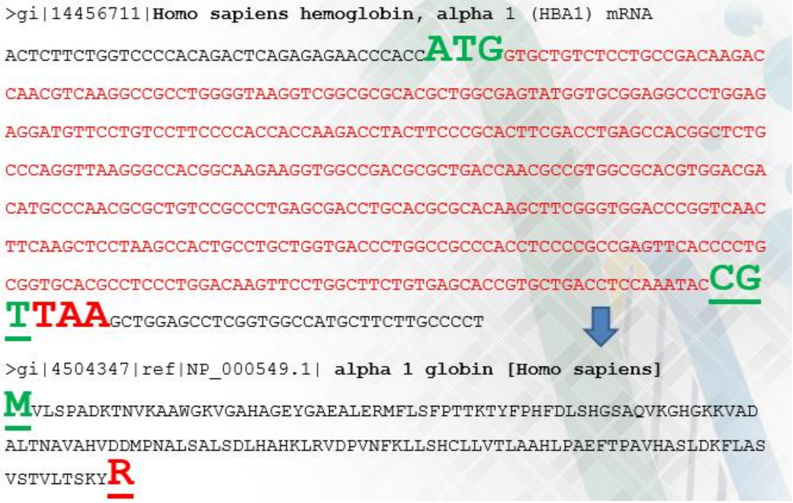


Figure 3 coding of amino acids



```
>gi|14456711|Homo sapiens hemoglobin, alpha 1 (HBA1) mRNA
ACTCTTCTGGTCCCCACAGACTCAGAGAGAACCCACCATGGTGCTGTCTCCTGCCGACAAGAC
CAACGTCAAGGCCGCCTGGGGTAAGGTCGGCGCGCACGCTGGCGAGTATGGTGCGGAGGCCCTGGAG
AGGATGTTCCTGTCCTTCCCCACCACCAAGACCTACTTCCCGCACTTCGACCTGAGCCACGGCTCTG
CCCAGGTTAAGGGCCACGGCAAGAAGGTGGCCGACGCGCTGACCAACGCCGTGGCGCACGTGGACGA
CATGCCCAACGCGCTGTCCGCCCTGAGCGACCTGCACGCGCACAAGCTTCGGGTGGACCCGGTCAAC
TTCAAGCTCCTAAGCCACTGCCTGCTGGTGACCCTGGCCGCCCACCTCCCCGCCGAGTTCACCCCTG
CGGTGCACGCCTCCCTGGACAAGTTCCTGGCTTCTGTGAGCACCGTGCTGACCTCCAAATACCG
TTAAGCTGGAGCCTCGGTGGCCATGCTTCTTGCCCCT

>gi|4504347|ref|NP_000549.1| alpha 1 globin [Homo sapiens]
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVAD
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

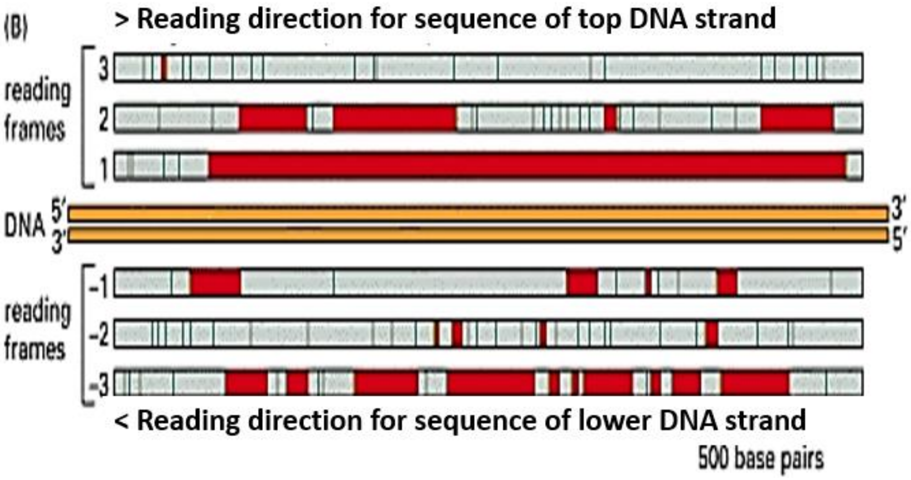
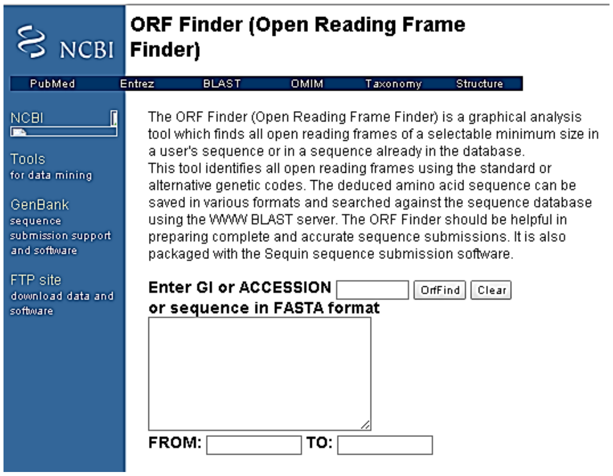

(B)
> Reading direction for sequence of top DNA strand
reading frames
3
2
1
DNA
5'
3'
3'
5'
reading frames
-1
-2
-3
< Reading direction for sequence of lower DNA strand
500 base pairs

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
OrfFind
Clear

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
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.

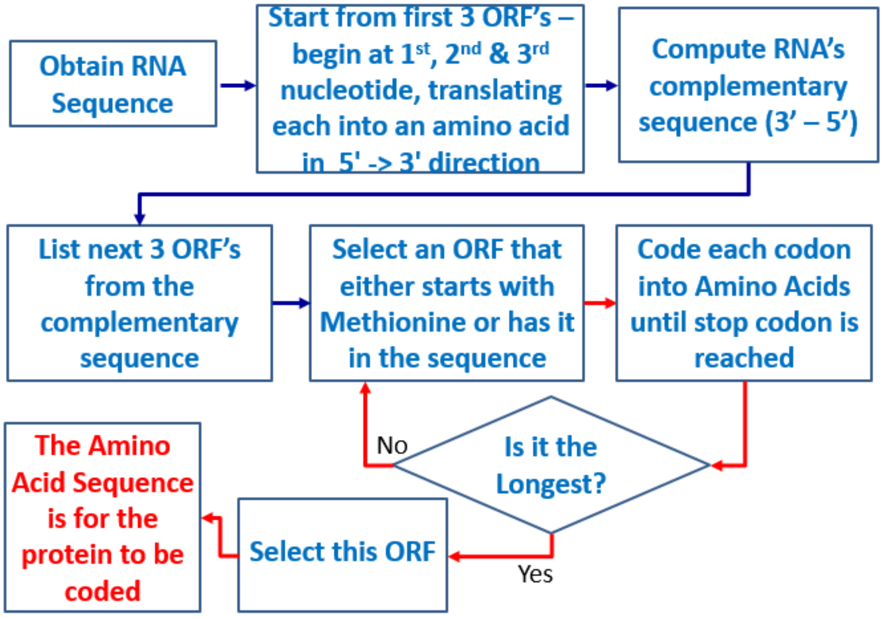

Obtain RNA Sequence
Start from first 3 ORF's – begin at 1st, 2nd & 3rd nucleotide, translating each into an amino acid in 5' -> 3' direction
Compute RNA's complementary sequence (3' – 5')
List next 3 ORF's from the complementary sequence
Select an ORF that either starts with Methionine or has it in the sequence
Code each codon into Amino Acids until stop codon is reached
Is it the Longest?
No
Yes
Select this ORF
The Amino Acid Sequence is for the protein to be coded

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 Adam 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 starting from the N-terminal and removing one amino acid at a time

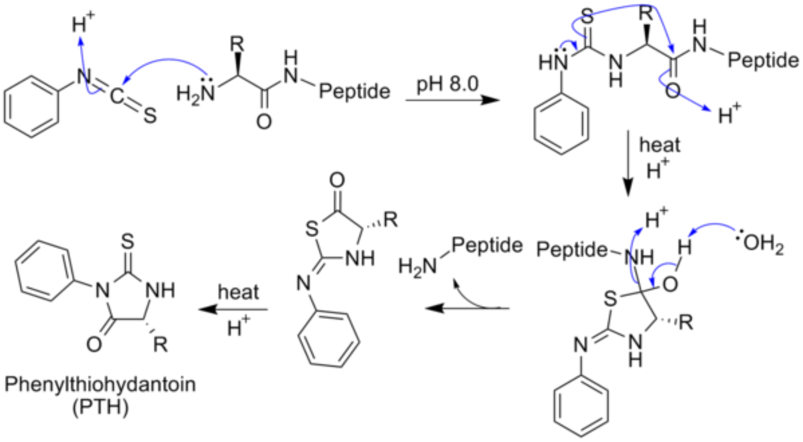
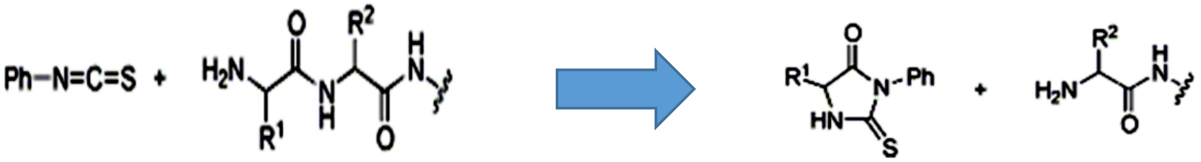


Figure 8 Mechanism

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.



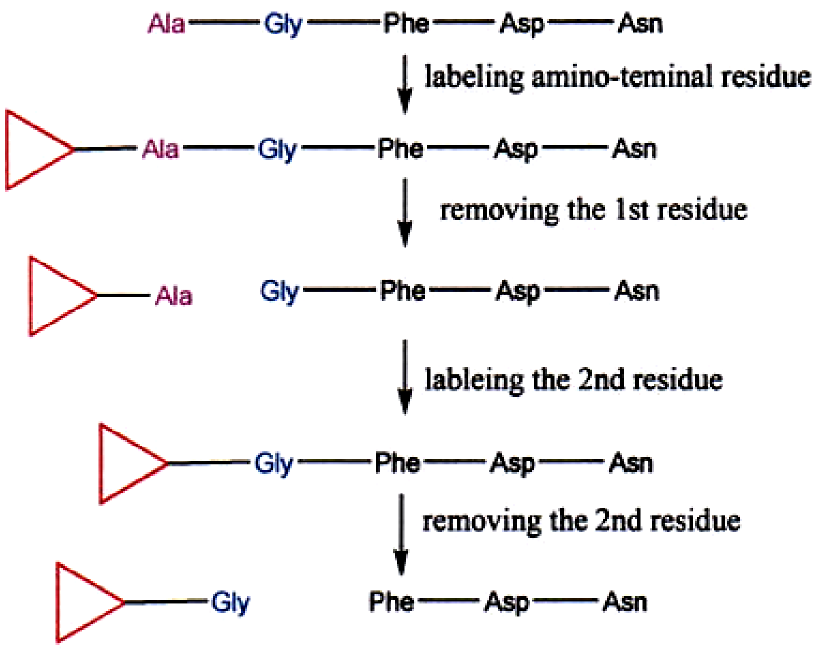


Figure 9 working of Edam degradation

- **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

Edam 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.

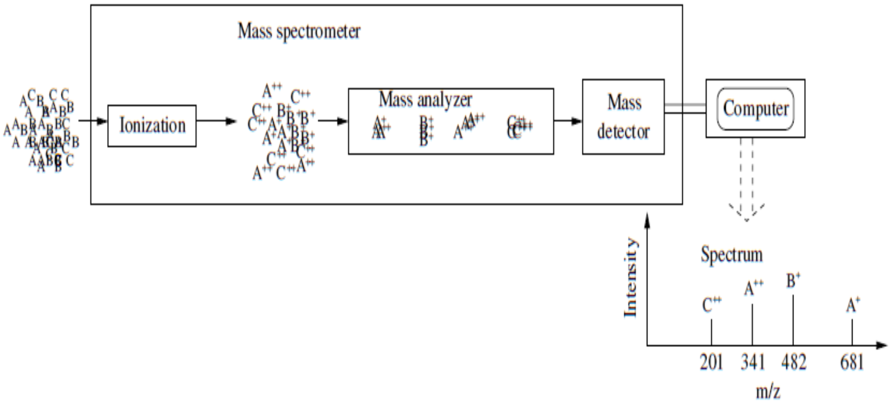
$$\mathbf{F} = Q(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \quad \text{(Lorentz force law)}$$

$$\mathbf{F} = m\mathbf{a} = m\frac{\mathrm{d}\mathbf{v}}{\mathrm{d}t} \quad \text{(Newton's second law of motion)}$$

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
- **v × 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

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

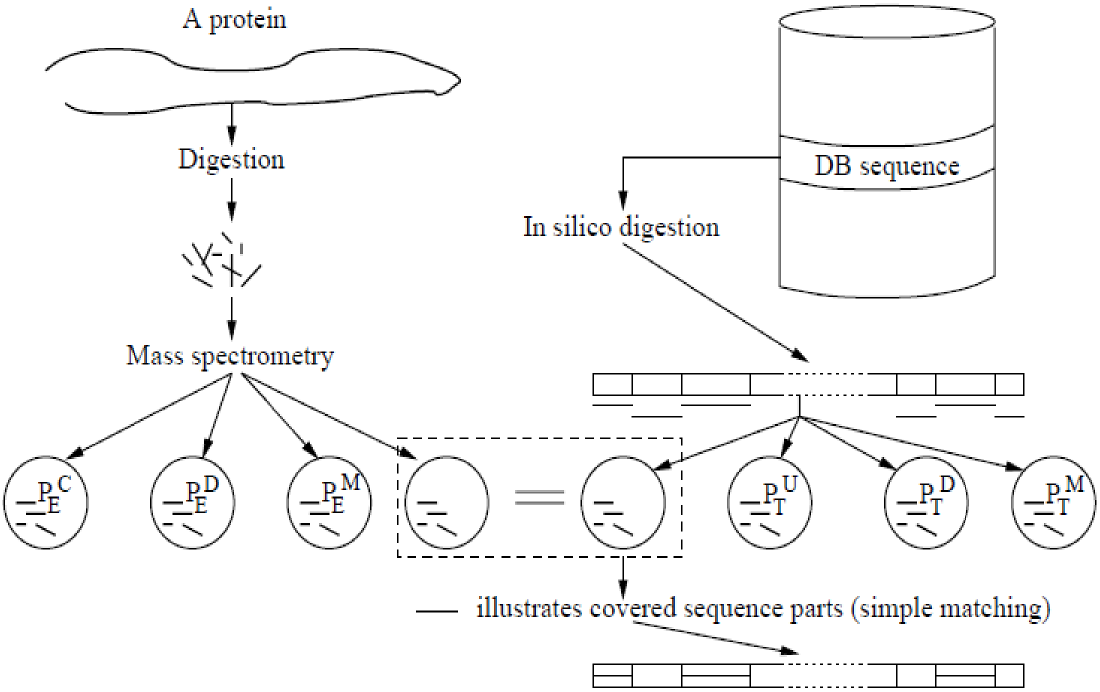
- **CONCLUSION**

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 Edam'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 than 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

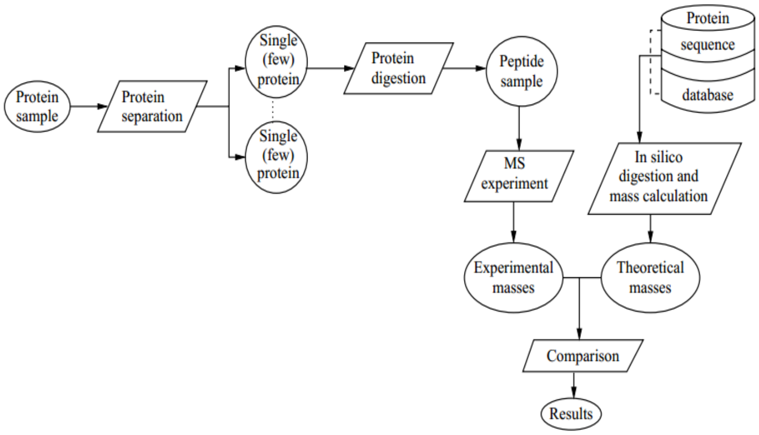


Figure 0.33 Peptide mass fingerprinting

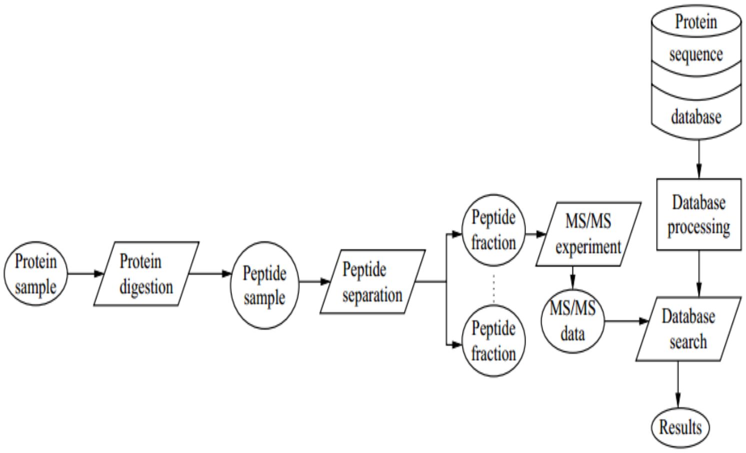


Figure 14 Shotgun Proteomics

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 masse.

- **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.

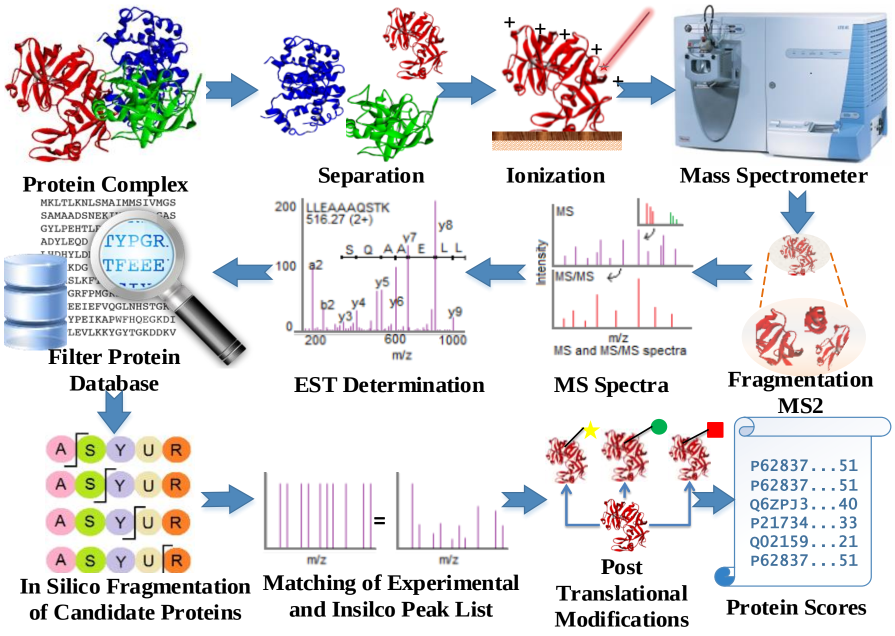
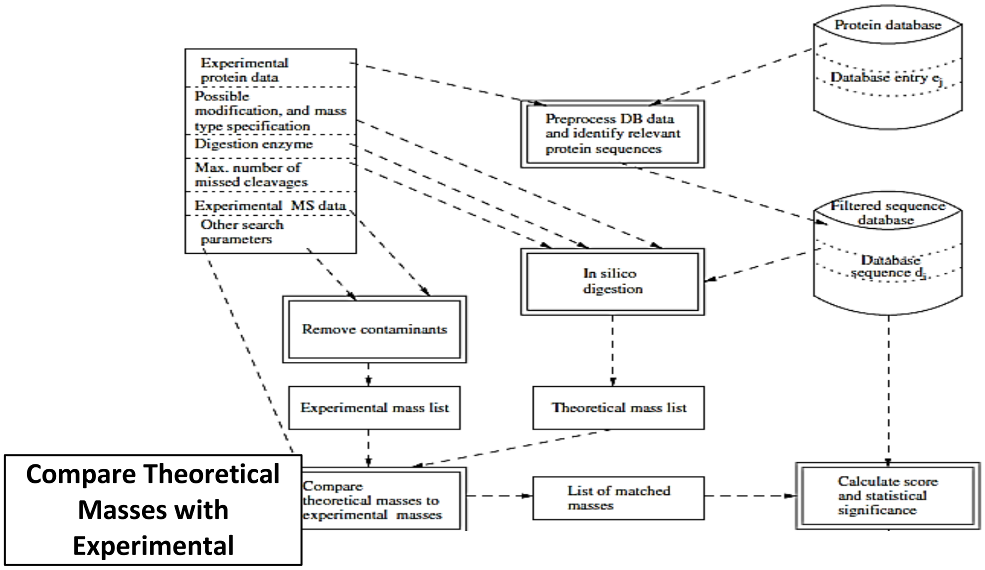


Figure 0.45 protein sequence identification flowchart


Experimental protein data
Possible modification, and mass type specification
Digestion enzyme
Max. number of missed cleavages
Experimental MS data
Other search parameters
Protein database
Database entry e_j
Preprocess DB data and identify relevant protein sequences
Filtered sequence database
Database sequence d_i
In silico digestion
Remove contaminants
Experimental mass list
Theoretical mass list
Compare theoretical masses to experimental masses
List of matched masses
Calculate score and statistical significance
Compare Theoretical Masses with Experimental

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

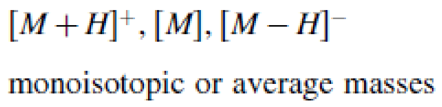
Module 013: PROTEIN IONIZATION TECHNIQUES

Protein ionization is used in Mass spectrometry based on proteomics protocols. Ionization involves loading of proton in protein or removal of protein. 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:



$[M+H]^+, [M], [M-H]^-$

monoisotopic or average masses

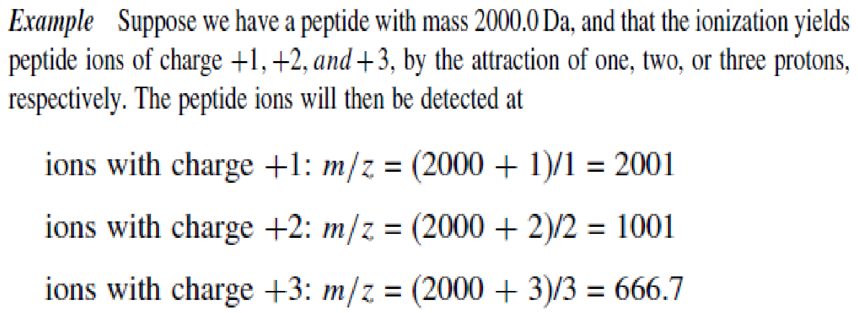
- **MALDI**

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

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

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

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

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.

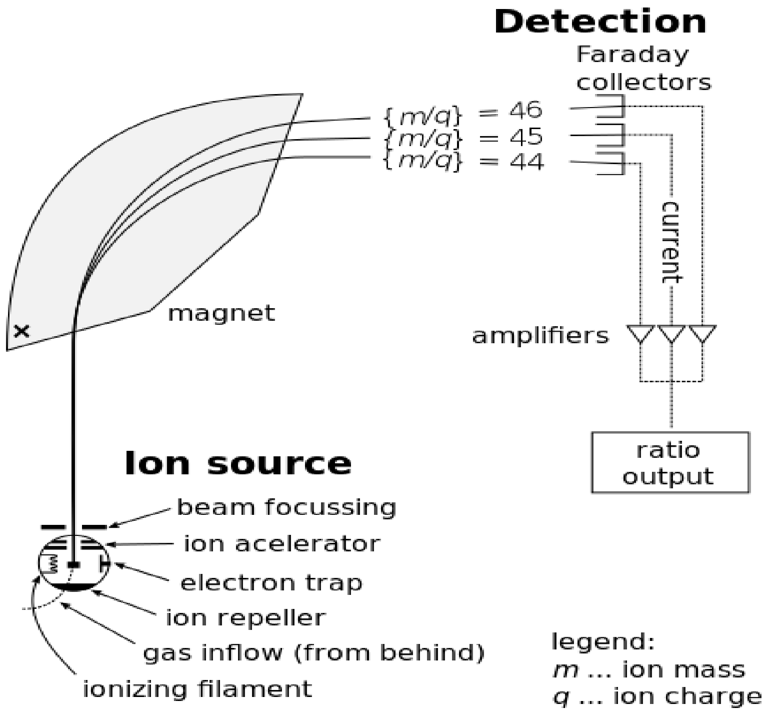


Figure 0.67 MS1 Schematic (Image courtesy Wikipedia)

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**

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

***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.

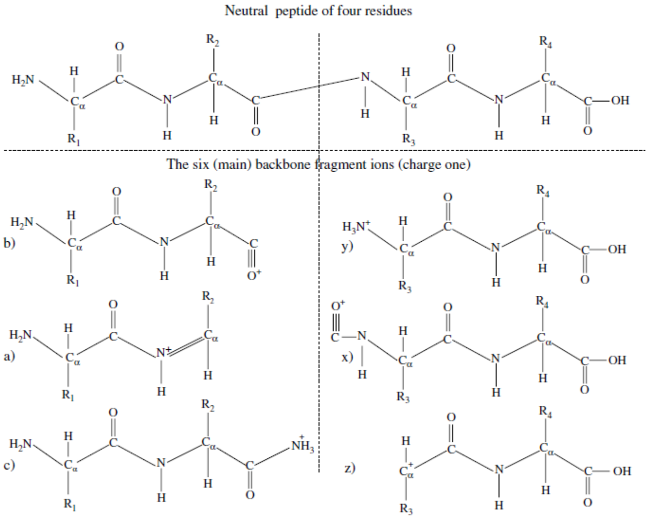
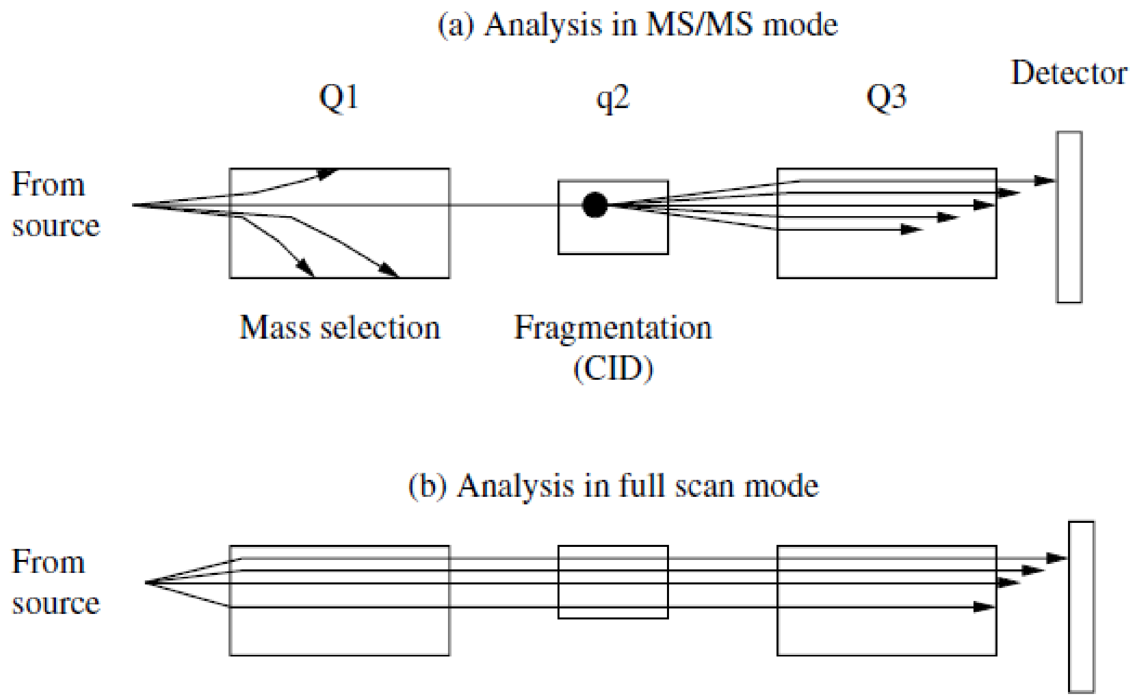

Neutral peptide of four residues
The six (main) backbone fragment ions (charge one)
b)
y)
a)
x)
c)
z)

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.



(a) Analysis in MS/MS mode

(b) Analysis in full scan mode

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 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.

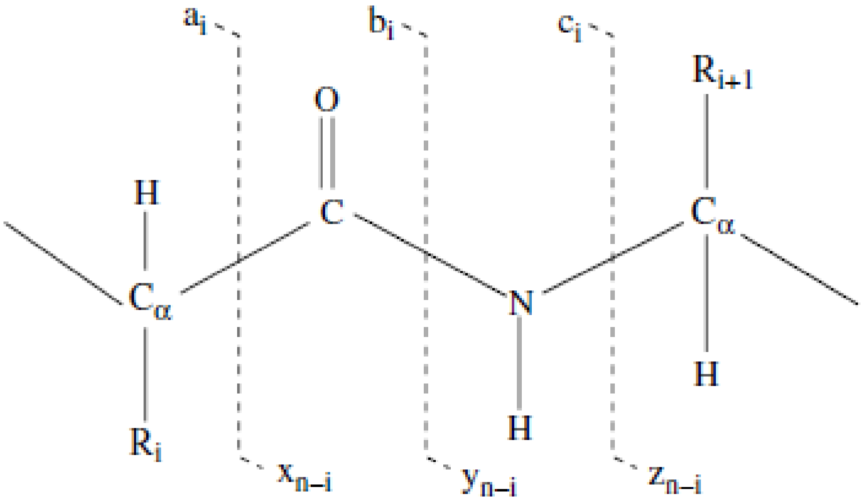
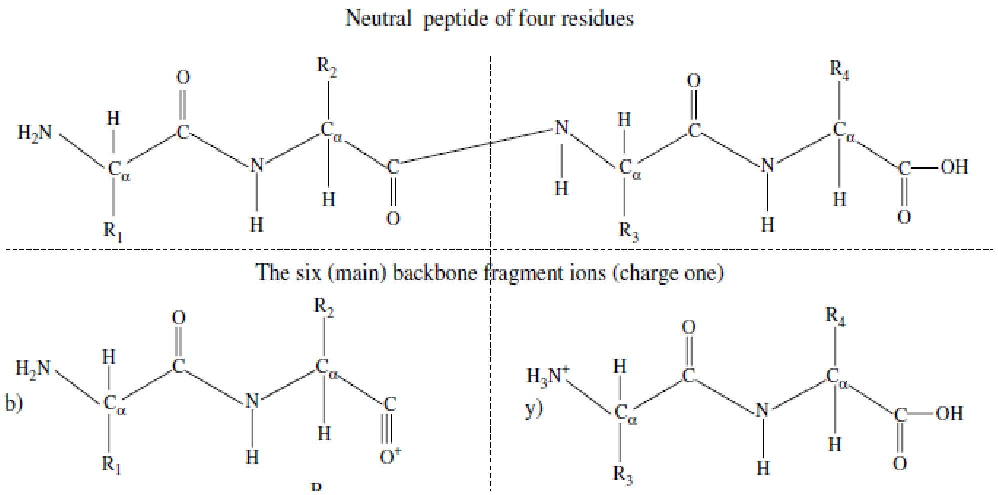


Figure 0.89 Masses after Fragmentation by ECD, CID & ETD

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.

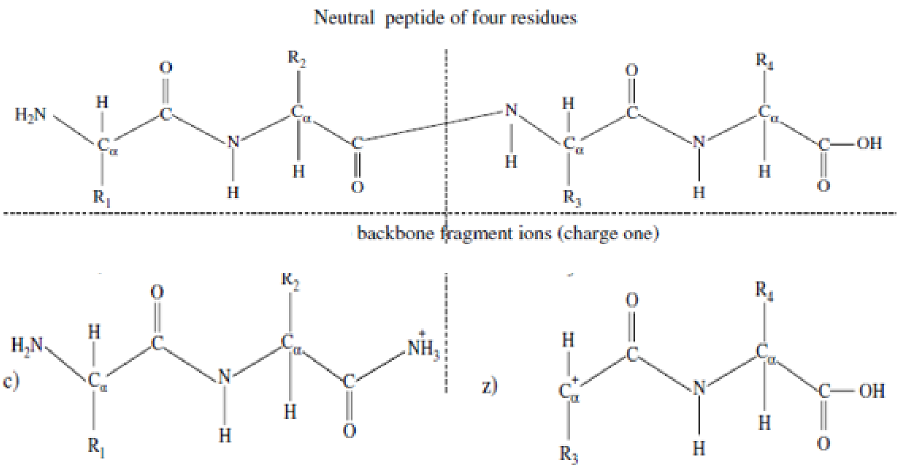
Module 019: Calculating Theoretical Fragment's Mass

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


Neutral peptide of four residues
The six (main) backbone fragment ions (charge one)
b)
y)

$$M_b = R_F + \langle N \rangle \qquad M_y = R_F + \langle C \rangle + 2\mathrm{H}$$

Figure 20 Masses after Fragmentation by CID


Neutral peptide of four residues
backbone fragment ions (charge one)
c)
z)

$$M_c = R_F + \langle N \rangle + \mathrm{N} + 3\mathrm{H} \qquad M_z = R_F + \langle C \rangle - \mathrm{NH}$$

Figure 0.91 Masses after Fragmentation by ECD

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

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

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

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

Figure 22 Masses after Fragmentation by ECD

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.

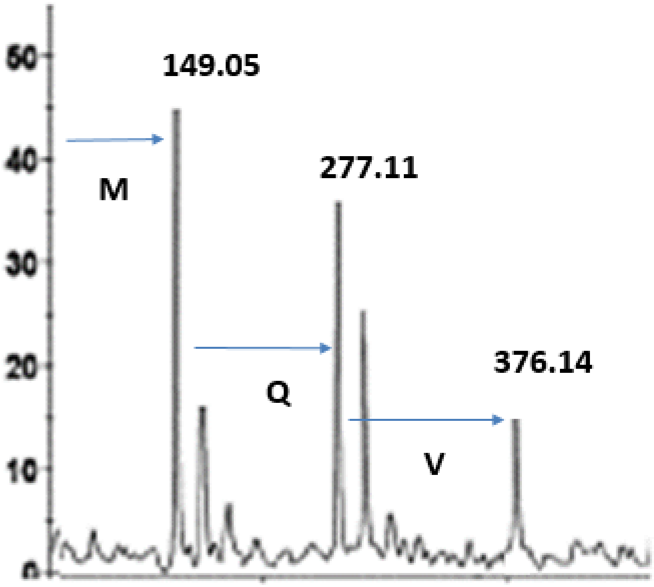


Figure 23 variation sites

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.

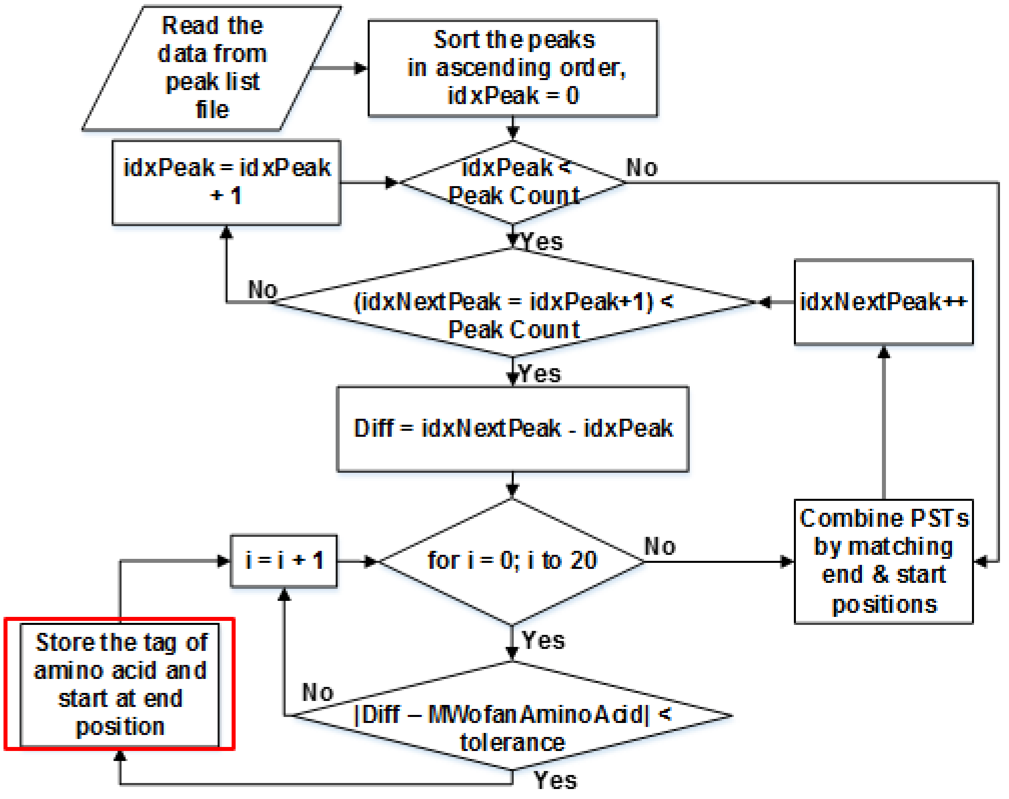

Read the data from peak list file
Sort the peaks in ascending order, idxPeak = 0
idxPeak = idxPeak + 1
idxPeak < Peak Count
No
Yes
No
(idxNextPeak = idxPeak+1) < Peak Count
idxNextPeak++
Yes
Diff = idxNextPeak - idxPeak
i = i + 1
for i = 0; i to 20
No
Combine PSTs by matching end & start positions
Store the tag of amino acid and start at end position
Yes
No
|Diff – MWofanAminoAcid| < tolerance
Yes

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

```
MAFSAEDVLKEYDRRRRMEALLLSLYYPNDRKLLDYKEWSPPRVQVECPKAPVEWNNPPSEKGLIVGHFSGI
KYKGEKAQASEVDVNKMCCWVSKFKDAMRRYQGIQTCKIPGQVLSDLDAKIKAYNLTVEGVEGFVRYSRVT
KQHVAAFLKELRHSKQYENVNLIHYILTDKRVDIQHLEKDLVKDFKALVESAHRMRQGHMQNVKYILYQLLK
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.

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.

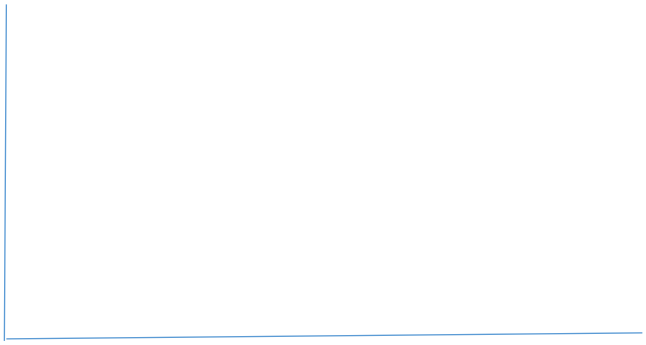


Figure 25 root mean square error

RMSE for a sequence tag ‘i’ of length ‘n’?

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

So, the updated relationship is:

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

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.

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:

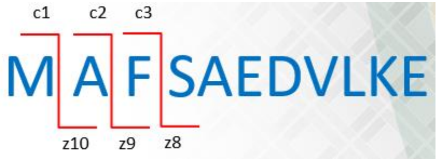


Figure 26 random protein sequence

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

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.

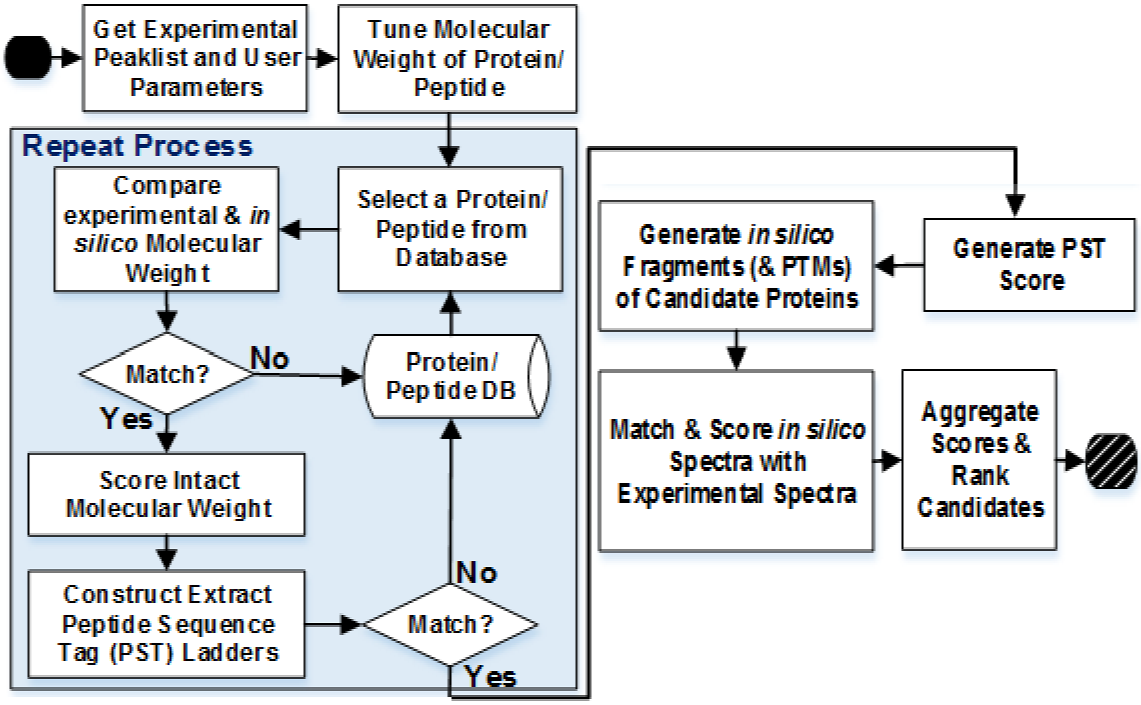

Get Experimental Peaklist and User Parameters
Tune Molecular Weight of Protein/ Peptide
Repeat Process
Compare experimental & *in silico* Molecular Weight
Select a Protein/ Peptide from Database
Match?
No
Yes
Protein/ Peptide DB
Score Intact Molecular Weight
Construct Extract Peptide Sequence Tag (PST) Ladders
Match?
No
Yes
Generate *in silico* Fragments (& PTMs) of Candidate Proteins
Generate PST Score
Match & Score *in silico* Spectra with Experimental Spectra
Aggregate Scores & Rank Candidates

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).

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

- 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.

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**

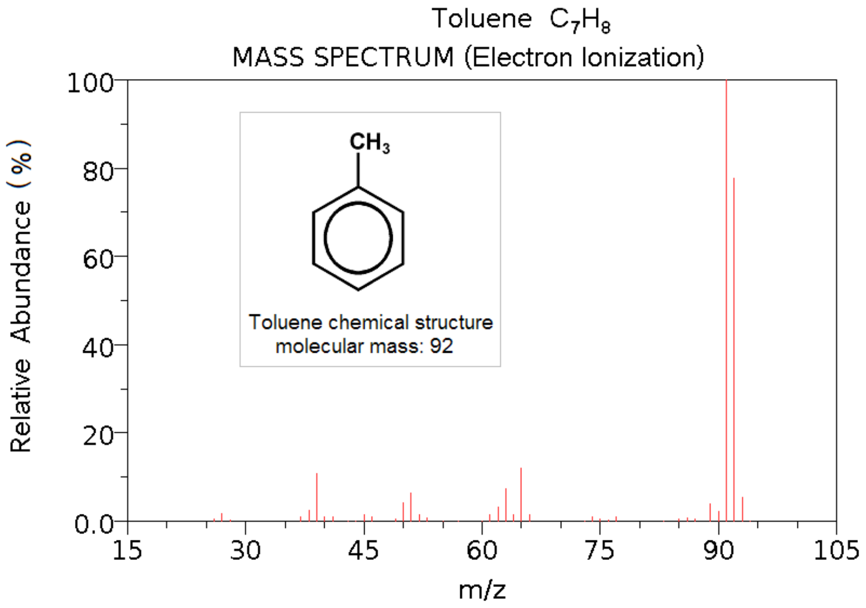
1. The instrument alternates between MS and MS/MS modes.
2. Three most intense peaks are chosen for MS/MS analysis.
3. After the initial MS scan, an MS/MS spectrum from peptide A is obtained by selectively fragmenting this mass only.
4. Next, a spectrum for peptide B is produced, followed by a recording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

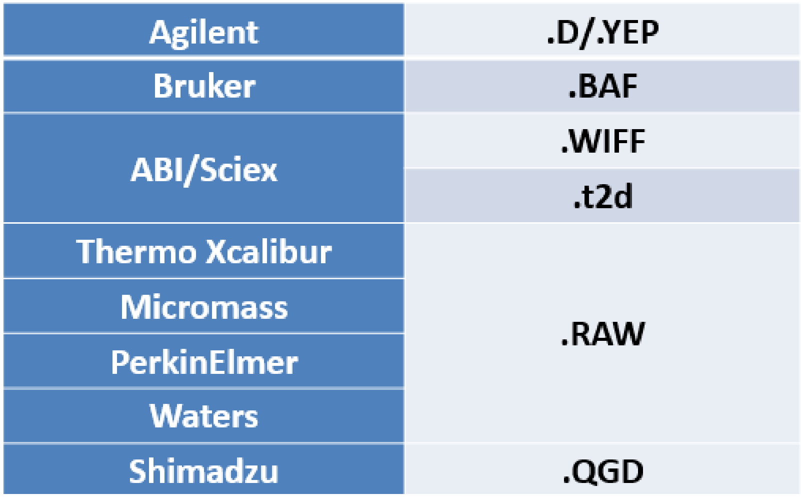
Module 029: PROTEOMICS DATA FILE FORMATS

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.



NIST Chemistry WebBook (http://webbook.nist.gov/chemistry)

Followings are the formats for proteomics data:



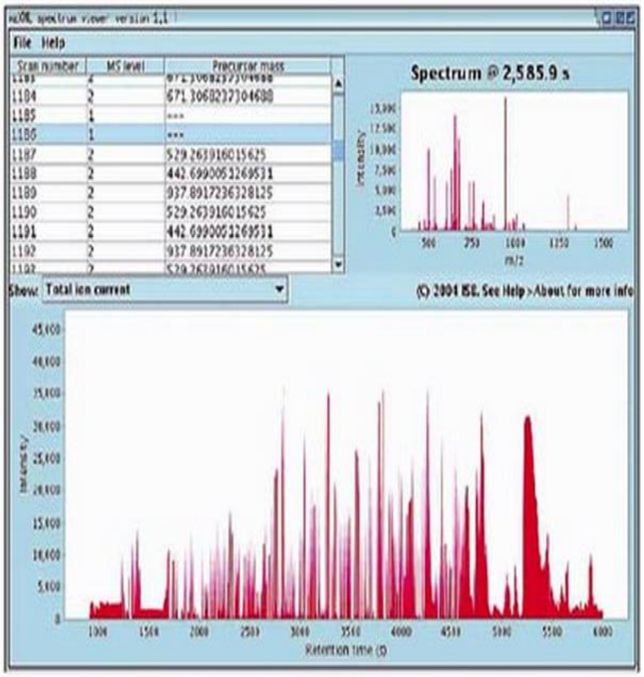
Agilent	.D/.YEP
Bruker	.BAF
ABI/Sciex	.WIFF
	.t2d
Thermo Xcalibur	.RAW
Micromass	
PerkinElmer	
Waters	
Shimadzu	.QGD

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)

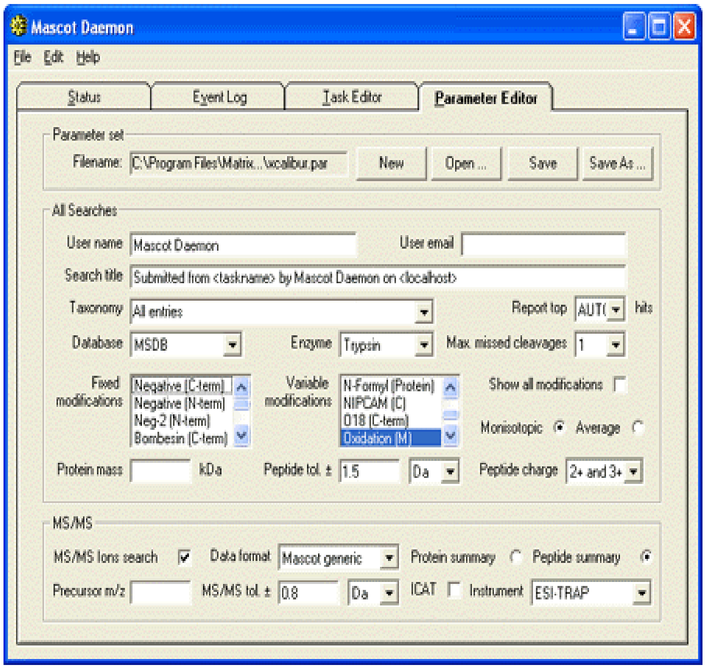


mzXML spectrum viewer version 1.1

Scan number	MS level	Precursor mass
1184	2	671.3068237304688
1185	1	***
1186	1	***
1187	2	529.263916015625
1188	2	442.6990051269531
1189	2	937.8917236328125
1190	2	529.263916015625
1191	2	442.6990051269531
1192	2	937.8917236328125

Show: Total ion current

(C) 2004 ISB. See Help>About for more info



Mascot Daemon

Parameter set

Filename: C:\Program Files\Matrix...\xcalibur.par

All Searches

User name: Mascot Daemon

User email:

Search title: Submitted from <taskname> by Mascot Daemon on <localhost>

Taxonomy: All entries

Report top: AUTO hits

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)

Show all modifications ☐

Monoisotopic ● Average ○

Protein mass: kDa

Peptide tol. ± 1.5 Da

Peptide charge: 2+ and 3+

MS/MS

MS/MS Ions search ☑

Data format: Mascot generic

Protein summary ○ Peptide summary ●

Precursor m/z:

MS/MS tol. ± 0.8 Da

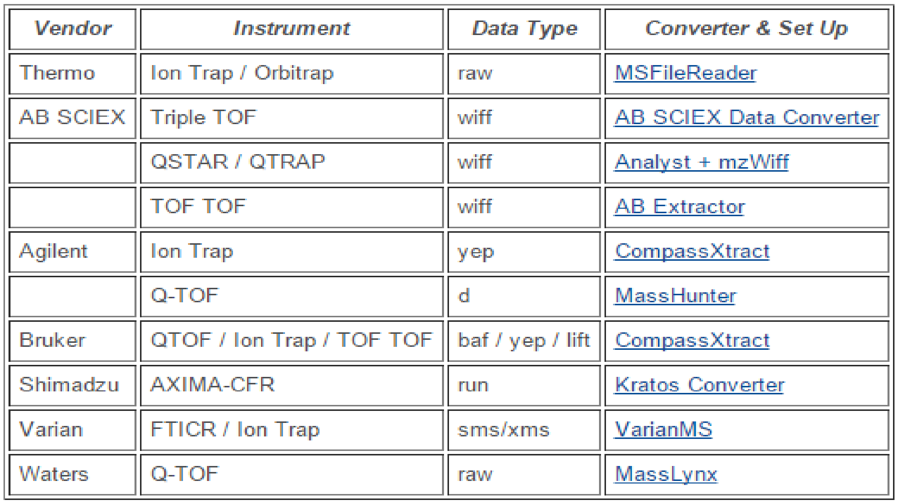
ICAT ☐

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.

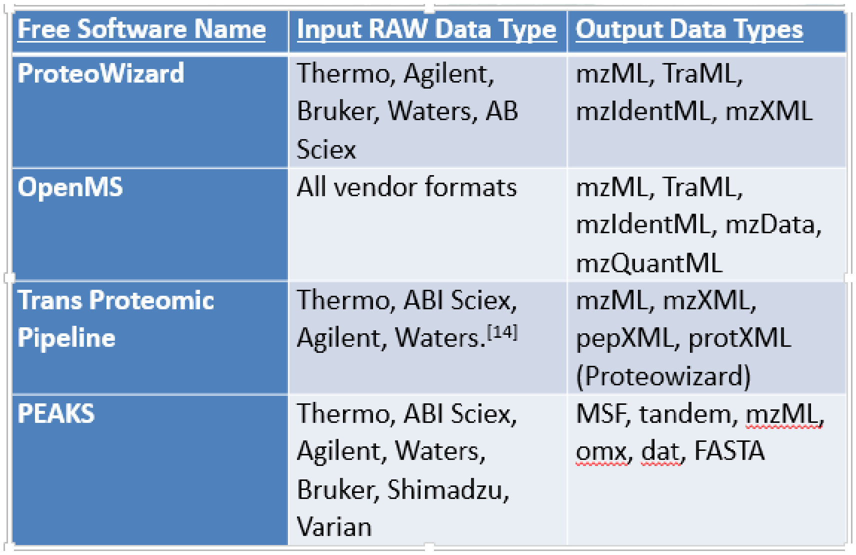
Module 030: RAW FILE FORMAT

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



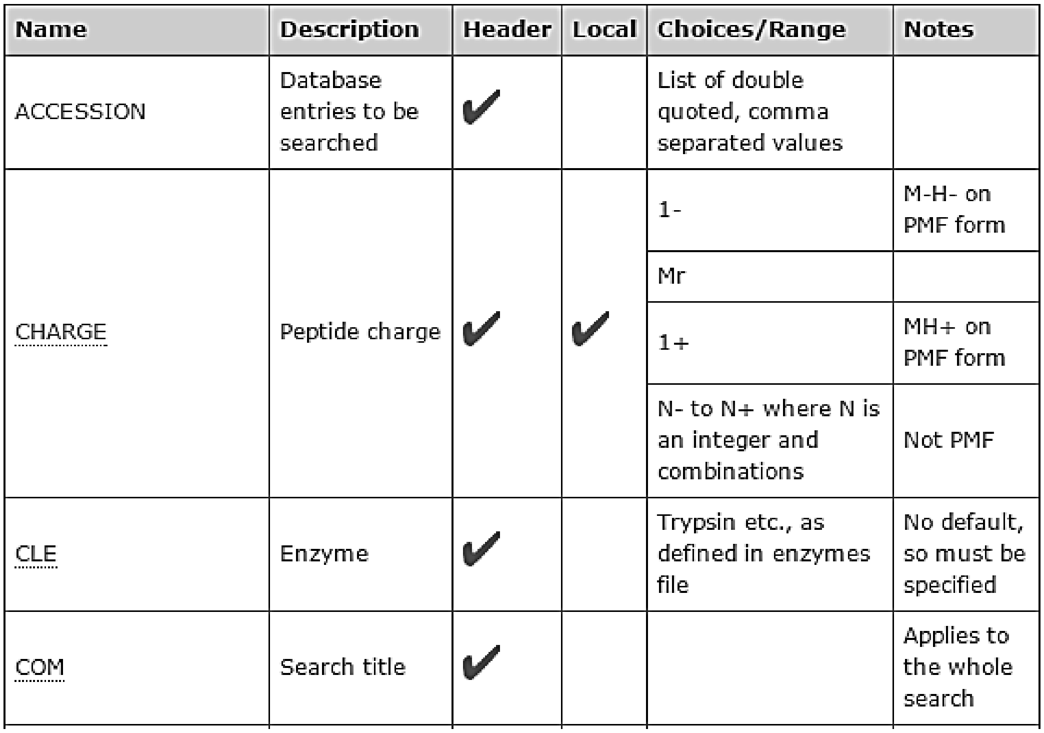
Free Software Name	Input RAW Data Type	Output Data Types
ProteoWizard	Thermo, Agilent, Bruker, Waters, AB Sciex	mzML, TraML, mzIdentML, mzXML
OpenMS	All vendor formats	mzML, TraML, mzIdentML, mzData, mzQuantML
Trans Proteomic Pipeline	Thermo, ABI Sciex, Agilent, Waters.[14]	mzML, mzXML, pepXML, protXML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

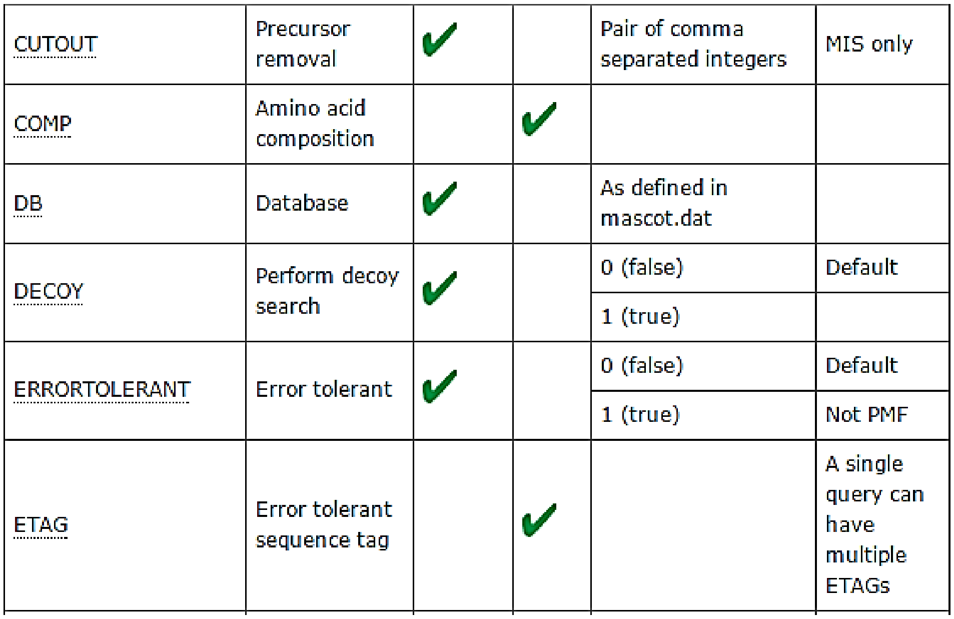
Module 031: MGF FILE FORMAT

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**

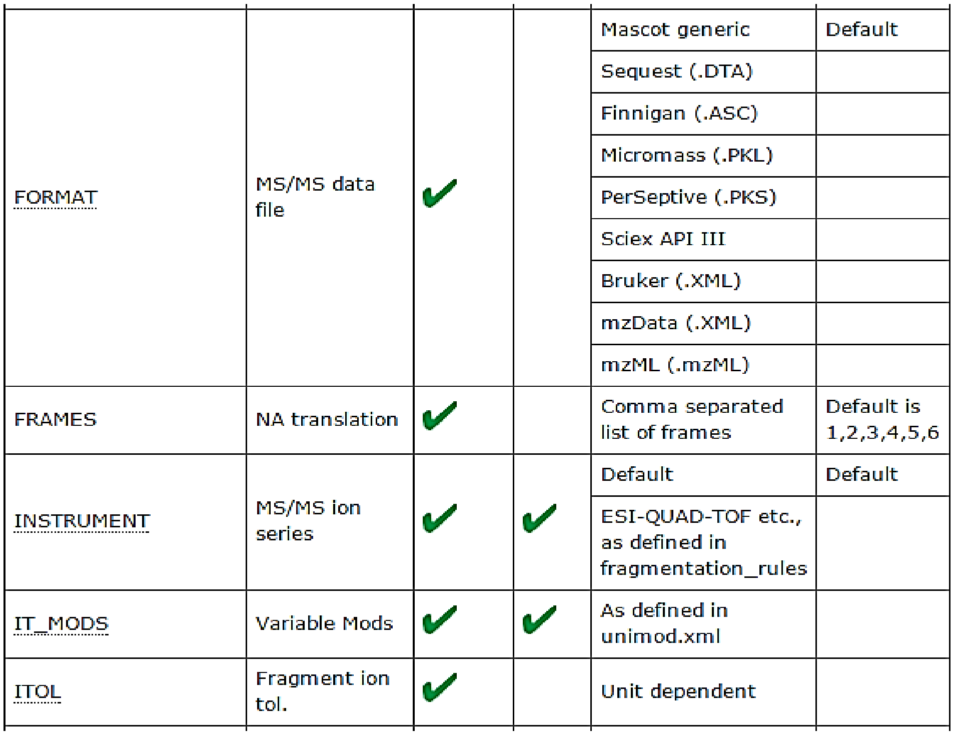


Name	Description	Header	Local	Choices/Range	Notes
ACCESSION	Database entries to be searched	✓		List of double quoted, comma separated values	
CHARGE	Peptide charge	✓	✓	1-	M-H- on PMF form
				Mr	
				1+	MH+ on PMF form
				N- to N+ where N is an integer and combinations	Not PMF
CLE	Enzyme	✓		Trypsin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

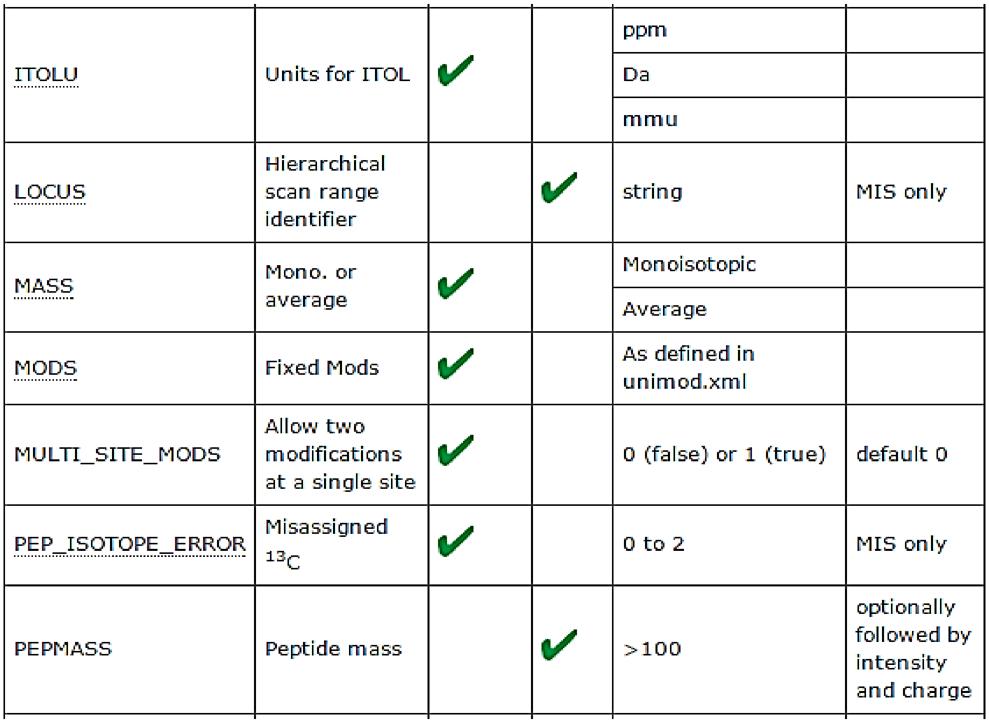
http://www.matrixscience.com/help/data_file_help.html



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



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



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

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

Module 032: OPEN MS DATA FORMAT

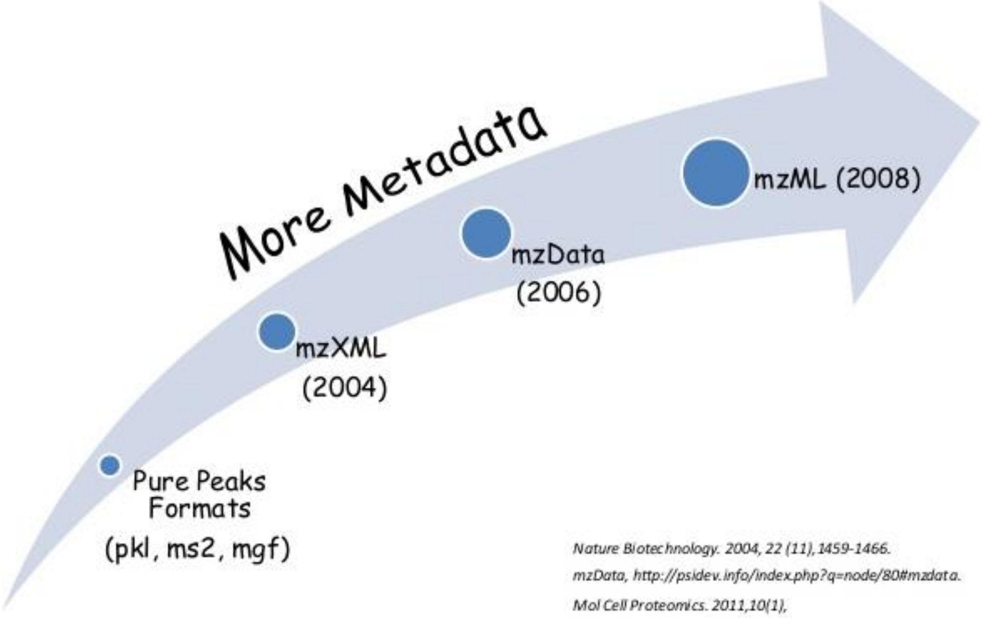
Mass spectrometer outputs data with mass/charge ratios & respective ion intensities. RAW file formats are specific to each instrument and each vendor has its own unique file format. Once an instrument is upgraded, data output from the instrument is also changed. Hence the underlying RAW file format needs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

NEED

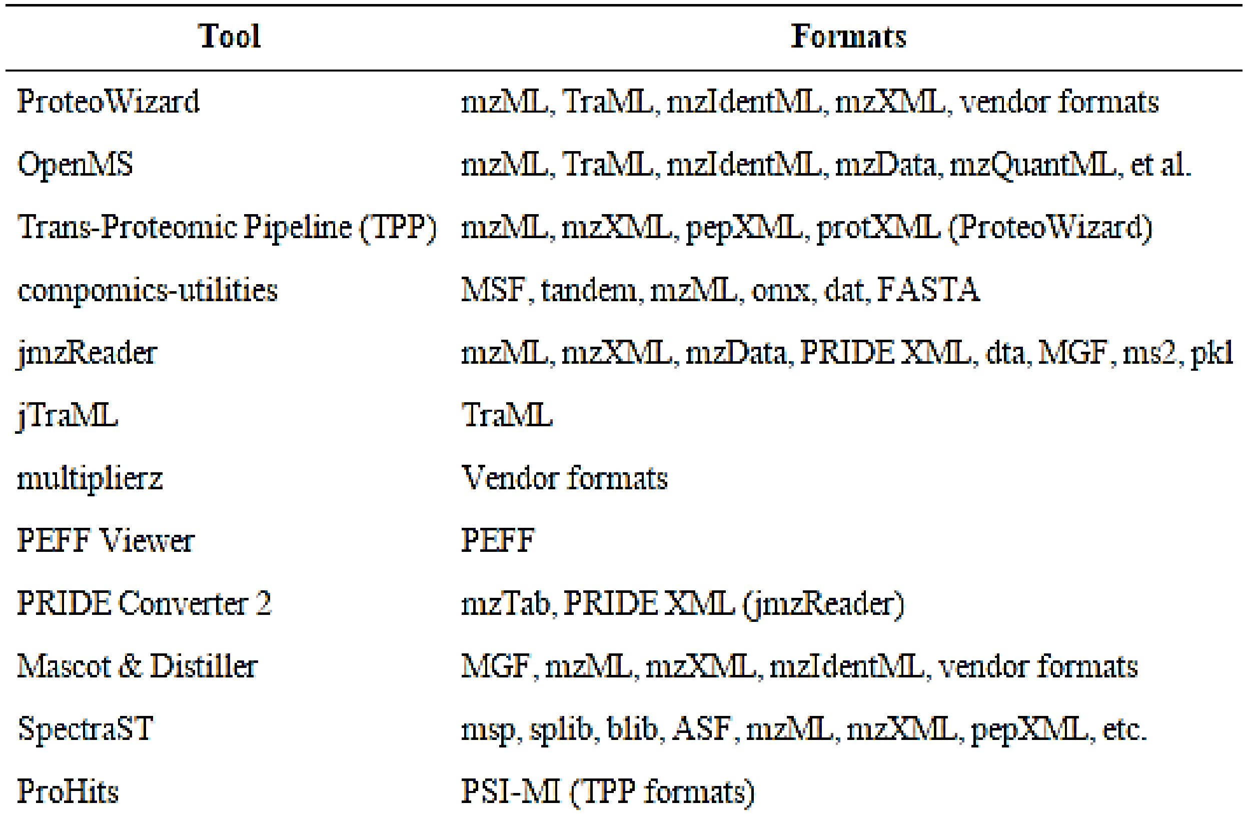
Proprietary RAW formats are binary formats which are difficult to read and parse. If you have the software from the maker of the MS then you can read the RAW data file as well.

SOLUTION

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

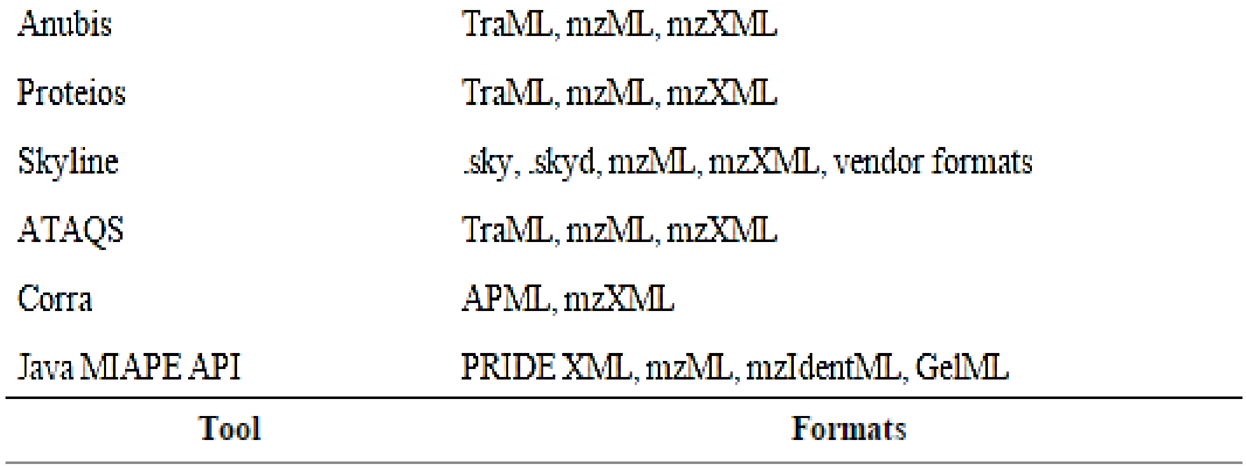


Nature Biotechnology. 2004, 22 (11), 1459-1466.
mzData, http://psidev.info/index.php?q=node/80#mzdata.
Mol Cell Proteomics. 2011,10(1),



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



Tool	Formats
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

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

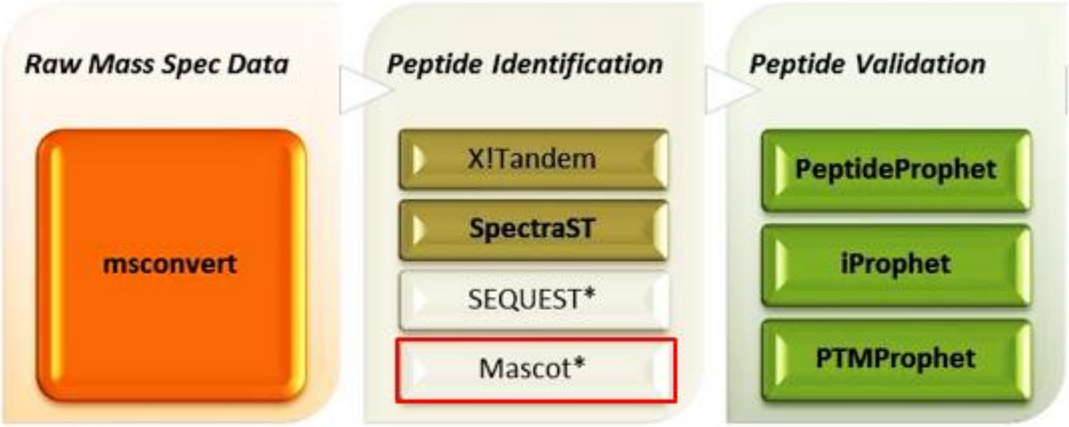
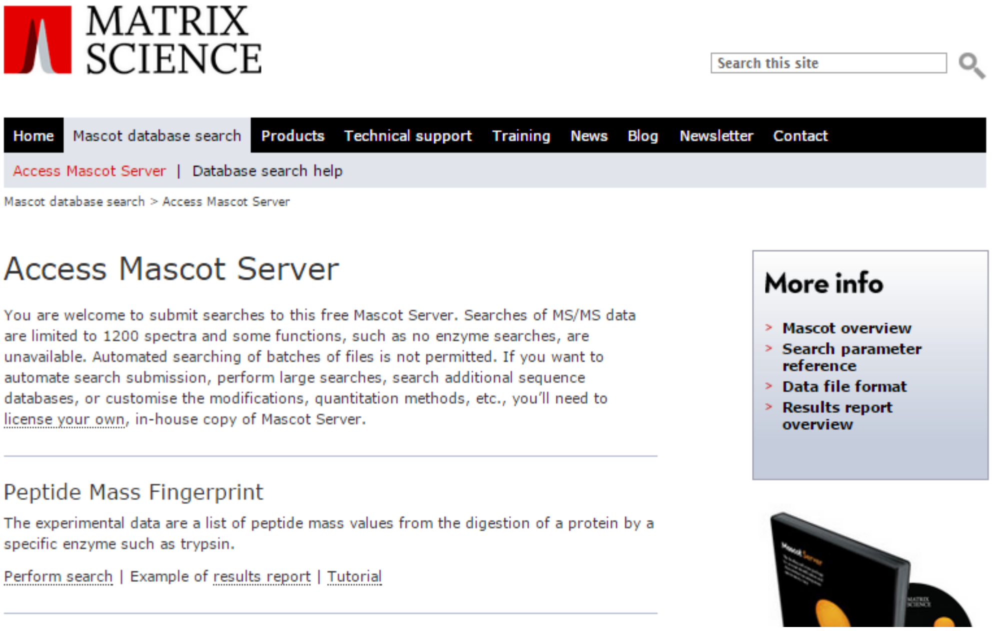


Figure 32 mzML http://tools.proteomecenter.org/software.php



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

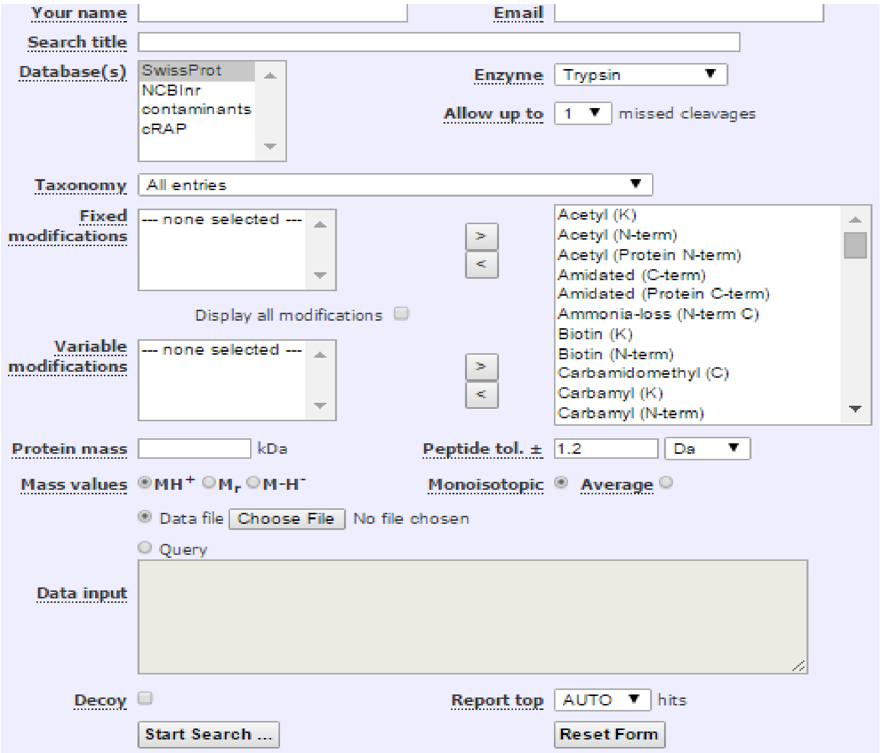
The experimental data are a list of peptide mass values from the digestion of a protein by a specific enzyme such as trypsin.

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. ±: 1.2 Da

Mass values: ● MH+ ○ Mr ○ M-H- 	Monoisotopic ● Average ○

● Data file Choose File No file chosen

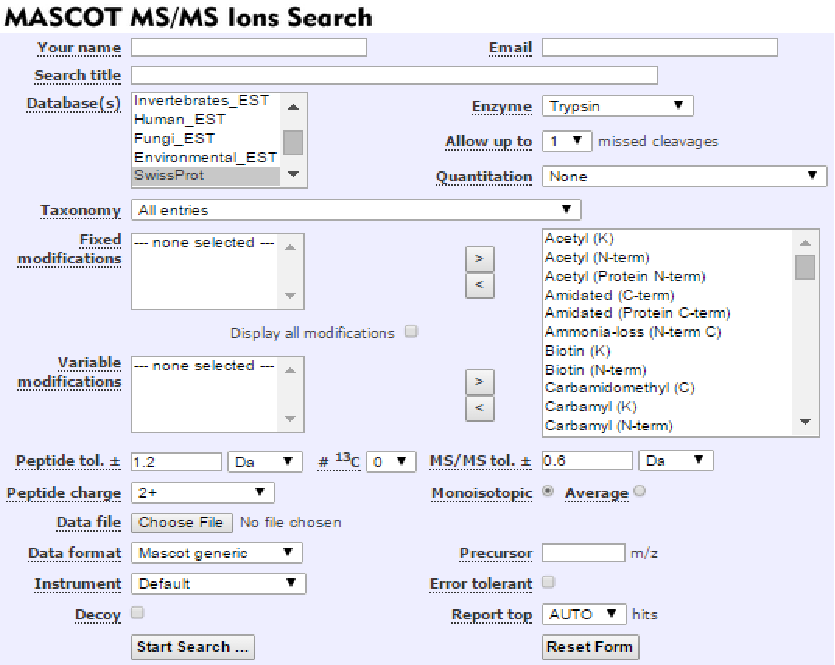
○ Query

Data input:

Decoy □ 	Report top: AUTO hits

Start Search ... 	Reset Form

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. ±: 1.2 Da 	# ^{13}C: 0 	MS/MS tol. ±: 0.6 Da

Peptide charge: 2+ 	Monoisotopic ● Average ○

Data file: Choose File No file chosen

Data format: Mascot generic 	Precursor: m/z

Instrument: Default 	Error tolerant □

Decoy □ 	Report top: AUTO hits

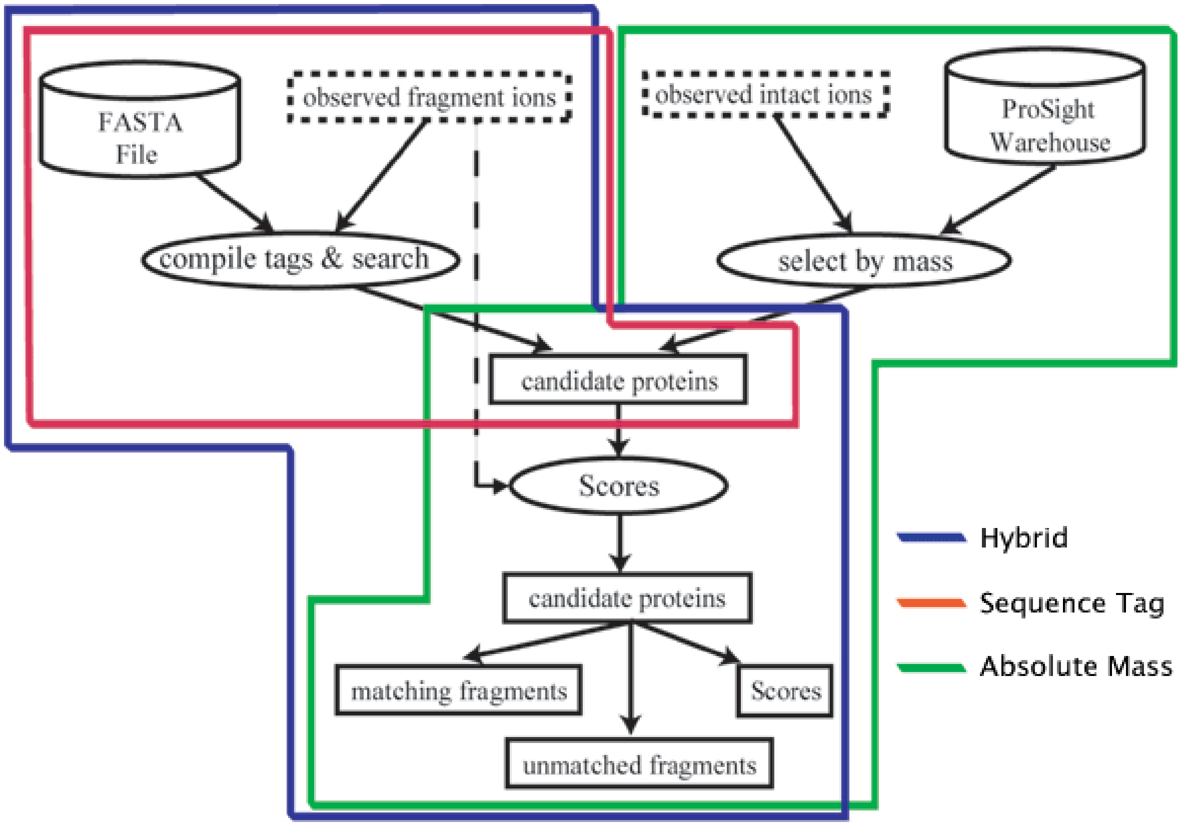
Start Search ... 	Reset Form

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.

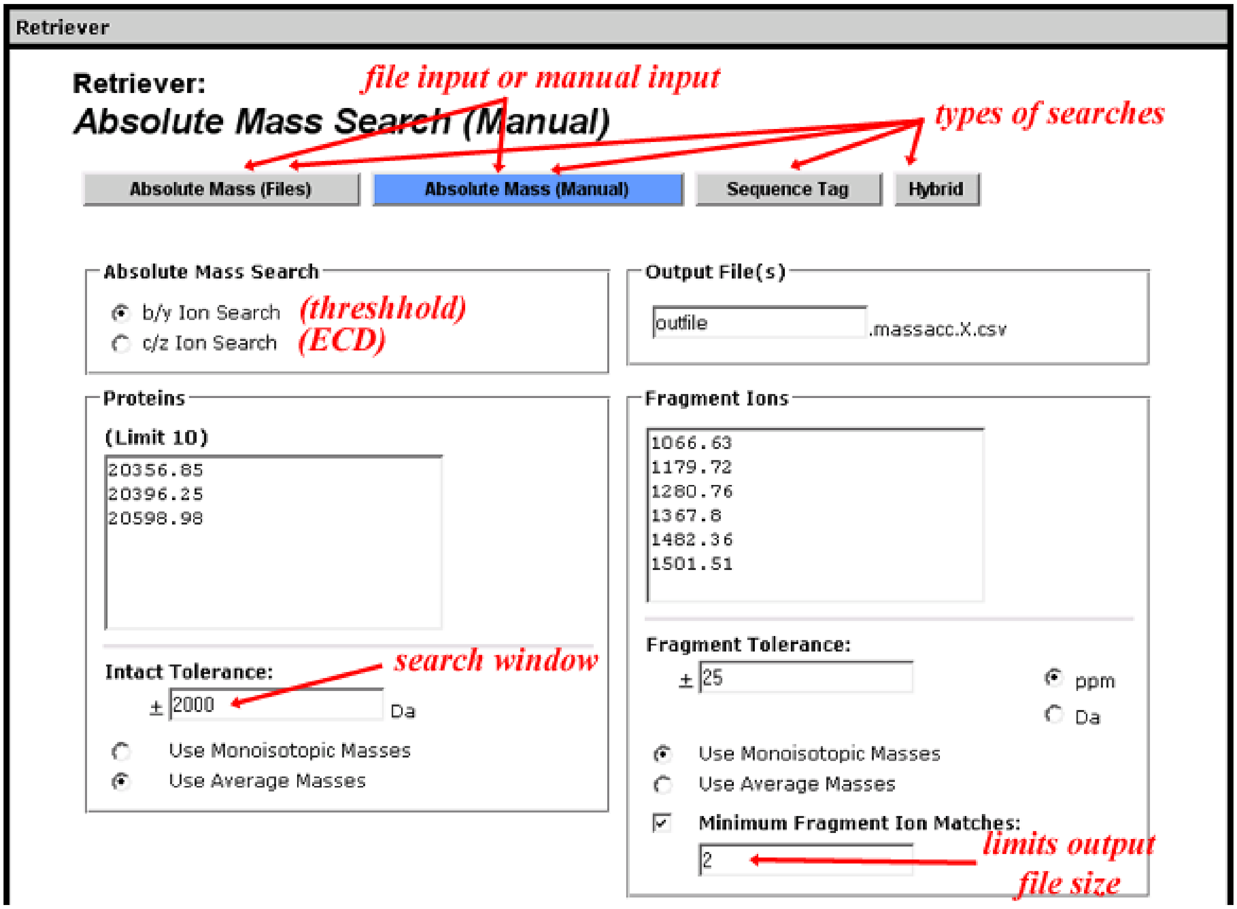
Module 034: Online Proteomics Tools – ProSight PTM

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:
Absolute Mass Search (Manual)

file input or manual input

types of searches

Absolute Mass Search

● b/y Ion Search (threshhold)
○ 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

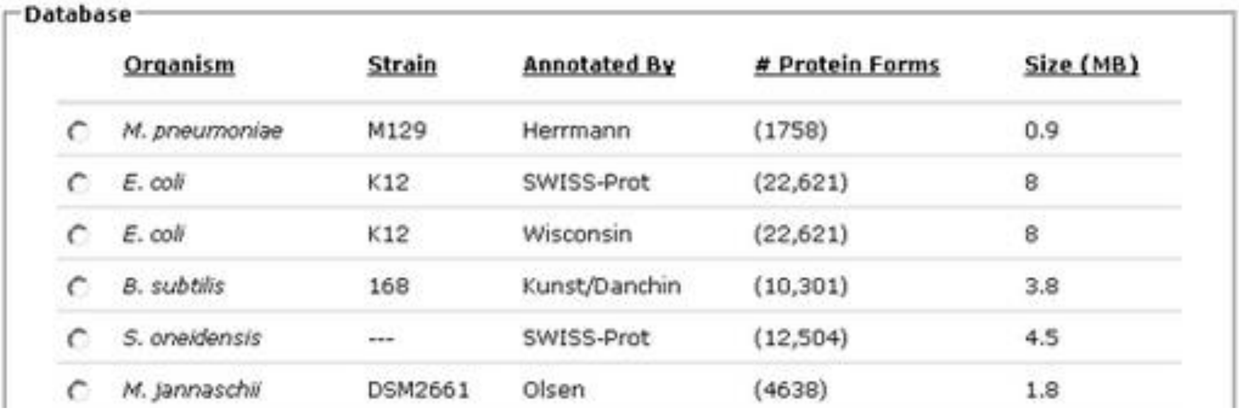
● Use Monoisotopic Masses
○ Use Average Masses

☑ Minimum Fragment Ion Matches:

2

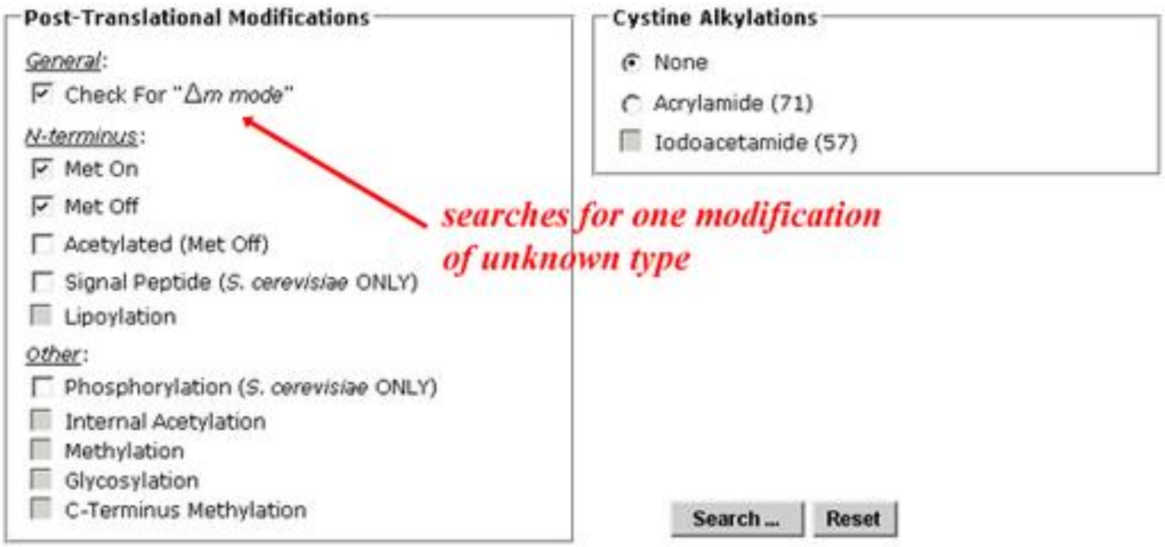
limits output file size

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



Database

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



Post-Translational Modifications

General:

☑ Check For "Δ*m 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

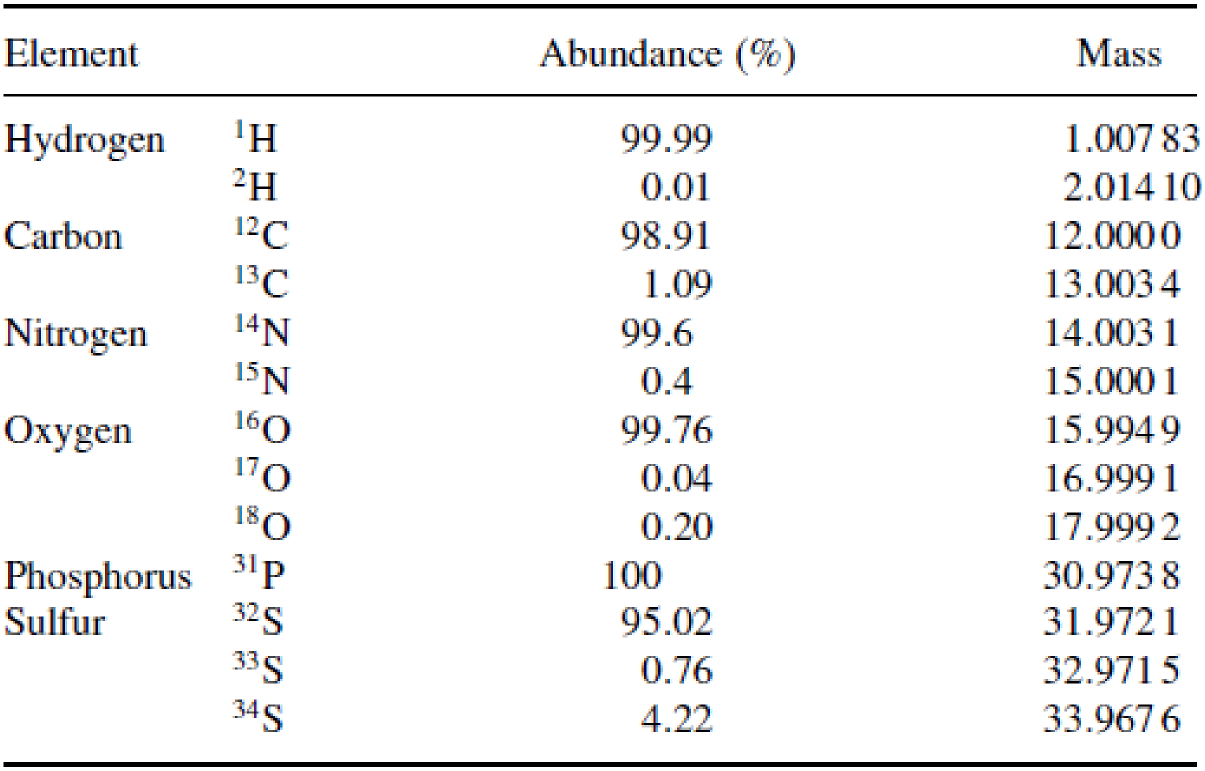
ProSight PTM is the state of the art in top down proteomics search. Using Prosight PTM, post-translational modifications can be accurately identified.

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.

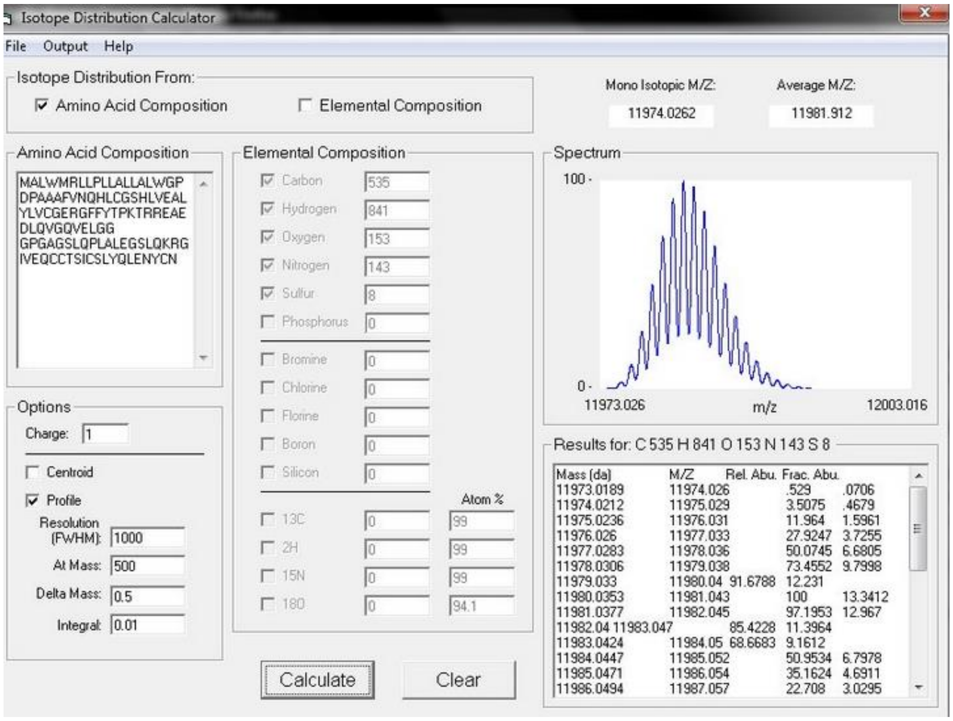


Element		Abundance (%)	Mass
Hydrogen	^{1}H	99.99	1.007 83
	^{2}H	0.01	2.014 10
Carbon	^{12}C	98.91	12.000 0
	^{13}C	1.09	13.003 4
Nitrogen	^{14}N	99.6	14.003 1
	^{15}N	0.4	15.000 1
Oxygen	^{16}O	99.76	15.994 9
	^{17}O	0.04	16.999 1
	^{18}O	0.20	17.999 2
Phosphorus	^{31}P	100	30.973 8
Sulfur	^{32}S	95.02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



Isotope Distribution From:

☑ Amino Acid Composition ☐ Elemental Composition

Mono Isotopic M/Z: 11974.0262

Average M/Z: 11981.912

Amino Acid Composition

```
MALWMRLLPLLALLALWGP
DPAAAFVNQHLCGSHLVEAL
YLVCGERGFFYTPKTRREAE
DLQVGQVELGG
GPGAGSLQPLALEGSLQKRG
IVEQCCTSICSLYQLENYCN
```

Elemental Composition

☑ Carbon	535	
☑ Hydrogen	841	
☑ Oxygen	153	
☑ Nitrogen	143	
☑ Sulfur	8	
☐ Phosphorus	0	
☐ Bromine	0	
☐ Chlorine	0	
☐ Florine	0	
☐ Boron	0	
☐ Silicon	0	
		Atom %
☐ 13C	0	99
☐ 2H	0	99
☐ 15N	0	99
☐ 18O	0	94.1

Options

Charge: 1

☐ Centroid

☑ Profile

Resolution (FWHM): 1000

At Mass: 500

Delta Mass: 0.5

Integral: 0.01

Results for: C 535 H 841 O 153 N 143 S 8

```
Mass (da)       M/Z       Rel. Abu. Frac. Abu.
11973.0189      11974.026           .529     .0706
11974.0212      11975.029           3.5075   .4679
11975.0236      11976.031           11.964   1.5961
11976.026       11977.033           27.9247  3.7255
11977.0283      11978.036           50.0745  6.6805
11978.0306      11979.038           73.4552  9.7998
11979.033       11980.04  91.6788   12.231
11980.0353      11981.043           100      13.3412
11981.0377      11982.045           97.1953  12.967
11982.04 11983.047        85.4228   11.3964
11983.0424      11984.05  68.6683   9.1612
11984.0447      11985.052           50.9534  6.7978
11985.0471      11986.054           35.1624  4.6911
11986.0494      11987.057           22.708   3.0295
```

Figure 35 Detecting Monoisotopic Peaks

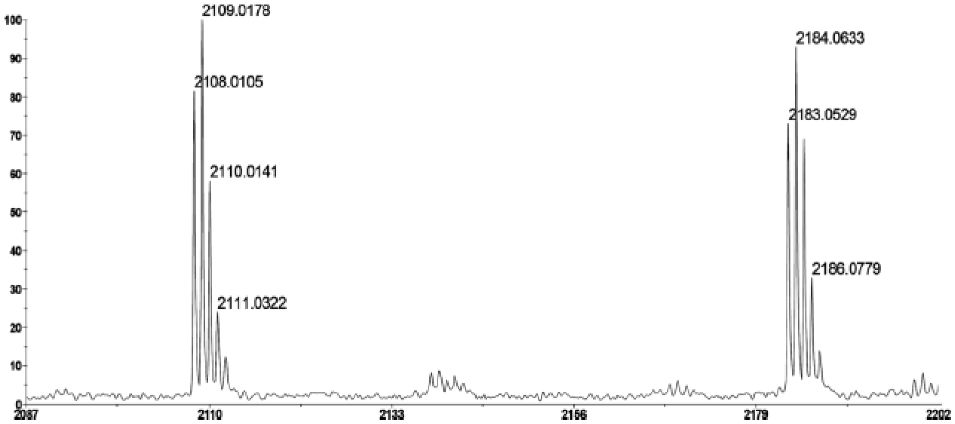


Figure 36 Detecting Monoisotopic Peaks

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.

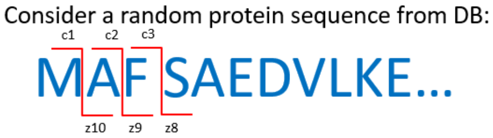
To ensure that all fragments of precursor molecule are also charged, we can use Electrospray ionization (ESI).ESI induces multiple charges on the intact molecule

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:

MAF SAEDVLKE…

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…

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

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)

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

For comprehensive scoring

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

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.

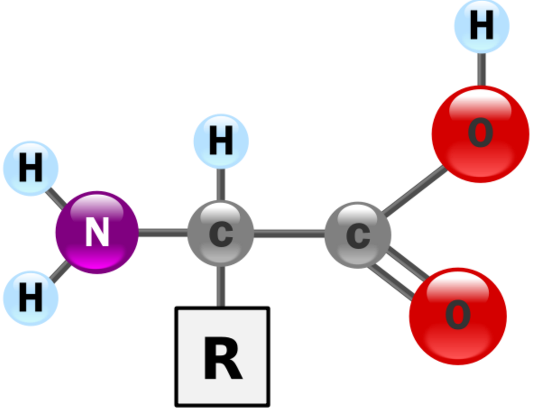


Figure 0.1 chemical structure of amino acid

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

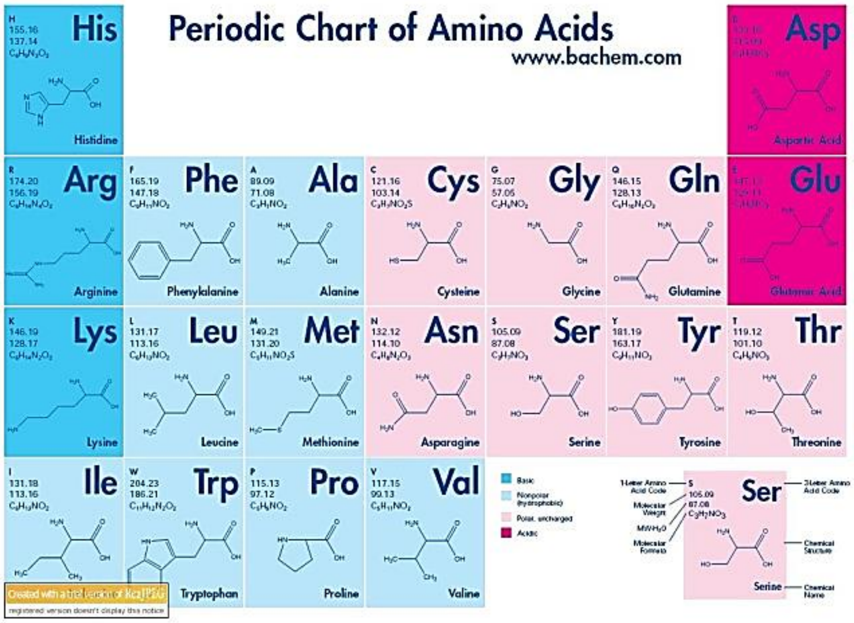

Periodic Chart of Amino Acids
www.bachem.com
H 155.16 137.14 $C_6H_9N_3O_2$ His Histidine
Asp Aspartic Acid
R 174.20 156.19 $C_6H_{14}N_4O_2$ Arg Arginine
F 165.19 147.18 $C_9H_{11}NO_2$ Phe Phenylalanine
A 89.09 71.08 $C_3H_7NO_2$ Ala Alanine
C 121.16 103.14 $C_3H_7NO_2S$ Cys Cysteine
G 75.07 57.05 $C_2H_5NO_2$ Gly Glycine
Q 146.15 128.13 $C_5H_{10}N_2O_3$ Gln Glutamine
Glu Glutamic Acid
K 146.19 128.17 $C_6H_{14}N_2O_2$ Lys Lysine
L 131.17 113.16 $C_6H_{13}NO_2$ Leu Leucine
M 149.21 131.20 $C_5H_{11}NO_2S$ Met Methionine
N 132.12 114.10 $C_4H_8N_2O_3$ Asn Asparagine
S 105.09 87.08 $C_3H_7NO_3$ Ser Serine
Y 181.19 163.17 $C_9H_{11}NO_3$ Tyr Tyrosine
T 119.12 101.10 $C_4H_9NO_3$ Thr Threonine
I 131.18 113.16 $C_6H_{13}NO_2$ Ile
W 204.23 186.21 $C_{11}H_{12}N_2O_2$ Trp Tryptophan
P 115.13 97.12 $C_5H_9NO_2$ Pro Proline
V 117.15 99.13 $C_5H_{11}NO_2$ Val Valine
Basic
Nonpolar (hydrophobic)
Polar, uncharged
Acidic
1-Letter Amino Acid Code
Molecular Weight
MW-H₂O
Molecular Formula
S 105.09 87.08 $C_3H_7NO_3$ Ser Serine
3-Letter Amino Acid Code
Chemical Structure
Chemical Name
Created with a trial version of KeyJPEG
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.

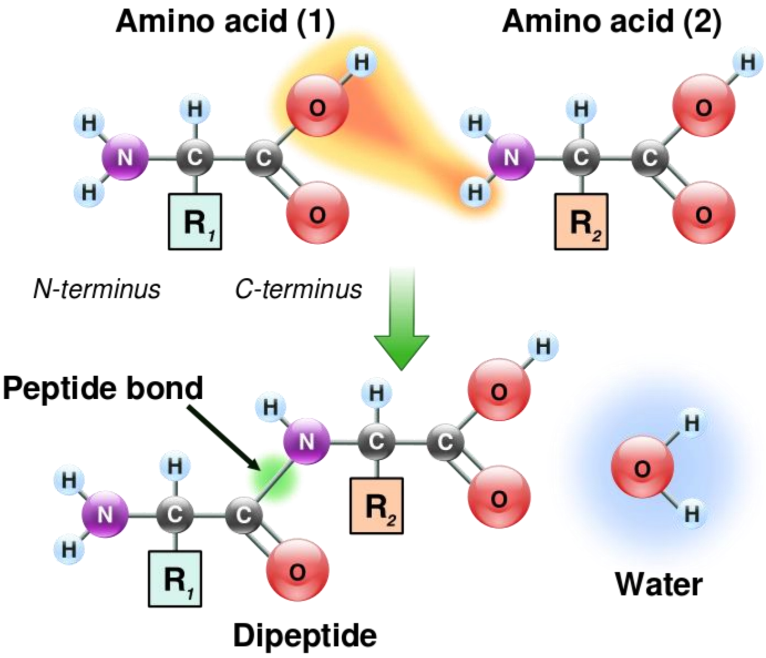


Figure 0.3 polymerization of amino acids

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).

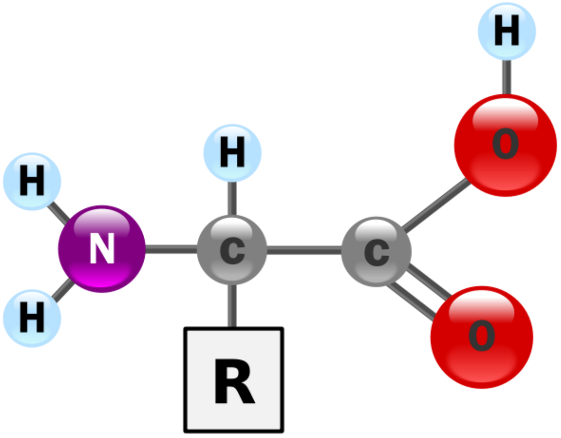


Figure 4 R group in amino acid

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.

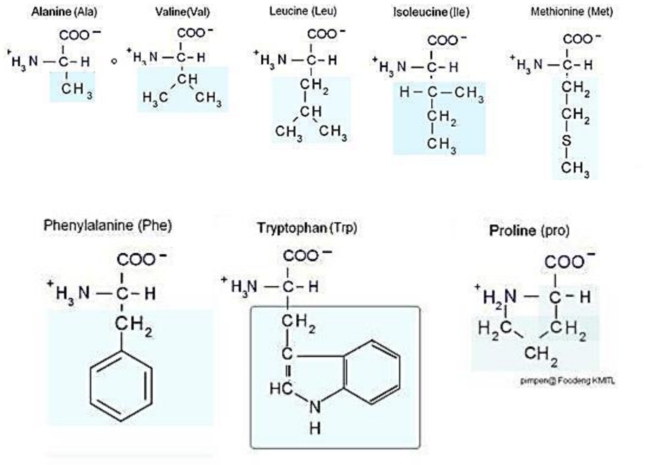


Figure 5 hydrophilic group

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.

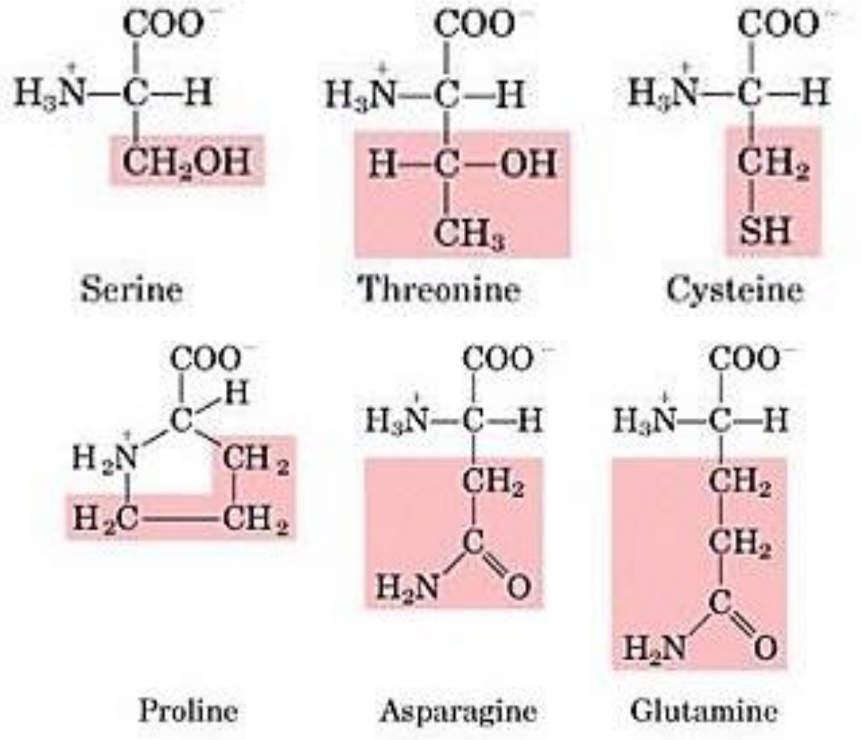


Figure 6 Polar amino acids

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.

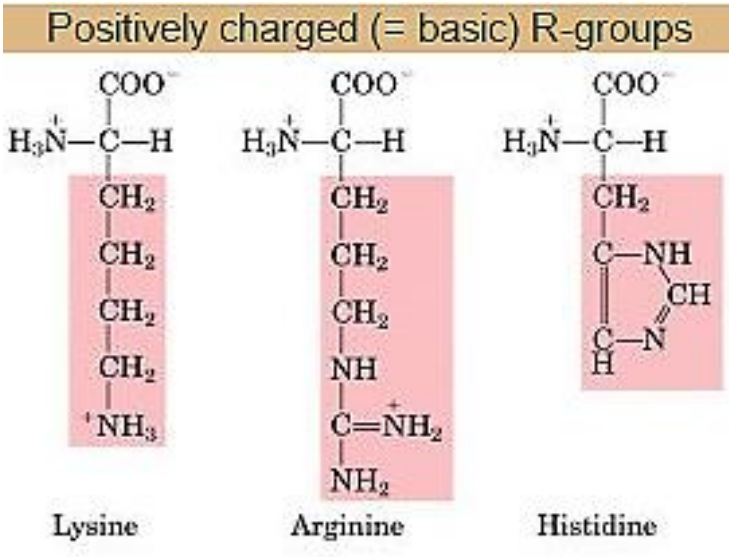


Figure 7 positively charged amino acids

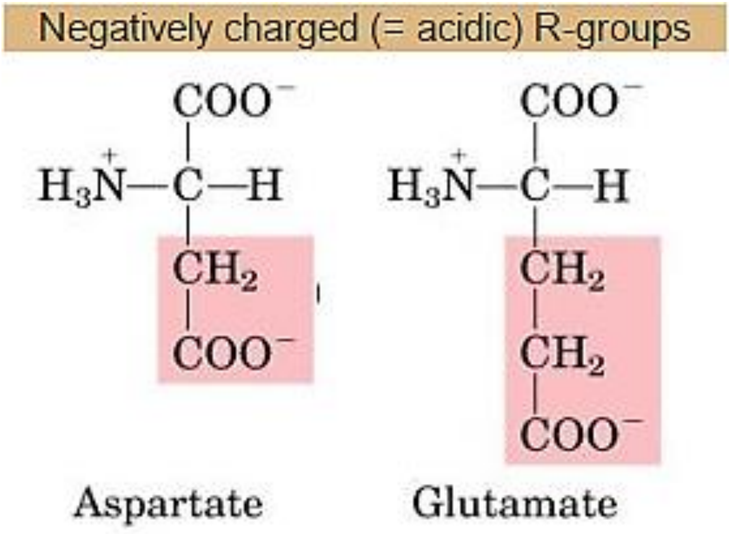
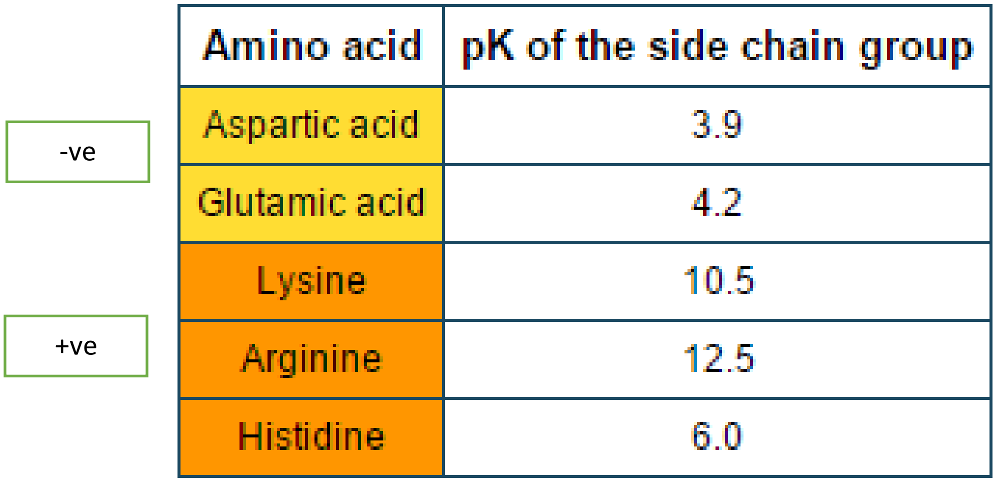


Figure 8 negative charge amino acid

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.

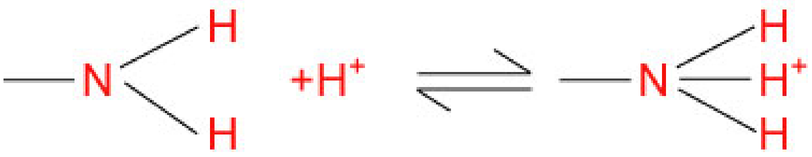
Module 004: PROPERTIES OF AMINO ACIDS-IV

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.

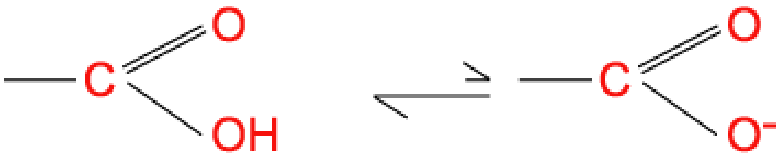


	Amino acid	pK of the side chain group
-ve	Aspartic acid	3.9
	Glutamic acid	4.2
+ve	Lysine	10.5
	Arginine	12.5
	Histidine	6.0

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



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



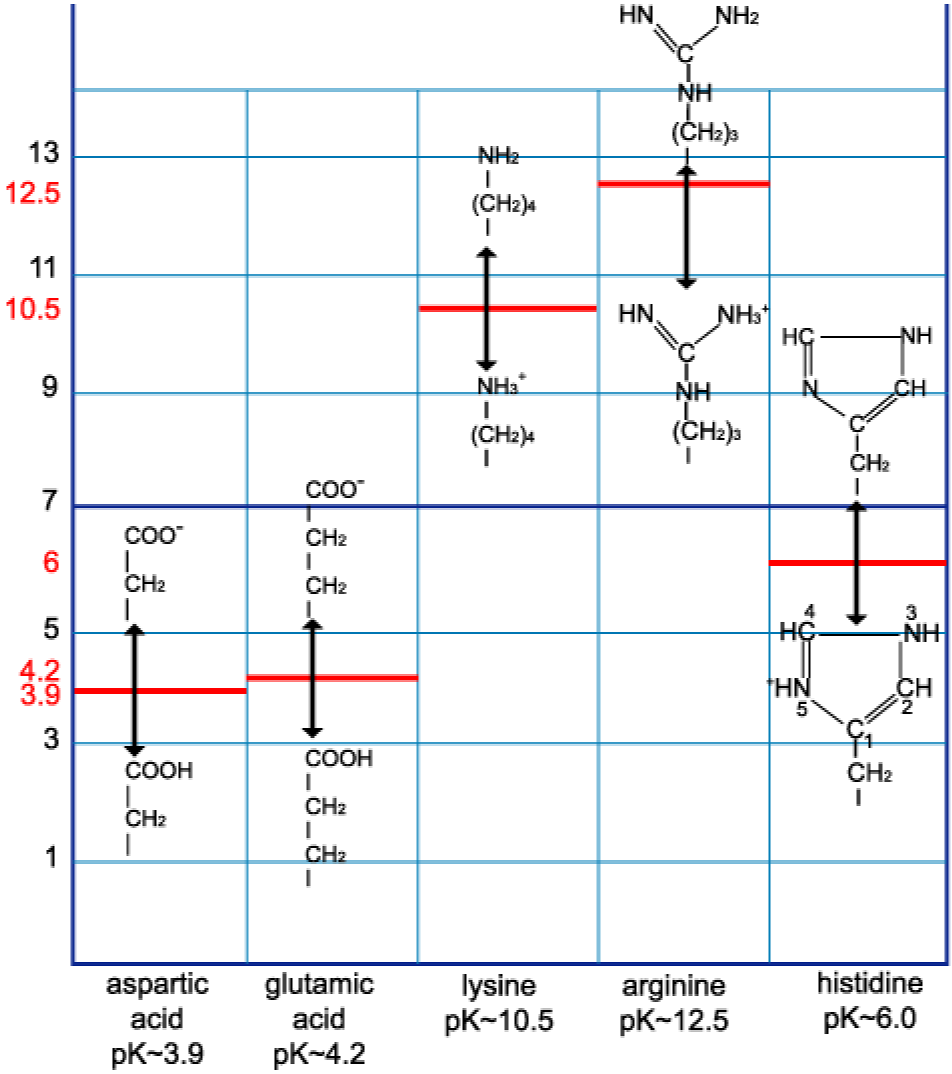


Figure 9 properties of amino acids according to pK and PH.

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.

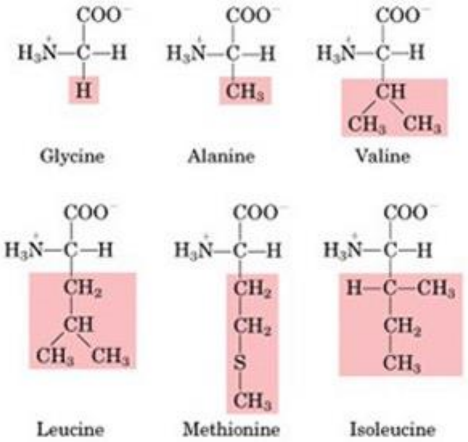


Figure 10 Aliphatic Amino Acids (Non polar C and H chains)

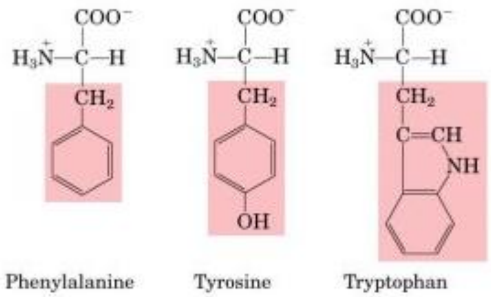


Figure 11 Aromatic R groups

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
H
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.

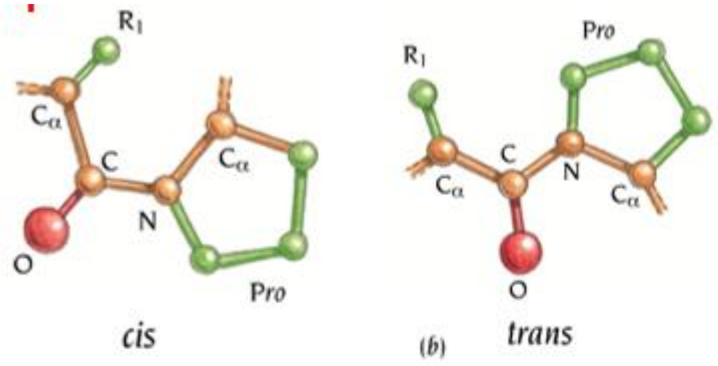


Figure 12 cis and Tran’s form of proline

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.

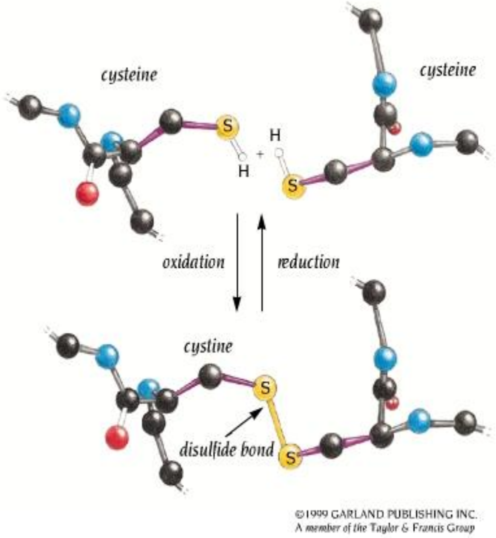

cysteine
cysteine
S
H
+
H
S
oxidation
reduction
cystine
S
disulfide bond
S
©1999 GARLAND PUBLISHING INC.
A member of the Taylor & Francis Group

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.

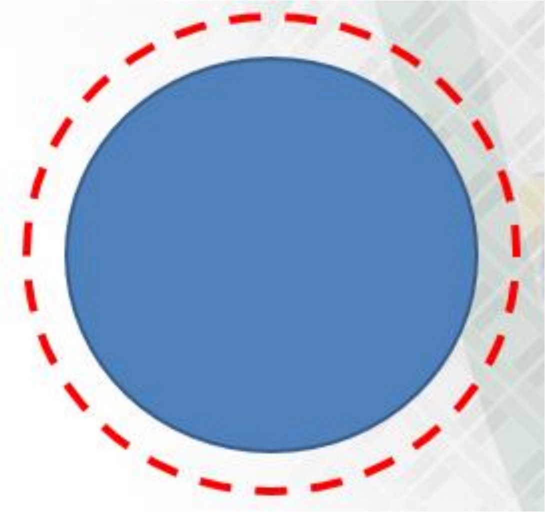
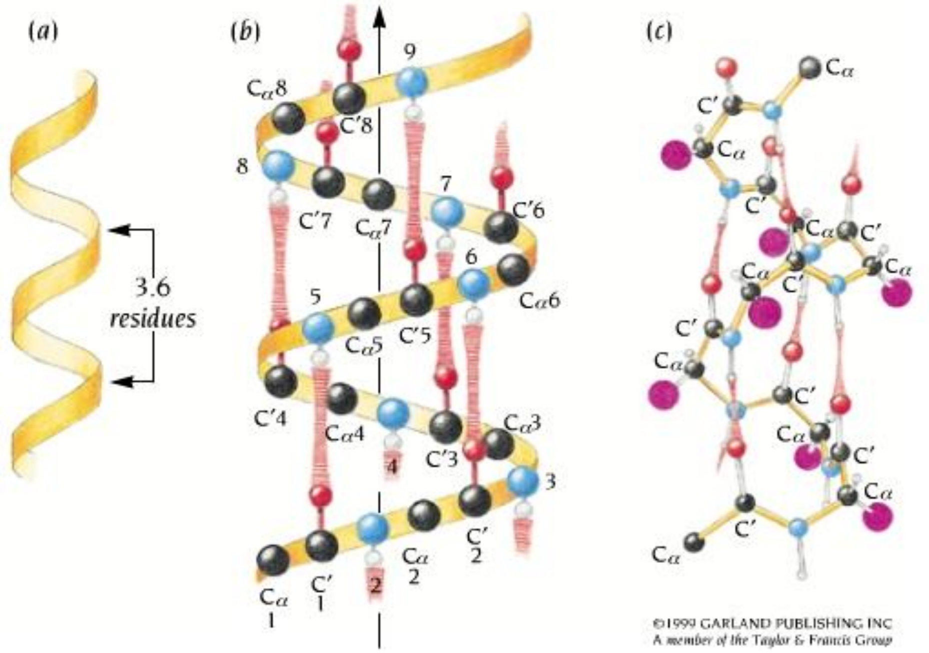


Figure 0.44 Organization of core and surface in a protein

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.

Module 09: STRUCTURAL TRAIT OF AMINO ACID-IV

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.



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

*Figure 0.55 Alpha Helix C = black **O = red N = blue***

Alpha Helix is an example of amino acid folding. Stabilized by H-bonds between every ~ 4th residues in backbone. Reactive amino acids are exposed for external interactions.

Module 010: INTRODUCTION TO PROTEIN FOLDING

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.

But how does a protein actually fold? The answer is still unknown. Scientists have spent decades in trying to find a definite answer to this question, but to no avail. After polymerization of amino acids, linear chains are formed. When these chains of amino acids are put in water, the proteins fold spontaneously.

The folded protein molecule should have the lowest possible energy. Anfinsen's dogma (also known as the thermodynamic hypothesis) is a postulate in molecular biology that, at least for small globular proteins, the native structure is determined only by the protein's amino acid sequence. Unique, stable and kinetically accessible minimum free energy

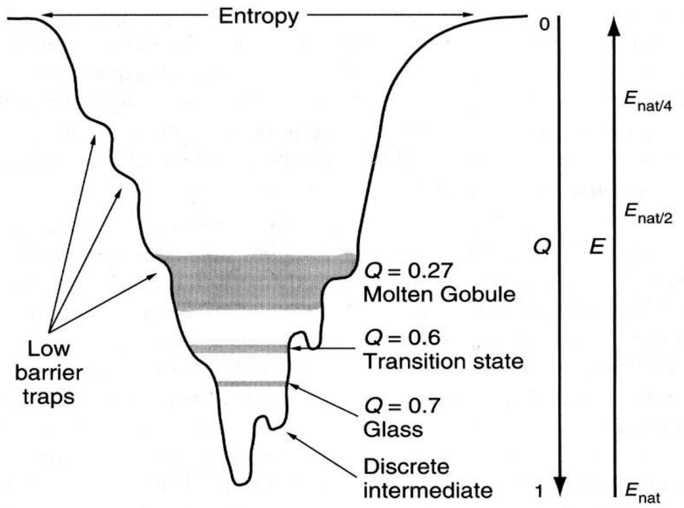


Figure 19.16 Cross section through a folding funnel. *E* corresponds to free energy. [Courtesy of P. G. Wolynes]

Figure 0.66 Overall Energy (stability) of the Protein

Proteins fold spontaneously in water. Proteins fold to achieve thermodynamic stability. Proteins fold to organize themselves for performing functions in cells.

Module 011: IMPORTANCE OF PROTEIN FOLDING

Proteins are like functional machines in cell, therefore understanding the folding behavior of proteins can helps us in designing the suitable drug. If a protein is misfolded, then it can lead to a lack of function in the protein. To study anomalies in structures and to discover newer structural forms, computational algorithms are used.

We can study the folding behavior of protein computationally First, we collect clues & evidences from experimentally reported structures. We utilize these observations to analyze unknown structures. The manner in which a newly synthesized chain of amino acids transforms itself into a perfectly folded protein depends both on the intrinsic properties of the amino-acid sequence (Dobson 2003)

Dobson, C. M. (2003). "Protein folding and misfolding." Nature **426**(6968): 884-890.

Module 012: COMPUTING PROTEIN FOLDING POSSIBILITIES

Computing the protein folding can help us study misfolding, interaction between drugs and proteins etc. However, first, it is important to know the number of the protein folding possibilities.

Let’s assume that each amino acid can fold into three different conformations. They are Alpha Helices, Beta Sheets and Loops. We know that proteins comprise of 100s of amino acids

If each amino acid can take 3 different conformations, and its parent protein has 100 amino acids, then 100^3 = 5 x 10^{47} will be the combination. If it take 1/10th of a Nano-second (10^{-10}), then to compute all the folding possibilities will take 1.6 x 10^{30} years.

In fact, it take a protein less than a second to fold. It’s the Amazing speed of folding.

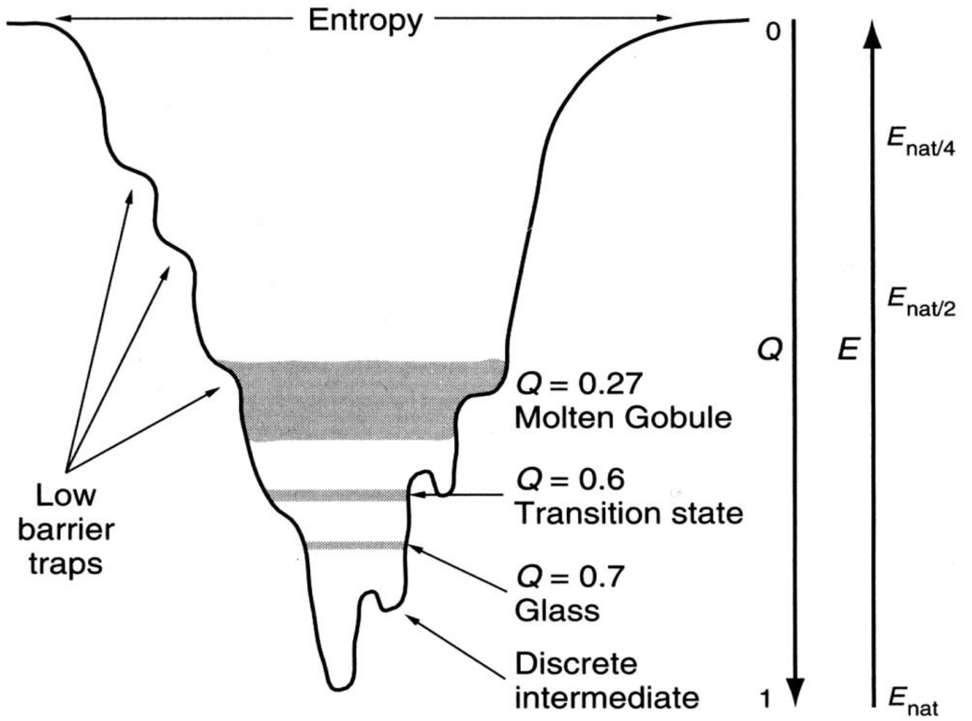


Figure 19.16 Cross section through a folding funnel. *E* corresponds to free energy. [Courtesy of P. G. Wolynes]

Figure 0.77 Overall Energy (stability) of the Protein

This is called “Levinthal’s Paradox”. We will try to understand this folding process using experimental datasets and algorithms. Molecular simulations are also helpful for it.

Module 013: PROCESSING OF PROTEIN FOLDING

Levinthal's Paradox- enormous time required to compute all folding possibilities. It's impossible to consider all the possibilities computationally. So, we are trying to understand the folding process.

The forces involved in protein folding include:

- Electrostatic interactions
- van der Waals interactions
- Hydrogen bonds
- Hydrophobic interactions

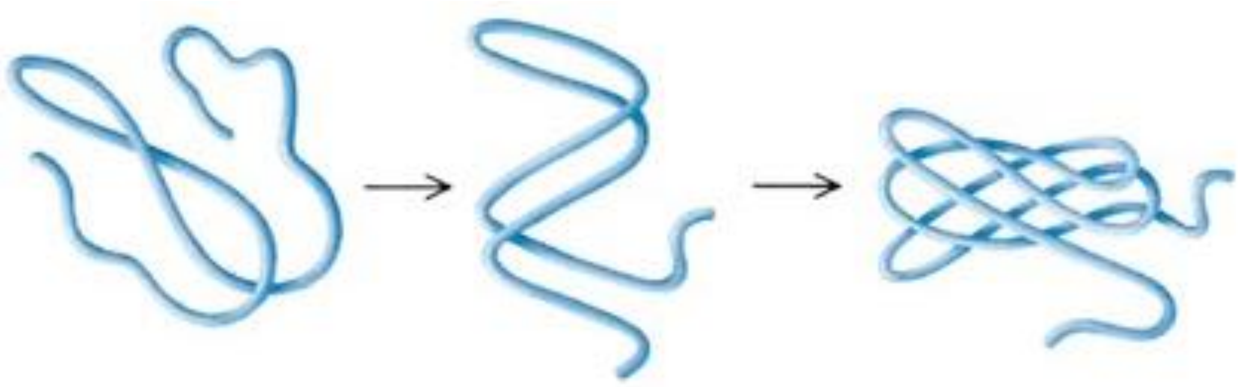


Figure 0.88 Protein folding

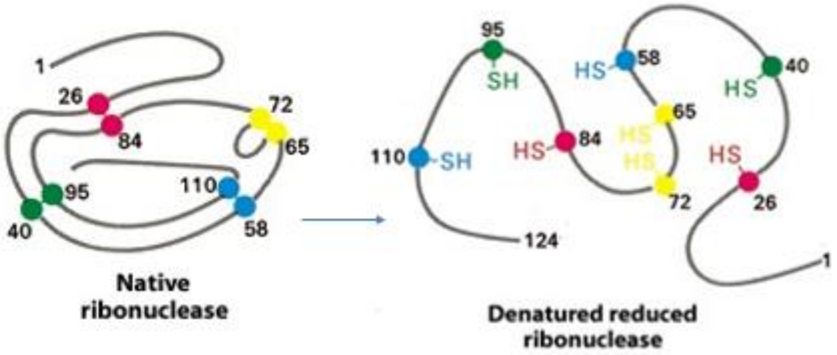


Figure 19 Anfinsen's Experiment

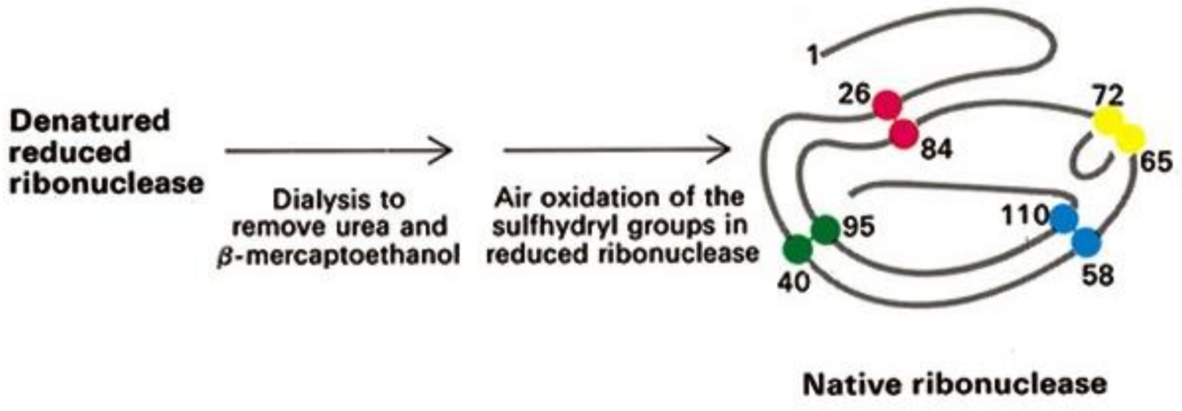

Denatured reduced ribonuclease
Dialysis to remove urea and β-mercaptoethanol
Air oxidation of the sulfhydryl groups in reduced ribonuclease
1
26
84
72
65
95
110
40
58
Native ribonuclease

Figure 20 Anfinsen's Experiment

All the information required for folding a protein into its native structure is present within the protein's amino acid sequence. The native folded form of protein is thermodynamically most stable as compared to others

Module 014: MODELS OF PROTEIN FOLDING

Information required for folding a protein into its native structure is present within the protein's amino acid sequence. The native folded form of protein is thermodynamically most stable as compared to others.

FRAME WORK MODEL

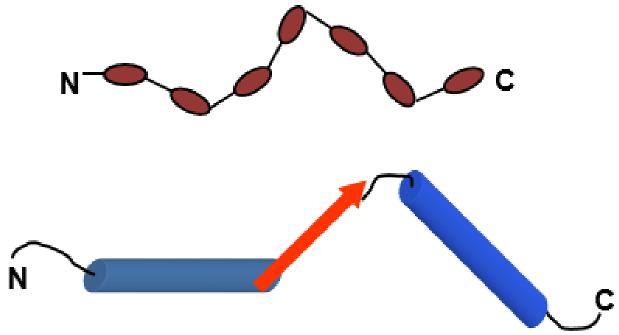


Figure 21 Step 1: Formation of secondary structures

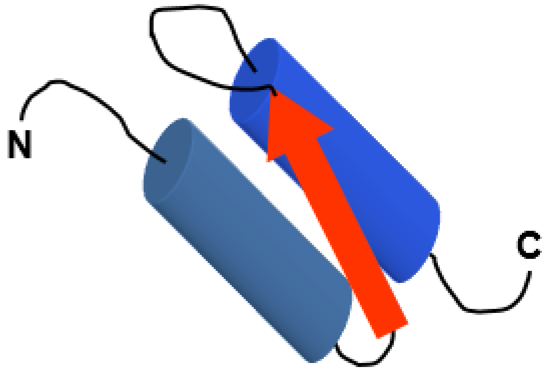


Figure 0.92 Step 2: Arrangement of secondary structures

NUCLEAR CONDENSATION MODEL

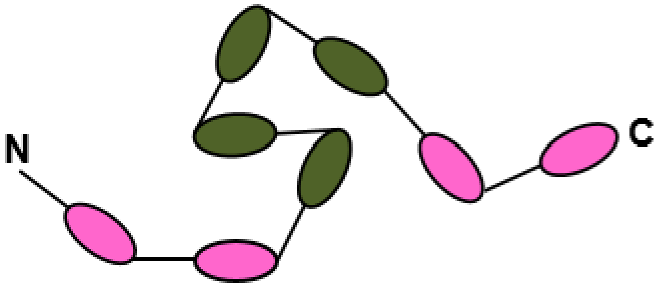


Figure 20.10 Step 1: Formation of a Hydrophobic Core

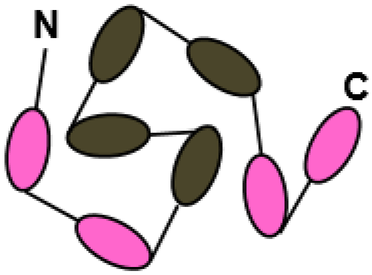


Figure 20.11 Step 2: Including remaining amino acids and ***expanding the nucleus***

Several models exist for folding a protein given its amino acid sequence. The fundamental requirement is that the folding process remain spontaneous. There is still no definitive folding hypothesis.

Module 015: PROTEIN STRUCTURE

Proteins spontaneously fold to take 3D forms. It's a fast yet specific process which leads to a folded protein. Several forces act together to fold the protein structure.

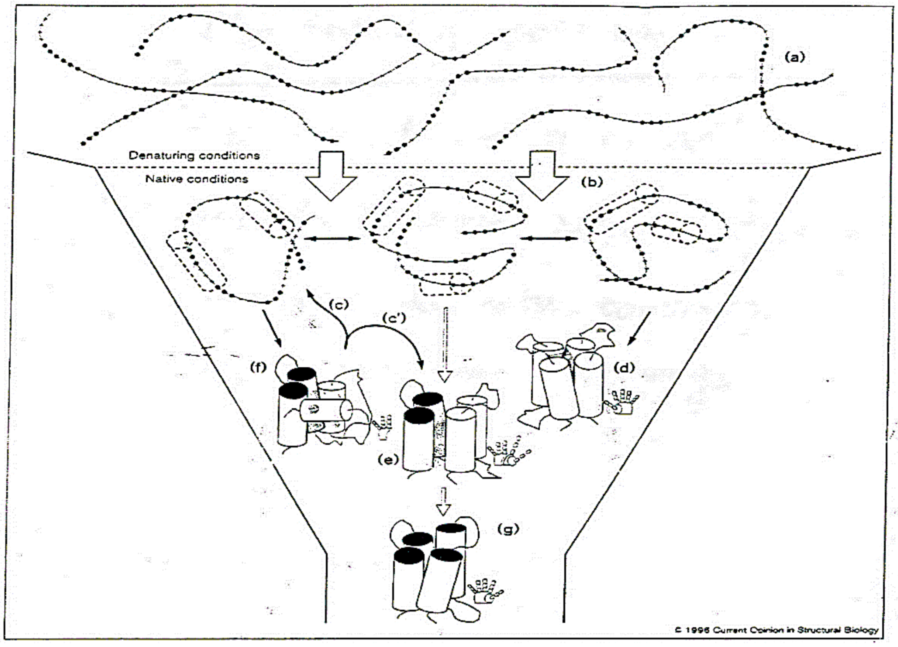


Figure 25 Folding funnel

Bond Type	kJ/mol
Covalent Bond	250
Electrostatic	5
van der Waals	5
Hydrogen bond	20

Figure 0.126 Energies of Various Bonds & Interactions

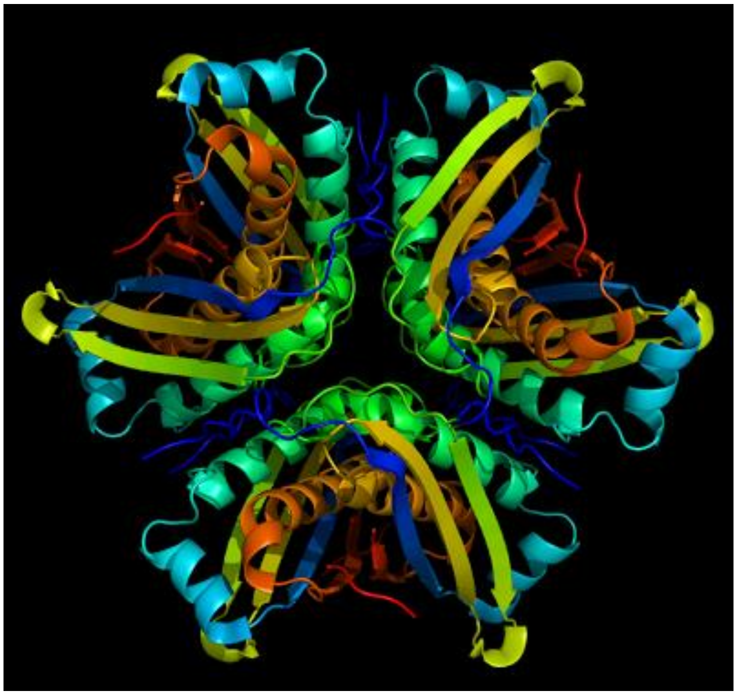


Figure 27 Hydro peroxide resistance protein OsmC (1vla)

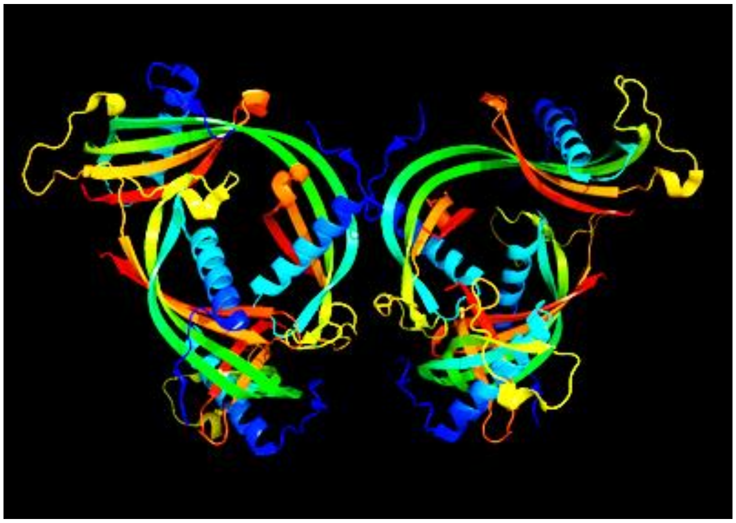


Figure 28 Cystatin – 3 (C) http://beautifulproteins.blogspot.com/

Protein structures are very complex yet they form spontaneously. We will investigate how to develop algorithms to predict such structures.

Module 016: Primary, Secondary, Tertiary and Quaternary Structures

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.

Complex protein structures form spontaneously as a protein folds. A huge variety of protein structures exist. Each structure is designed to perform a specific function. Interestingly, each protein mega structure gets built out of only a few sub-structures. Combinations from the SMALL substructure set are used to construct larger protein structures.

There are many types of structure Single Alphabet Amino acid tags can be put together linearly to represent a protein sequence. This sequence is also called the primary sequence. Primary sequence can also be referred to as 1' structure. Sub-structures are formed as a result of 1' structure's folding. Folded sub structures are called secondary protein structures .Secondary structures are also referred to as 2' structures.

2' sub-structures are packed together to form super structures. These protein super structures are called tertiary structures .Tertiary structures are also referred to as 3' structures.

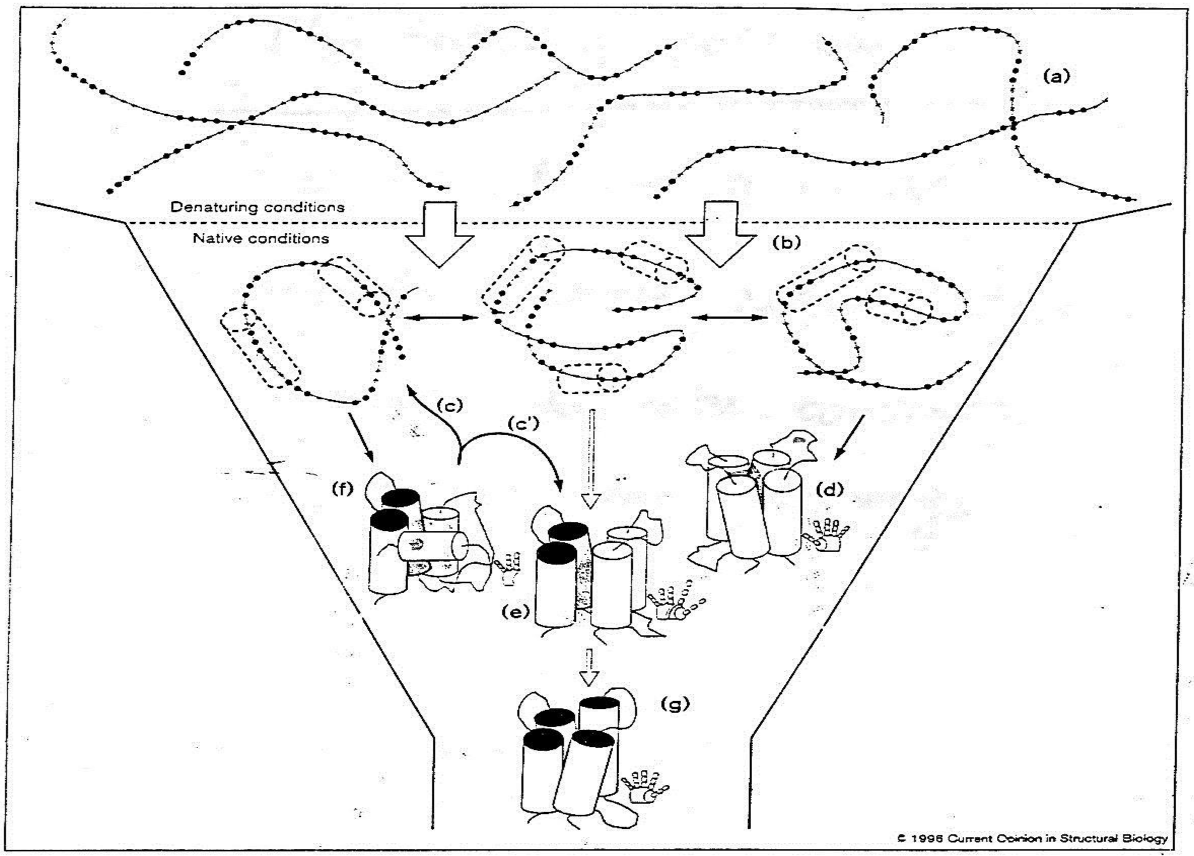
3' structures represent the complete monomeric protein structure.3' structures can combine with other polypeptide units to form a quaternary structure.

Quaternary structures are also called 4' structures. 4' structures are exemplified by protein complexes etc.

Protein structures are organized into 1', 2', 3' and 4' modular conformations. We will investigate how to develop algorithms to predict these structures

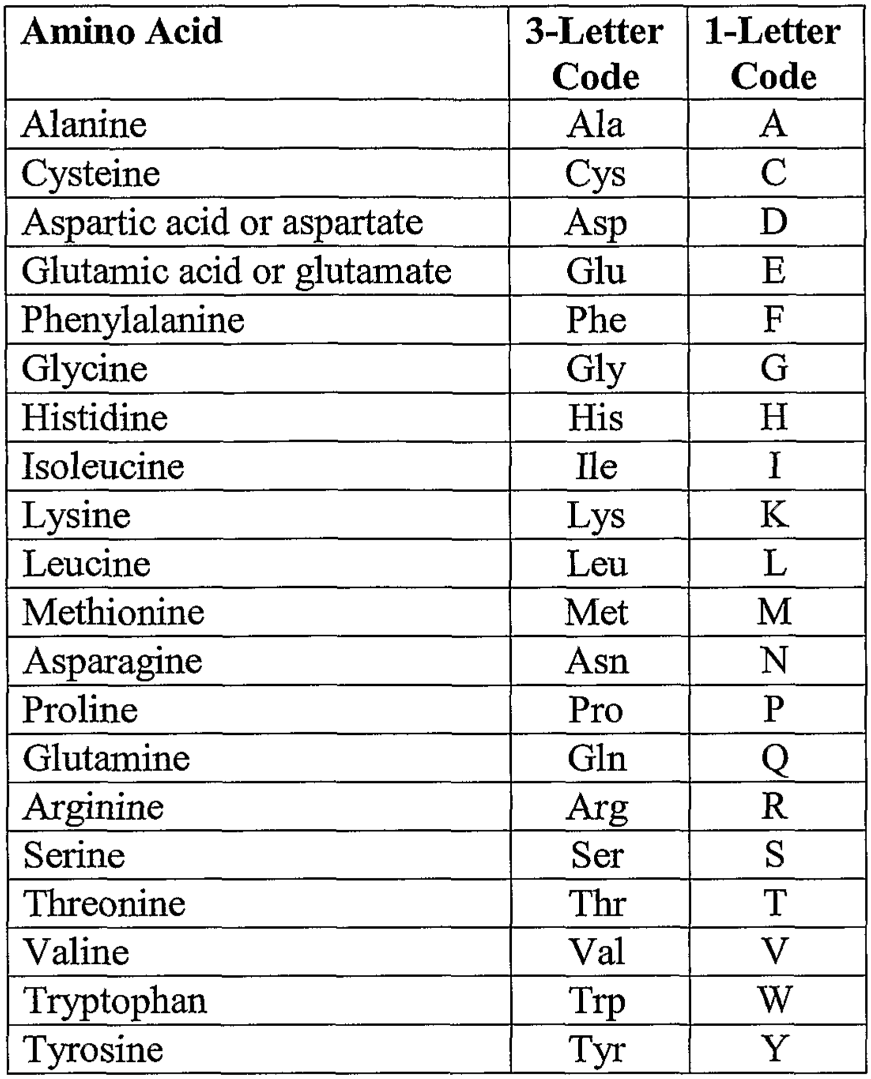
Module 017: Primary Structures of Protein

Protein structures are organized into 1’, 2’, 3’ and 4’ modular conformations. 1’ structures are essentially the amino acid sequence of the proteins.



© 1996 Current Opinion in Structural Biology

Figure 29 protein folding funnel



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

Figure 30 list of amino acids

There are two methods for obtaining 1’ structure.

- Edman Degradation
- Tandem Mass Spectrometry

1’ structure databases are essentially protein sequence databases. Examples include Uniprot, Swissprot amongst several others.

Protein sequences are the primary structures of proteins. The primary or 1’ structure of a protein determines its initial properties.1’ structure lays the foundation for 2’ structures

Module 018: Secondary Structures of Proteins - I

The primary or 1’ structure of a protein determines its basic properties and 1’ structure lays the foundation for 2’ structures. 2’ structures are also referred to as secondary structures.

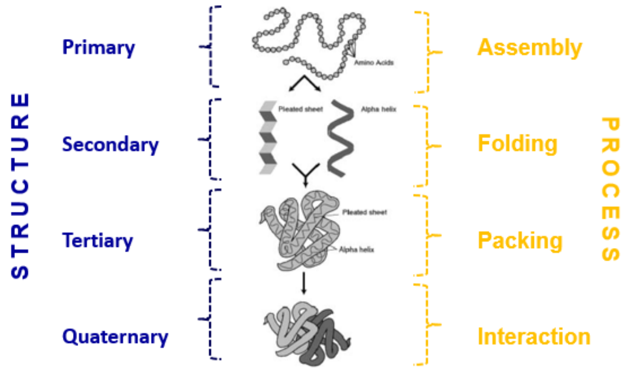


Figure 30.13 Organization of Secondary Structure

Formation of 2’ structure

C- Terminus is negatively charged .N-terminus is positively charged. C and N termini can therefore make Hydrogen Bonds. Hydrogen Bonds are the reason of 2’ structure formation.

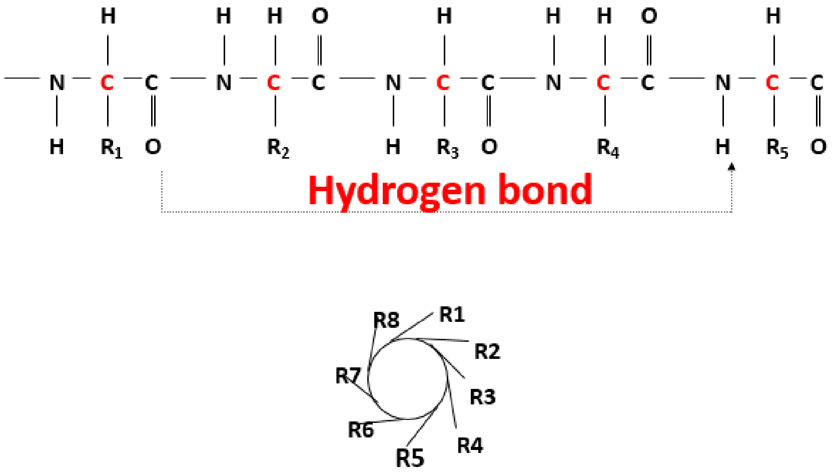


Figure 30.14 Forming Secondary Structure

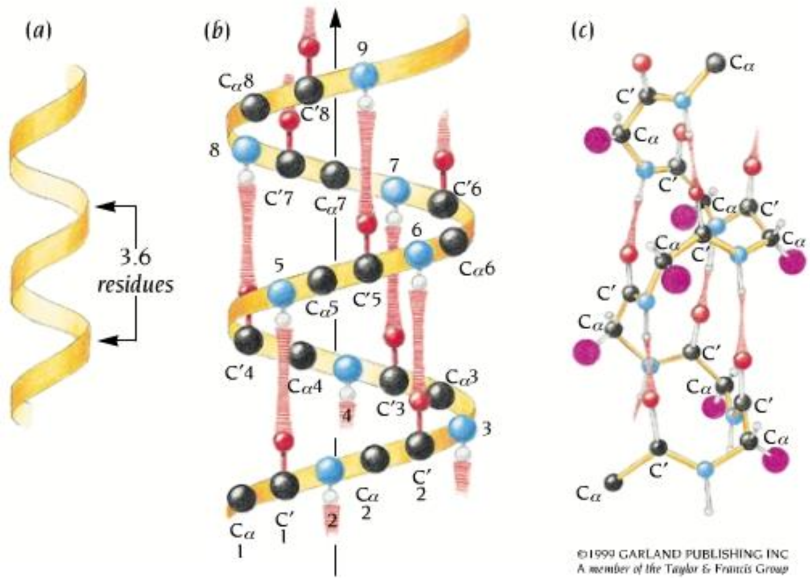


Figure 30.15 Types of Secondary Structures – Alpha Helix

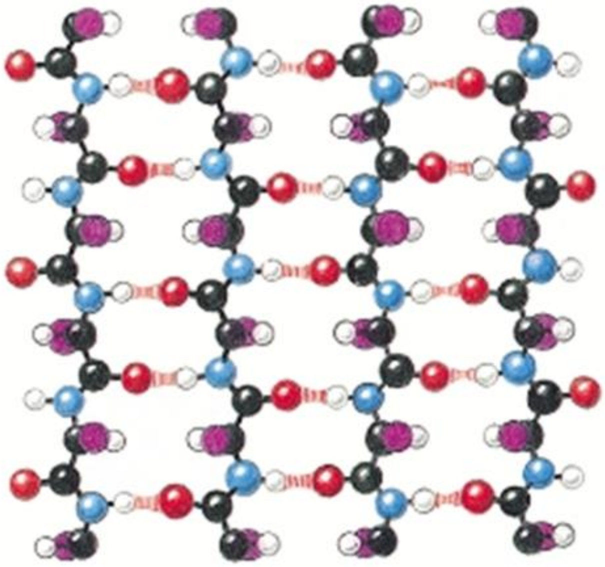
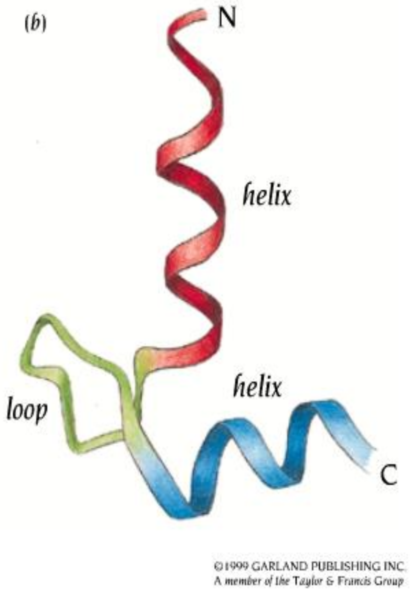


Figure 30.16 Types of Secondary Structures – Beta Sheets

Protein sequences fold onto themselves and make H-Bonds to create 2’ structures. Several types of 2’ structures exist. These include Alpha Helices and Beta Sheets.

Module 019: Secondary Structures of Proteins – II

2’ structures or secondary protein structures are formed as a result of H-Bond formation between N and C termini in a protein backbone. Types of 2’ structures include Alpha helices and Beta sheets.



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

Figure 35 A Special Secondary Structure

Properties of Loop

- Loops connect helices and sheets
- Loops vary in length and 3-D configurations
- Loops are mostly located on the surface of proteins
- Loops are more “acceptable" of mutations
- Loops are flexible and can adopt multiple conformations
- Loops tend to have charged and polar amino acids
- Loops are frequently components of active sites

Coils

- Secondary structure that are not helices, sheets, or recognizable turns
- Disordered regions, but also appear to play important functional roles

Loops and Coils are also secondary structure which form the first structures after folding of protein’s amino acids. Loops and Coils are very important 2’ structures in that they form active sites of proteins.

Module 020: Tertiary Structures of Proteins

2’ structures include alpha helices, beta sheets, loops and coils. Upon combination of 2’ structures, a tertiary or 3’ structure is formed.3’ structure is next level of structure organization.

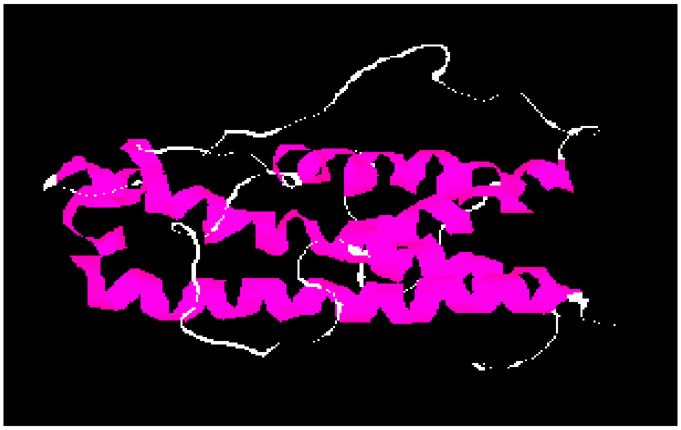


Figure 36 Example of Tertiary Structure

Formation of 3’ structure

- Hydrophobic interactions between nonpolar R-groups
- Covalent bonds in the form of Disulphide bridges

Combinations of Alpha helices, Beta sheets, coils and loops help form 3’ structures. Covalent bonds, Hydrogen bonds and hydrophobic interactions enforce the 3’ structure.

Module 021: Quaternary Structures of Proteins

4’ structures or quaternary structures are formed by different peptide chains that make up the protein. Multimeric proteins which comprise of multiple peptides form 4’ structures.

Monomeric vs. Multimeric Proteins

Protein comprised of only a single chain (monomeric) do not have a quaternary structure. Proteins with multiple chains can form 4’ structures.

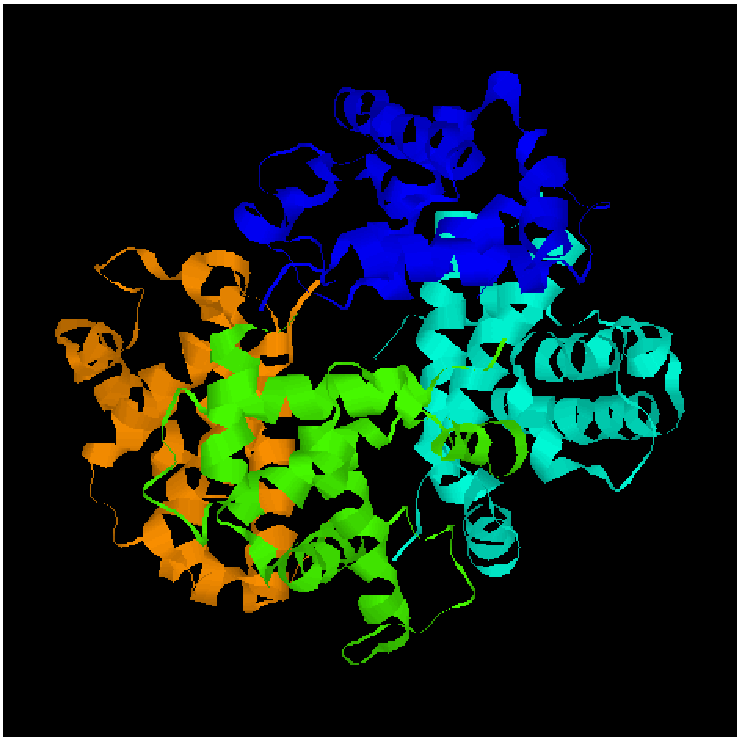


Figure 37 Example of Quaternary Structure See how 2’ and 3’ structures come together

4’ structures are kept in conformation by Hydrogen Bonds, Covalent Disulphide Bonds, Hydrophobic Interactions and ionic bonds. In terms of stability 4’ > 3’ > 2’ > 1’

Module 022: Introduction to Bond Angles in Proteins

Protein folding results in a linear chain of amino acids getting packed into a compact 3D structure. This leads to a reduction in bond angles from an initial of 180 degrees (protein's linear form)

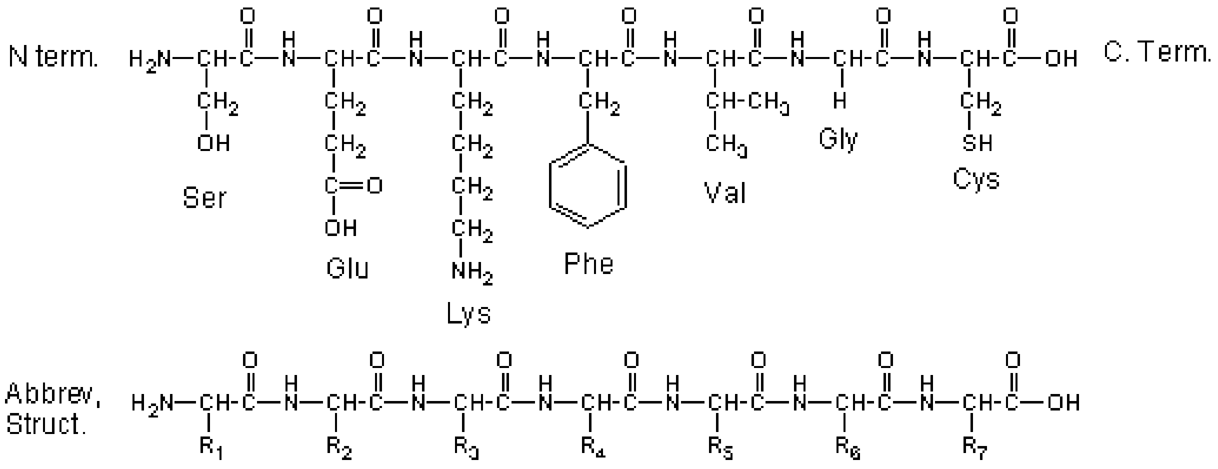
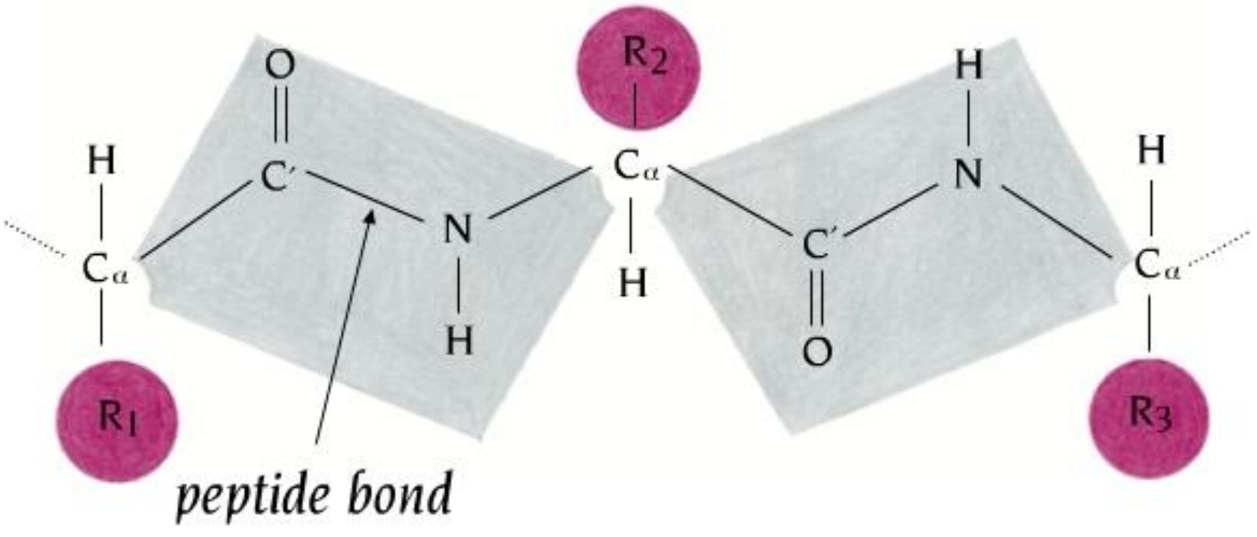


Figure 38 Linear Protein



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

Figure 39 Formation of Planar Peptide Bond

The resultant chain gets its own set of attributes and Peptide bond is planar & rigid.

Dihedral Angles

- Angle between two planes (i.e. 4 points)!

- Considering the middle two points to be aligned (or overlapped), the angle between the 1st, overlapped and the 4th points forms a dihedral angle.

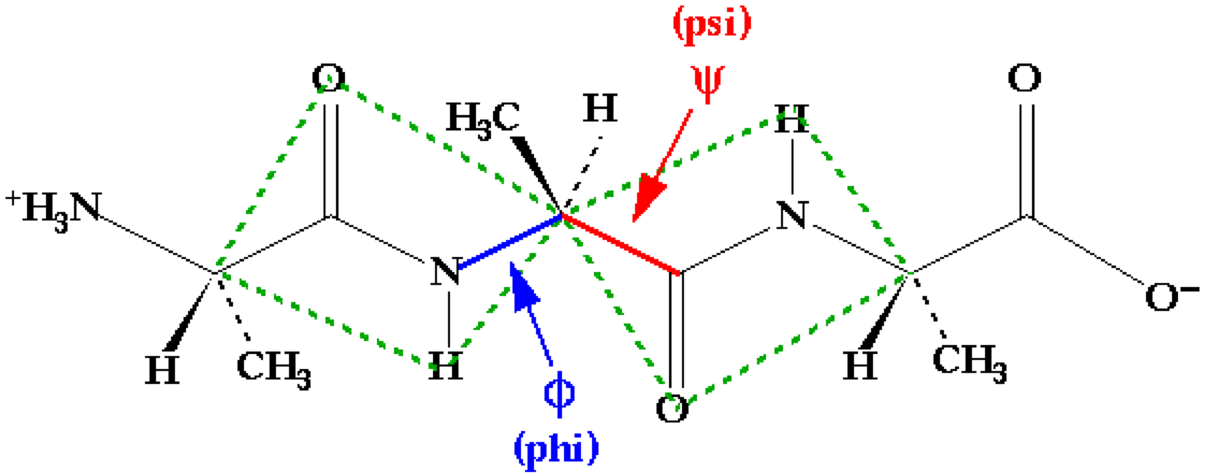


Figure 40 Protein after Folding: Phi and Psi Angles

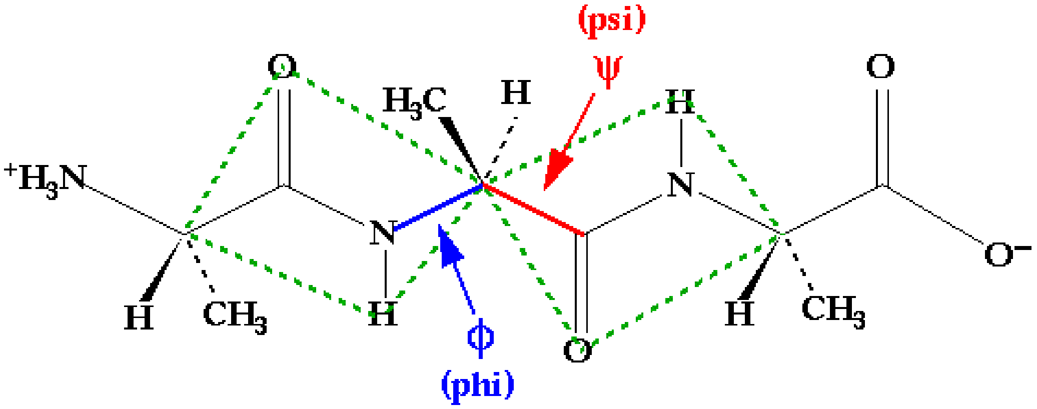


Figure 0.171 Protein after Folding: Phi and Psi Angles

Φ (*phi*, involving C'-N-C^α^-C')
ψ (*psi*, involving N-C^α^-C'-N)

Proteins fold into 3D structures. Phi and psi angles are taken up as a result of folding. These angles can be measured towards understanding the protein structure.

Module 023: Ramachandran Plot

Phi and Psi angles can be measured with in the folded structures like:

- ϕ - *phi*
- Involves C'-N-C$^{\alpha}$-C'
- ψ – *psi*
- Involves N-C$^{\alpha}$-C'-N

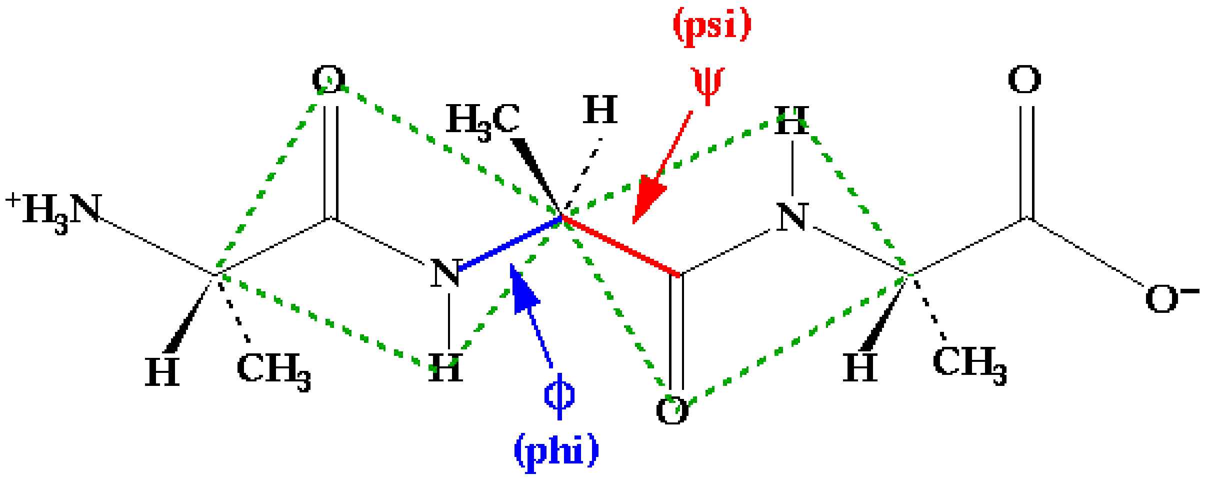


Figure 42 Phi and Psi Angles

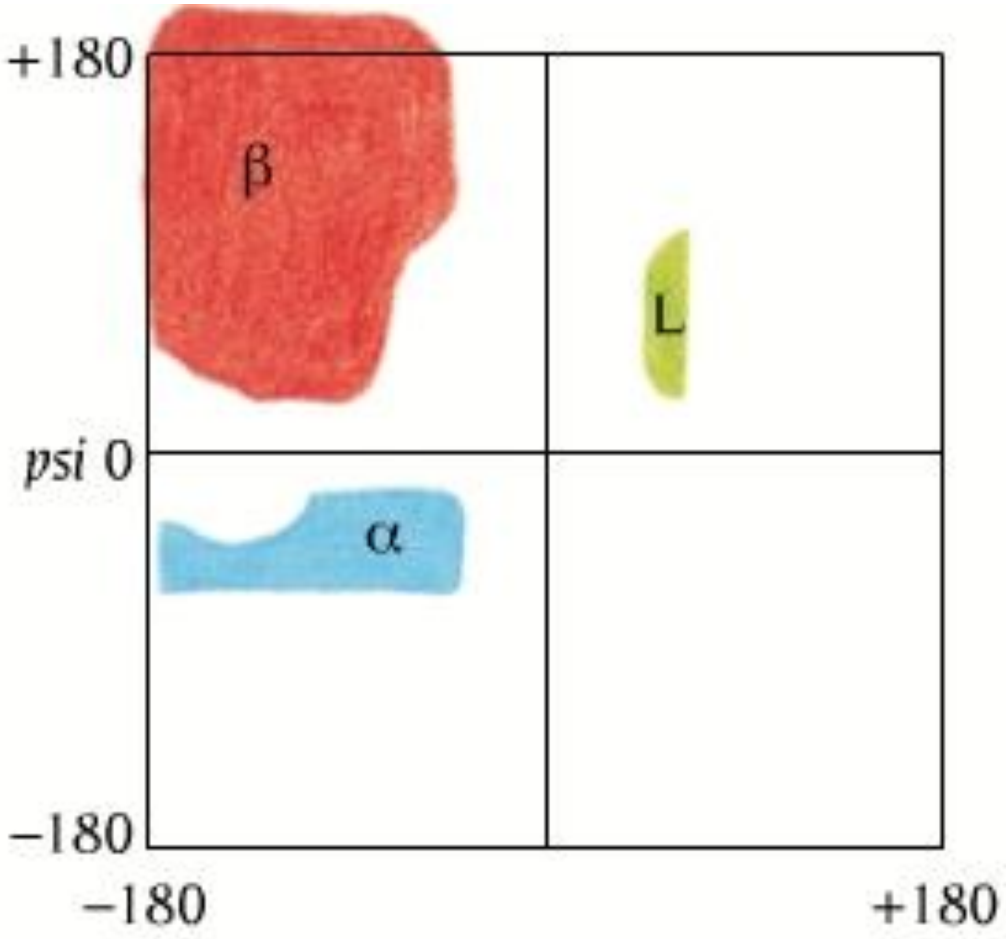
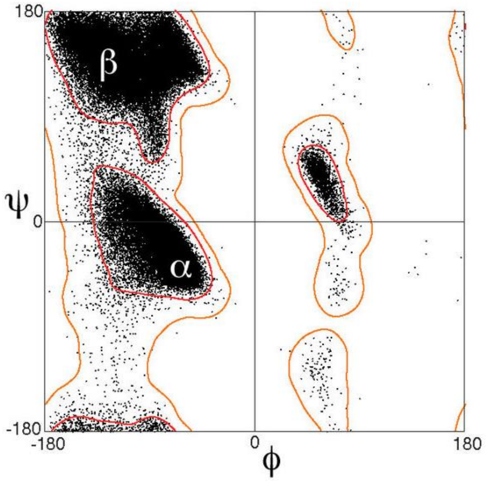


Figure 43 Allowable Phi and Psi Angles



Data as in (Lovell et al. 2003) showing about 100,000 data points for several amino-acids

A limited range of Phi and psi angles are taken up as a result of folding. This range of angles constitutes the allowable range of torsion or rotation angles that are taken up by the protein.

Module 024: Structure Visualization - I

We know that protein backbone takes up specific rotation angles after folding. A protein consists of multiple amino acids. Each amino acid has a C-terminus and an N-Terminus.

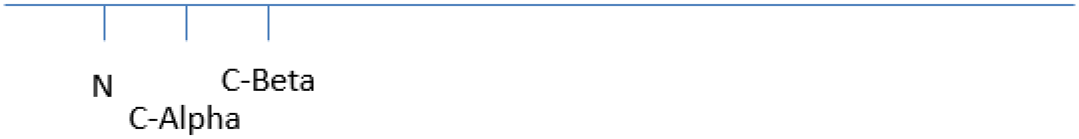


Figure 44 Protein Backbone and C atoms

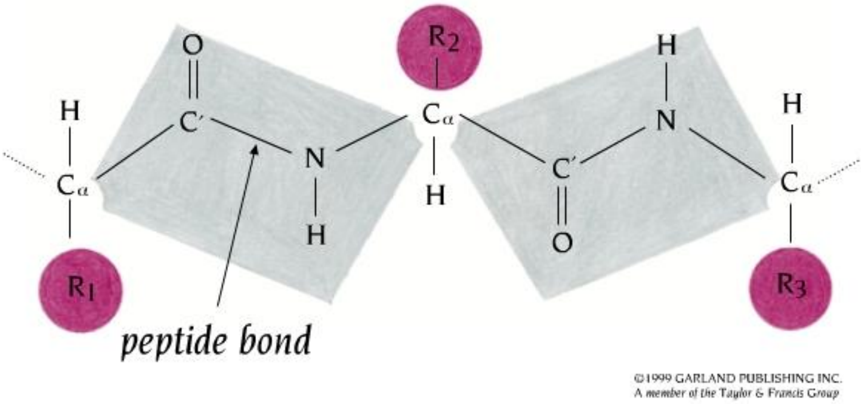
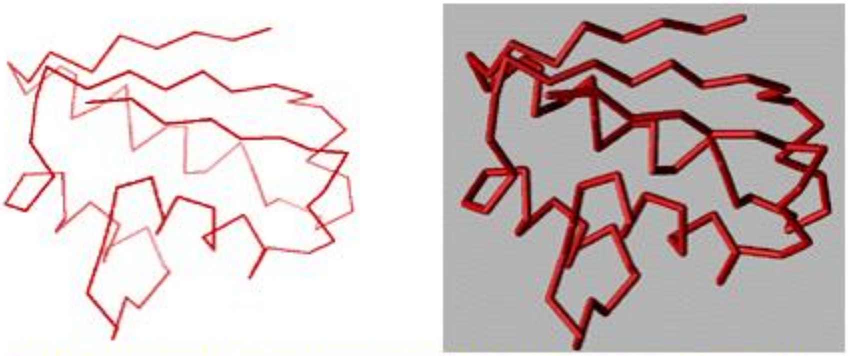

O
R2
H
H
C′
Cα
N
Cα
C′
N
H
H
Cα
H
O
Cα
R1
R3
peptide bond
©1999 GARLAND PUBLISHING INC.
A member of the Taylor & Francis Group

Figure 45 Omitting Planar bonds and Tracing C-Alpha atoms in backbone



http://www.danforthcenter.org/smith/MolView/Over/overview.html

Figure 46 C-Alpha Backbone visualization

C-Alpha atoms are traced to recreate a 3D protein structure. The choice is made while keeping planar nature of the peptide bond in view. Later we will see how to insert side chains into the visual models as well.

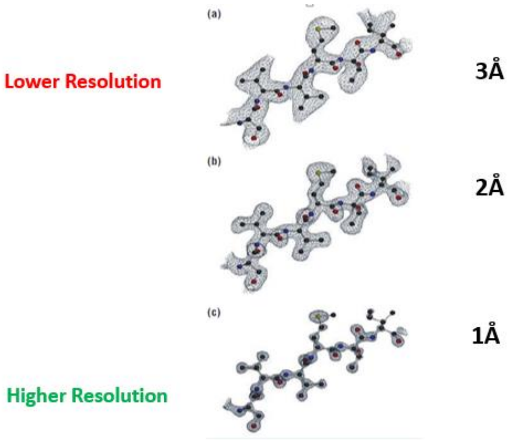
Module 025: Structure Visualization – II

C-Alphas can be used to construct the backbone of a protein towards its visualization. We also need a representation of measurements for assigning the atomic distances. The ångström is used to express the size of atoms, molecules and extremely small biological structures, the lengths of chemical bonds, the arrangement of atoms in crystals.

1 angstrom is a unit of length equal to 10^{-10} m (one ten-billionth of a meter) or 0.1 nm

Atoms of phosphorus, sulfur, and chlorine are ~1 Å in covalent radius, while a hydrogen atom is 0.25 Å

Figure 47 Ansedel Anders Ångström (1814–1874)



C-Alpha atoms are traced to recreate a 3D protein structure. Each C-Alpha atom is at a distance which can be represented in the unit “Angstrom”.1 A resolution is better than 10 A.

Module 026: Experimental Determination of Protein Structure

C-Alpha atoms are traced to recreate a 3D protein structure. Distances between C-Alphas are measured in the unit "Angstrom".

X-Ray Crystallography

- Crystallography data gives relative positions of atomic coordinates
- The data is obtained from diffractions by the atoms in a protein structure
- The coordinates of each atom in x,y and z axis are output

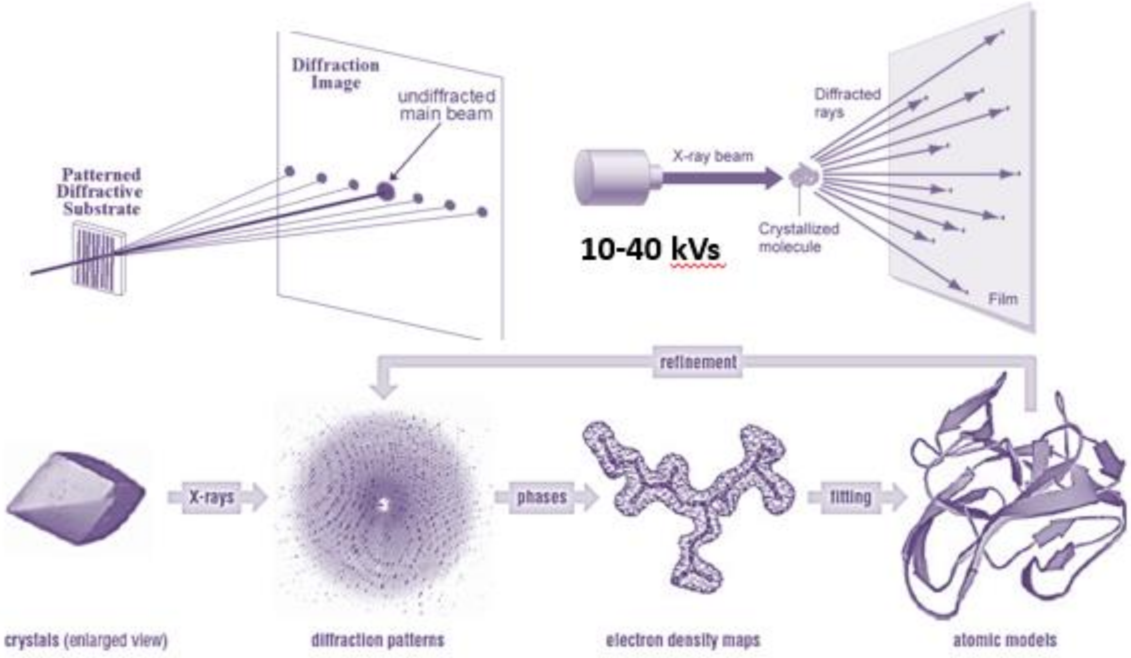
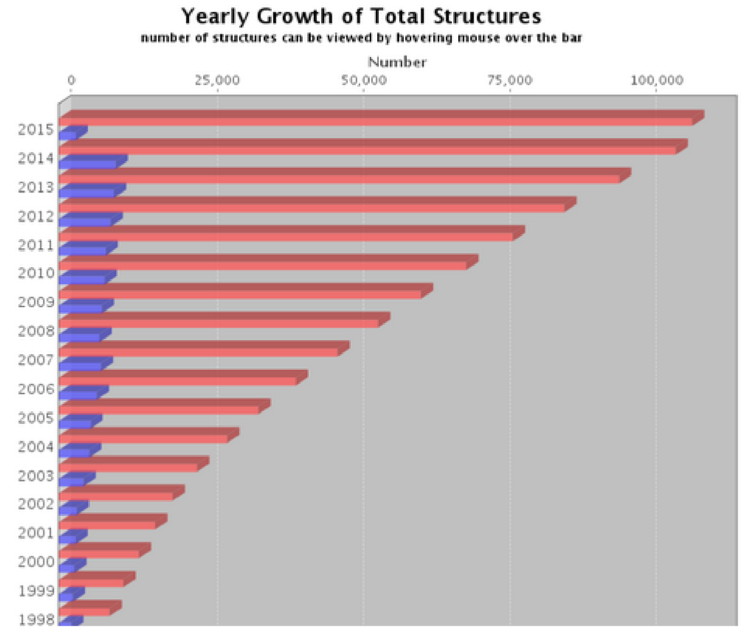


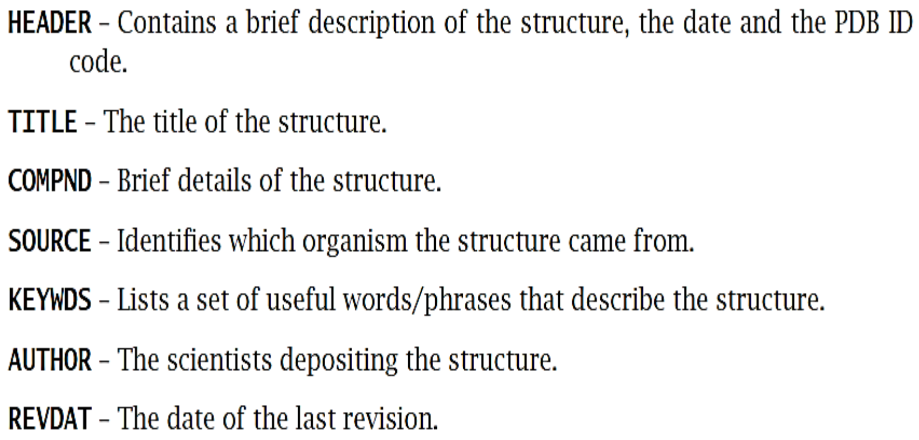
Figure 48 x-ray crystallography

Crystallized proteins are used to determine protein structures. As X-rays diffract from the atoms in a protein, the atomic distances are noted. These distance in 3D are measured in Angstroms.

Module 027: Protein Databank

Position of C-Alpha atoms are used to construct 3D protein structure. X-Ray diffraction data helps measure the atomic positions. X, Y and Z positions of several proteins are available online.





HEADER – Contains a brief description of the structure, the date and the PDB ID code.

TITLE – The title of the structure.

COMPND – Brief details of the structure.

SOURCE – Identifies which organism the structure came from.

KEYWDS – Lists a set of useful words/phrases that describe the structure.

AUTHOR – The scientists depositing the structure.

REVDAT – The date of the last revision.

Figure 49 PDB File Format

JRNL – One or more literature references that describe the structure.

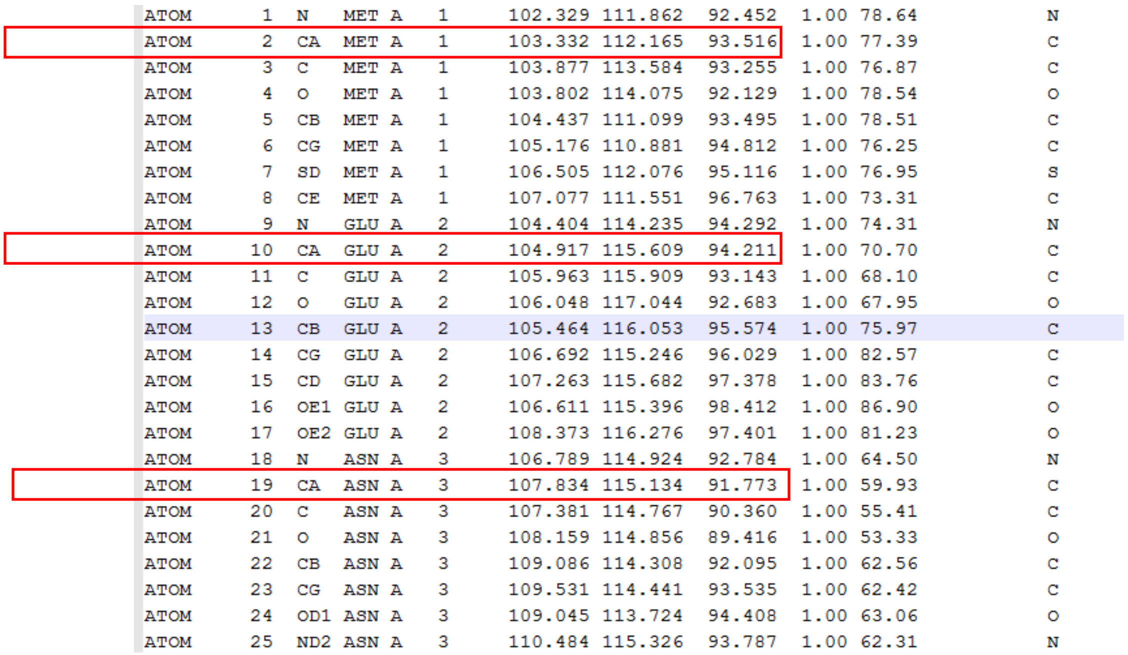
REMARK 1 through REMARK 999 – Details of the experimental methods used to determine the structure are contained in this subsection (see the example in the next section).

DBREF – Cross links to other databases.

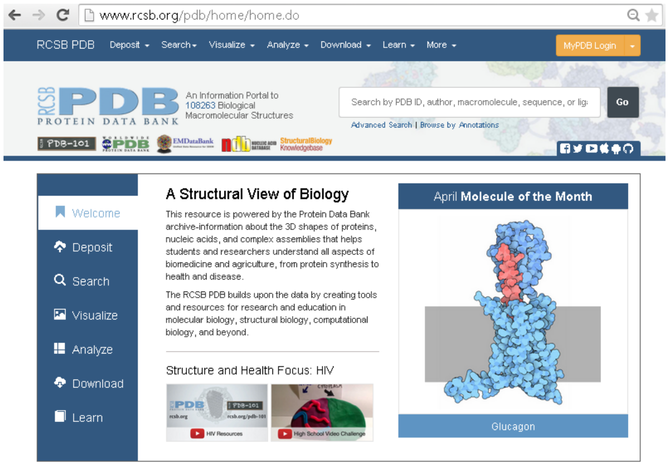
SEQRES – The official amino acid sequence (protein, RNA or DNA) of the structure.

HELIX/SHEET – Details of the regions of secondary structure found in the protein.

ATOM/HETATM – The 3D spatial coordinates of particular atoms in the protein structure (the "ATOM" lines) or other molecules such as water or co-factors (the "HETATM" lines).



```
ATOM      1  N   MET A   1     102.329 111.862  92.452  1.00 78.64           N
ATOM      2  CA  MET A   1     103.332 112.165  93.516  1.00 77.39           C
ATOM      3  C   MET A   1     103.877 113.584  93.255  1.00 76.87           C
ATOM      4  O   MET A   1     103.802 114.075  92.129  1.00 78.54           O
ATOM      5  CB  MET A   1     104.437 111.099  93.495  1.00 78.51           C
ATOM      6  CG  MET A   1     105.176 110.881  94.812  1.00 76.25           C
ATOM      7  SD  MET A   1     106.505 112.076  95.116  1.00 76.95           S
ATOM      8  CE  MET A   1     107.077 111.551  96.763  1.00 73.31           C
ATOM      9  N   GLU A   2     104.404 114.235  94.292  1.00 74.31           N
ATOM     10  CA  GLU A   2     104.917 115.609  94.211  1.00 70.70           C
ATOM     11  C   GLU A   2     105.963 115.909  93.143  1.00 68.10           C
ATOM     12  O   GLU A   2     106.048 117.044  92.683  1.00 67.95           O
ATOM     13  CB  GLU A   2     105.464 116.053  95.574  1.00 75.97           C
ATOM     14  CG  GLU A   2     106.692 115.246  96.029  1.00 82.57           C
ATOM     15  CD  GLU A   2     107.263 115.682  97.378  1.00 83.76           C
ATOM     16  OE1 GLU A   2     106.611 115.396  98.412  1.00 86.90           O
ATOM     17  OE2 GLU A   2     108.373 116.276  97.401  1.00 81.23           O
ATOM     18  N   ASN A   3     106.789 114.924  92.784  1.00 64.50           N
ATOM     19  CA  ASN A   3     107.834 115.134  91.773  1.00 59.93           C
ATOM     20  C   ASN A   3     107.381 114.767  90.360  1.00 55.41           C
ATOM     21  O   ASN A   3     108.159 114.856  89.416  1.00 53.33           O
ATOM     22  CB  ASN A   3     109.086 114.308  92.095  1.00 62.56           C
ATOM     23  CG  ASN A   3     109.531 114.441  93.535  1.00 62.42           C
ATOM     24  OD1 ASN A   3     109.045 113.724  94.408  1.00 63.06           O
ATOM     25  ND2 ASN A   3     110.484 115.326  93.787  1.00 62.31           N
```



www.rcsb.org/pdb/home/home.do
RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
More
MyPDB Login
RCSB PDB PROTEIN DATA BANK
Search by PDB ID, author, macromolecule, sequence, or lig
Go
Advanced Search | Browse by Annotations
PDB-101
WORLDWIDE PDB PROTEIN DATA BANK
EMDataBank
ndb NUCLEIC ACID DATABASE
StructuralBiology Knowledgebase
Welcome
Deposit
Search
Visualize
Analyze
Download
Learn
HIV Resources
High School Video Challenge
April Molecule of the Month
Glucagon

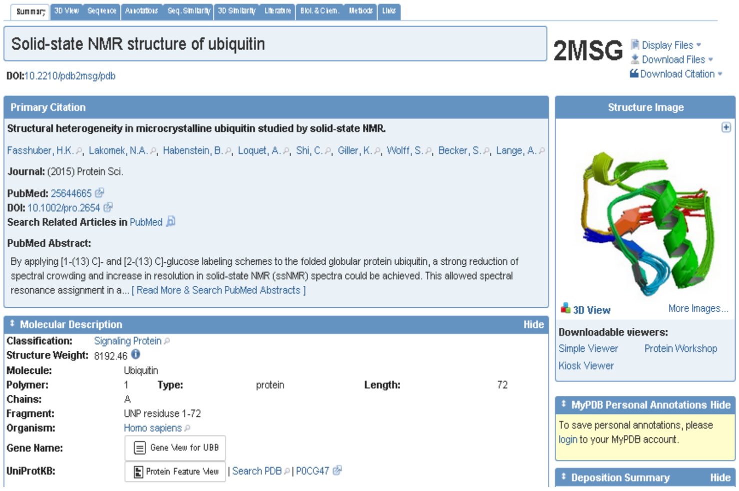
An Information Portal to 108263 Biological Macromolecular Structures

A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

Structure and Health Focus: HIV



Solid-state NMR structure of ubiquitin

2MSG

DOI:10.2210/pdb2msg/pdb

Primary Citation

Structural heterogeneity in microcrystalline ubiquitin studied by solid-state NMR.

Fasshuber, H.K., Lakomek, N.A., Habenstein, B., Loquet, A., Shi, C., Giller, K., Wolff, S., Becker, S., Lange, A.

Journal: (2015) Protein Sci.

PubMed: 25644665
DOI: 10.1002/pro.2654
Search Related Articles in PubMed

PubMed Abstract:

By applying [1-(13) C]- and [2-(13) C]-glucose labeling schemes to the folded globular protein ubiquitin, a strong reduction of spectral crowding and increase in resolution in solid-state NMR (ssNMR) spectra could be achieved. This allowed spectral resonance assignment in a... [Read More & Search PubMed Abstracts]

Molecular Description Hide

Classification:	Signaling Protein				
Structure Weight:	8192.46				
Molecule:	Ubiquitin				
Polymer:	1	**Type:**	protein	**Length:**	72
Chains:	A				
Fragment:	UNP residuse 1-72				
Organism:	Homo sapiens				
Gene Name:	Gene View for UBB				
UniProtKB:	Protein Feature View | Search PDB | P0CG47				

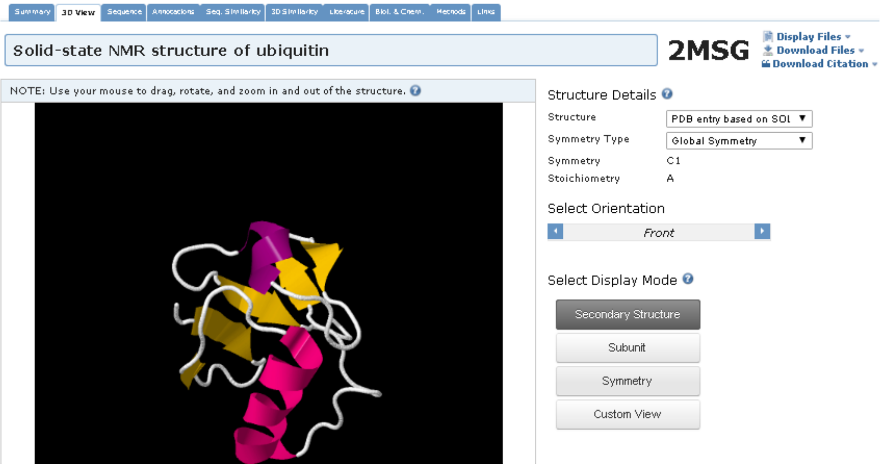
Structure Image

Downloadable viewers:
Simple Viewer Protein Workshop
Kiosk Viewer

MyPDB Personal Annotations Hide

To save personal annotations, please login to your MyPDB account.

Deposition Summary Hide



Solid-state NMR structure of ubiquitin

2MSG

NOTE: Use your mouse to drag, rotate, and zoom in and out of the structure.

Structure Details

Structure	PDB entry based on SOL
Symmetry Type	Global Symmetry
Symmetry	C1
Stoichiometry	A

Select Orientation

Front

Select Display Mode

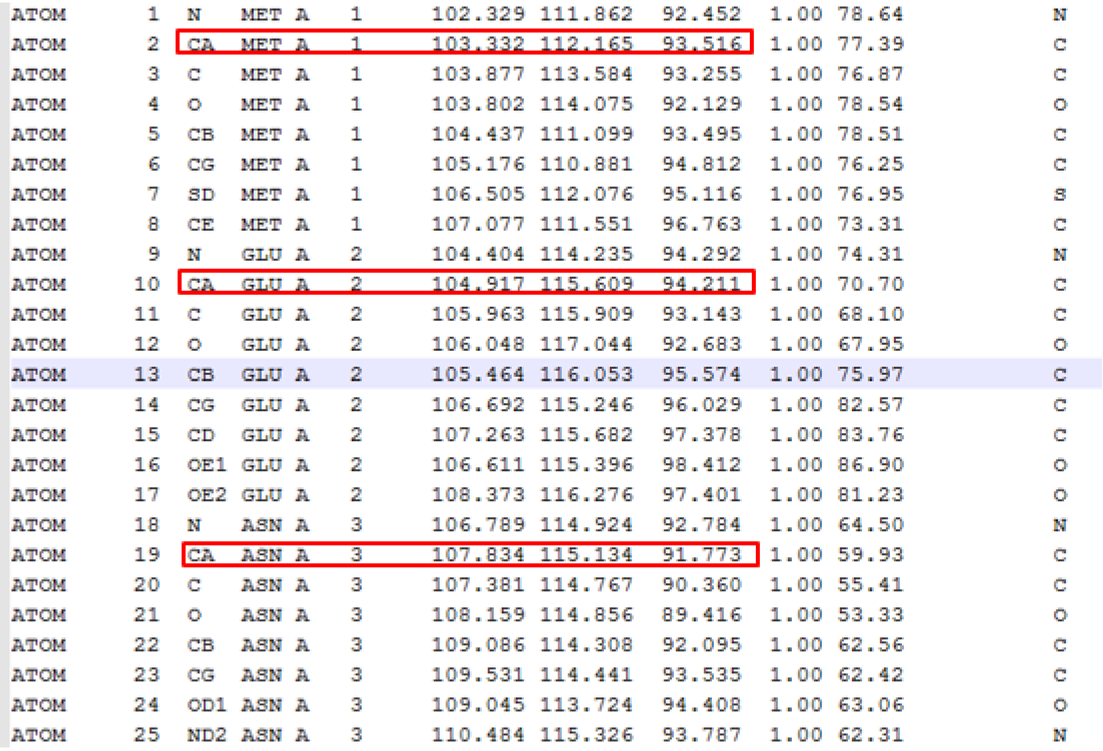
PDB contains protein structure information. It has the coordinates of C-Alphas for over 50,000 proteins. Protein structures can be visualized using this information.

Module 028: Visualization Technique

Proteins fold into 3D structures. Phi and psi angles are assumed as a result of folding. These angles can be measured and viewed towards understanding the protein structure. To view a protein, we need to evaluate the physical location of its atoms. Proteins have Carbon and Nitrogen in their backbone.

CA atomic coordinates

- To trace the backbone of a protein, CA atoms trace can be used
- Note that CA atoms have the side chains attached to them
- A coordinates can be found in the PDB file

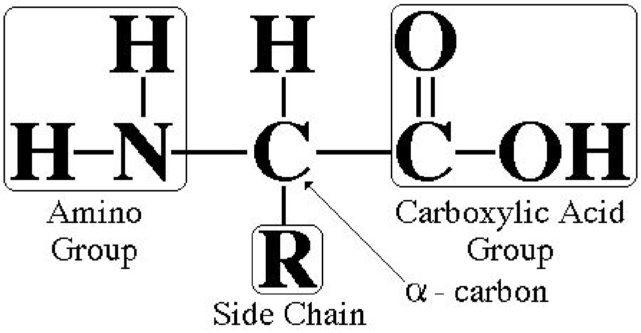


```
ATOM      1  N   MET A   1     102.329 111.862  92.452  1.00 78.64           N
ATOM      2  CA  MET A   1     103.332 112.165  93.516  1.00 77.39           C
ATOM      3  C   MET A   1     103.877 113.584  93.255  1.00 76.87           C
ATOM      4  O   MET A   1     103.802 114.075  92.129  1.00 78.54           O
ATOM      5  CB  MET A   1     104.437 111.099  93.495  1.00 78.51           C
ATOM      6  CG  MET A   1     105.176 110.881  94.812  1.00 76.25           C
ATOM      7  SD  MET A   1     106.505 112.076  95.116  1.00 76.95           S
ATOM      8  CE  MET A   1     107.077 111.551  96.763  1.00 73.31           C
ATOM      9  N   GLU A   2     104.404 114.235  94.292  1.00 74.31           N
ATOM     10  CA  GLU A   2     104.917 115.609  94.211  1.00 70.70           C
ATOM     11  C   GLU A   2     105.963 115.909  93.143  1.00 68.10           C
ATOM     12  O   GLU A   2     106.048 117.044  92.683  1.00 67.95           O
ATOM     13  CB  GLU A   2     105.464 116.053  95.574  1.00 75.97           C
ATOM     14  CG  GLU A   2     106.692 115.246  96.029  1.00 82.57           C
ATOM     15  CD  GLU A   2     107.263 115.682  97.378  1.00 83.76           C
ATOM     16  OE1 GLU A   2     106.611 115.396  98.412  1.00 86.90           O
ATOM     17  OE2 GLU A   2     108.373 116.276  97.401  1.00 81.23           O
ATOM     18  N   ASN A   3     106.789 114.924  92.784  1.00 64.50           N
ATOM     19  CA  ASN A   3     107.834 115.134  91.773  1.00 59.93           C
ATOM     20  C   ASN A   3     107.381 114.767  90.360  1.00 55.41           C
ATOM     21  O   ASN A   3     108.159 114.856  89.416  1.00 53.33           O
ATOM     22  CB  ASN A   3     109.086 114.308  92.095  1.00 62.56           C
ATOM     23  CG  ASN A   3     109.531 114.441  93.535  1.00 62.42           C
ATOM     24  OD1 ASN A   3     109.045 113.724  94.408  1.00 63.06           O
ATOM     25  ND2 ASN A   3     110.484 115.326  93.787  1.00 62.31           N
```

Protein structures can be visualized by tracing the CA atoms. Coordinates of CA atoms can be obtained from the PDB. Next, we need a tool to plot these coordinates.

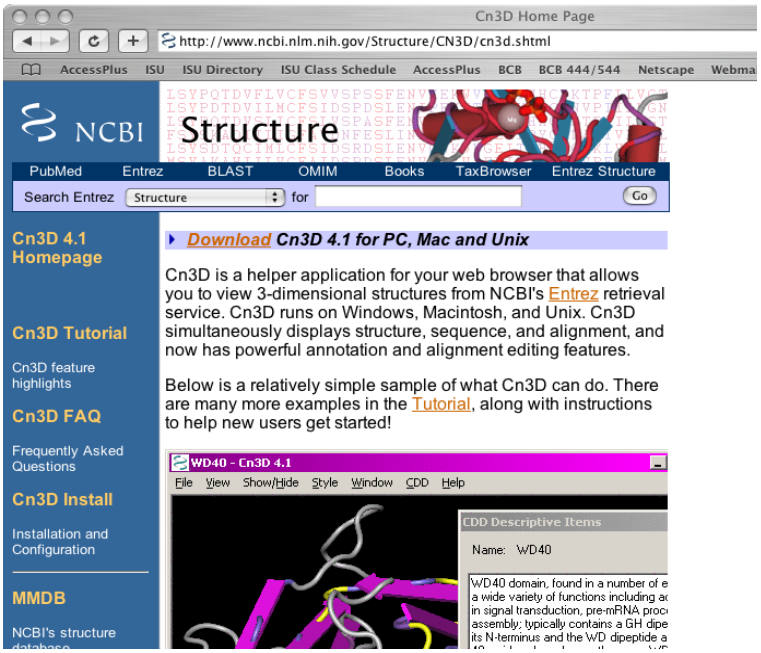
Module 029: Online Resources for Protein Visualization

Protein structures can be visualized by tracing the CA atoms. CA Coordinates can be taken from PDB.



Online Tools

- Rasmol and CHIME are basic tools for visualizing proteins
- Swiss PDB Viewer offers several features such as protein surface view, alignment of several proteins & modelling secondary structures
- PyMOL is a python-script based tool for visualizing the protein structure
- Cn3D is another tool which helps us visualize protein structures
- It also provides for annotating protein structures



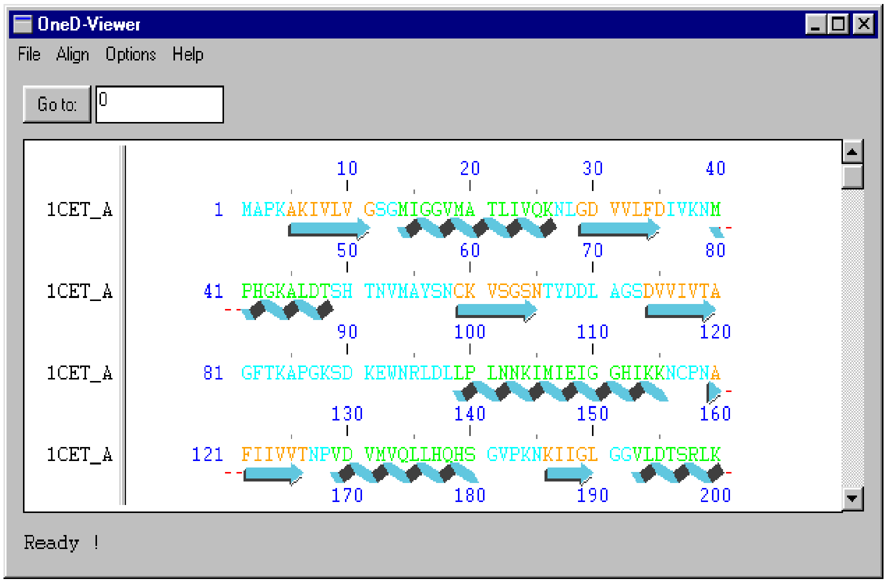
Download Cn3D 4.1 for PC, Mac and Unix

Cn3D is a helper application for your web browser that allows you to view 3-dimensional structures from NCBI's Entrez retrieval service. Cn3D runs on Windows, Macintosh, and Unix. Cn3D simultaneously displays structure, sequence, and alignment, and now has powerful annotation and alignment editing features.

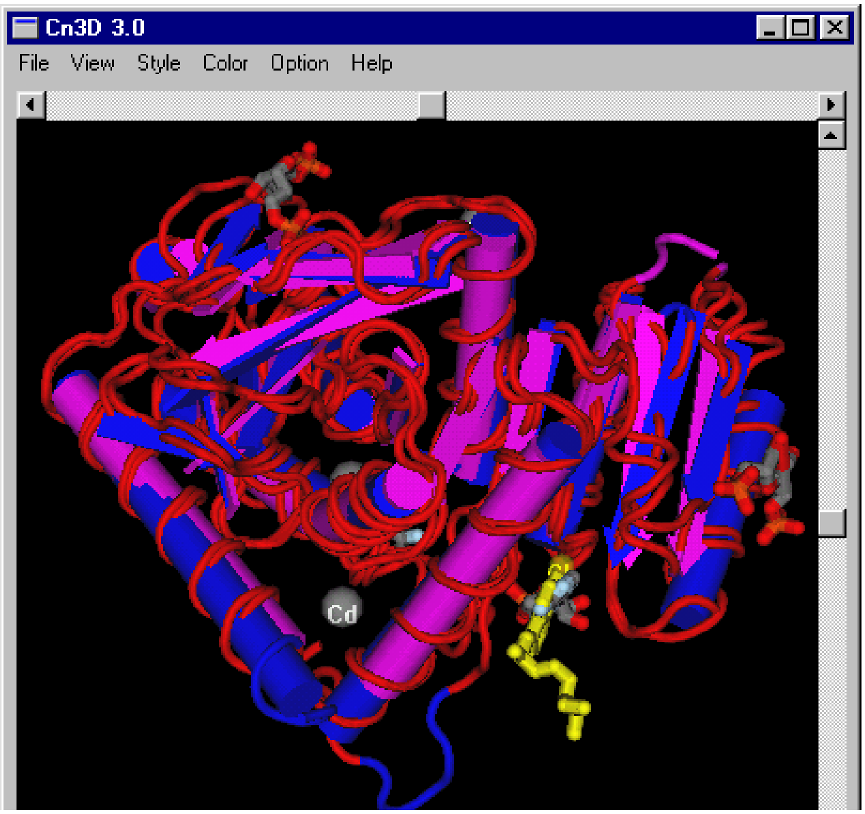
Below is a relatively simple sample of what Cn3D can do. There are many more examples in the Tutorial, along with instructions to help new users get started!

Name: WD40

WD40 domain, found in a number of e
a wide variety of functions including a
in signal transduction, pre-mRNA proc
assembly; typically contains a GH dipe
its N-terminus and the WD dipeptide a



```
                     10         20         30         40
1CET_A       1  MAPKAKIVLV GSGMIGGVMA TLIVQKNLGD VVLFDIVKNM
                     50         60         70         80
1CET_A      41  PHGKALDTSH TNVMAYSNCK VSGSNTYDDL AGSDVVIVTA
                     90        100        110        120
1CET_A      81  GFTKAPGKSD KEWNRLDLLP LNNKIMIEIG GHIKKNCPNA
                    130        140        150        160
1CET_A     121  FIIVVTNPVD VMVQLLHQHS GVPKNKIIGL GGVLDTSRLK
                    170        180        190        200
Ready !
```



Protein structures are visualized using several online tools. These tools include Rasmol, CHIME, Swiss PDB Viewer and Cn3D.

Module 030: Types of Protein Visualizations

To visualize proteins, we use CA coordinates or positions. We can use several online tools to view the resulting model.

CPK: Corey-Paulin-Koltun Diagrams. In CPK diagrams, each atom is represented by a solid sphere. Spheres are equal to atomic van der Waal radius (the volume of the atom).

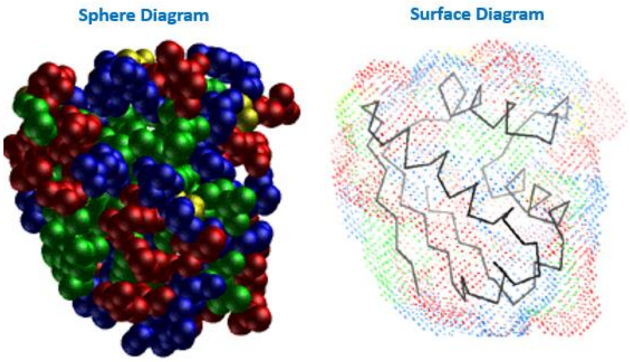


Figure 50 sphere and surface diagrams of protein

http://www.danforthcenter.org/smith/MolView/Over/overview.html

Ribbon Diagrams

Ribbon diagrams are an easy and frequently used technique for representing protein structures. Structure is represented by the secondary structures (fold) using simple cartoon figures.

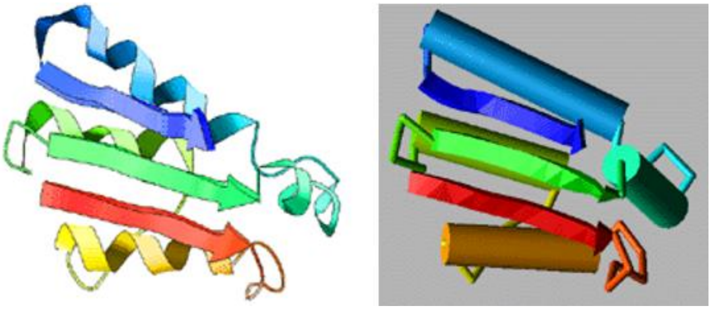


Figure 51 ribbon diagrams

http://www.danforthcenter.org/smith/MolView/Over/overview.html

Balls & Stick (BS) Models

BS model is another popular protein structure representation strategy. BS Models have atoms as colored balls and intermediate bonds as sticks.

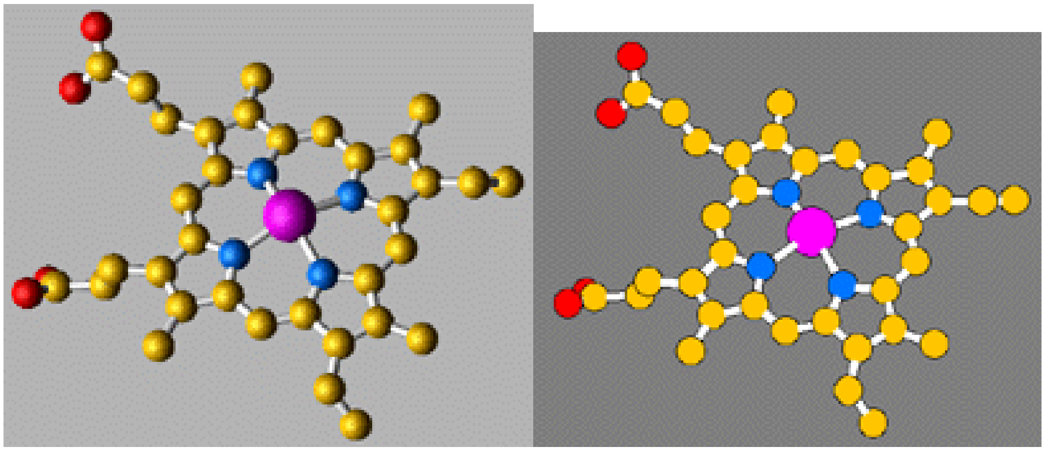


Figure 52 Balls and sticks model

http://www.danforthcenter.org/smith/MolView/Over/overview.html

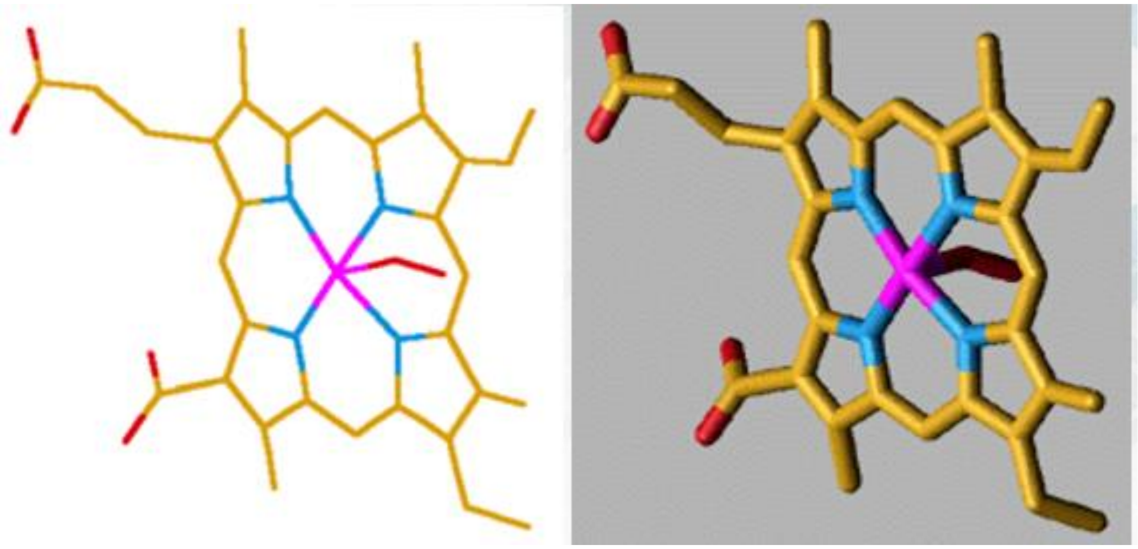


Figure 53 Colored Sticks Models

http://www.danforthcenter.org/smith/MolView/Over/overview.html

Protein Structure Visualization can be performed using several atomic representations. These include CPK, Ribbon and Balls & Stick Diagrams.

Module 031: Introduction to Energy of Protein Structures

Proteins come together as a result of peptide bond formation between various amino acids. The resulting polymer then goes through the step of folding which leads to the formation of a 3D structure.

https://folding.stanford.edu/home/the-science/

Role of Amino Acids

We know that amino acids can be polar, charged and hydrophobic. Role of polar and charged amino acids in folding. Role of hydrophobic amino acids in folding.

Overall Goal of Folding

Anfinsen's thermodynamic hypothesis: Proteins fold for a unique, stable and minimum free kinetic energy structure. What other factors may come into play for satisfying Anfinsen hypothesis.

Minimizing Energy

We know that if bonds can be formed between two atoms, then energy is released. This leads to a situation where there is lesser free energy accessible to each atom for further interactions. So, proteins maximize bonds that can be made between the side chains on each of their constituent amino acids

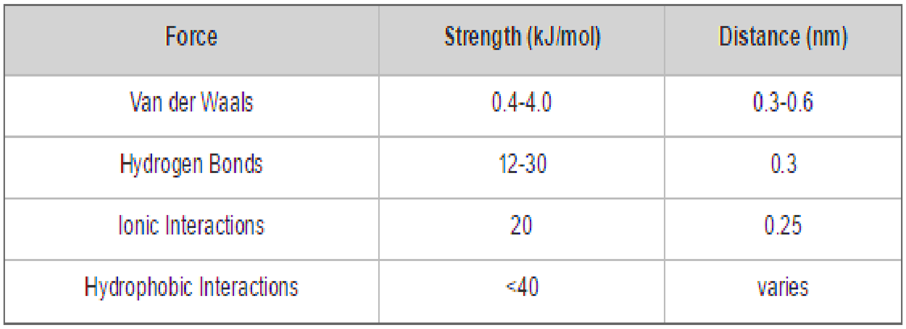
Such atomic interactions include:

- Disulphide bonds between Cysteine residues
- Hydrogen Bonds
- Van der Waals Forces
- Electrostatic Interactions between polar/charged amino acids

The greater the number of these bonds, the more stable a protein becomes. Hence, the basic idea of thermodynamic stability is to maximize bonding in order to minimize the free energy

Module 032: Calculating Energy of a Protein Structure

As we know the greater the number of bonds between the amino acids, the more stable a protein becomes.



Force	Strength (kJ/mol)	Distance (nm)
Van der Waals	0.4-4.0	0.3-0.6
Hydrogen Bonds	12-30	0.3
Ionic Interactions	20	0.25
Hydrophobic Interactions	<40	varies

Figure 54 Energies of Interactions www.ucdavis.edu

Comparison of bond energies

- Hydrophobic interactions >
- Electrostatic interactions>
- Hydrogen bond > van der Waals

Calculating overall energy of a protein structure

Given the number of atomic interactions in a protein, you can simply sum the energy in the protein molecule.

$$\begin{aligned}\text{Energy}_{\text{TOTAL}} = {} & \text{Atoms}_{\text{VWF}} \text{ x } \text{Energy}_{\text{VWF}} + \\ & \text{Atoms}_{\text{HB}} \text{ x } \text{Energy}_{\text{HB}} + \\ & \text{Atoms}_{\text{IonicInteraction}} \text{ x } \text{Energy}_{\text{IonicInteraction}}\end{aligned}$$

Energies of protein structures can be computed by first enumerating the types of interactions between each atom. Then, accumulating the energy of each interaction towards calculating an overall energy of a protein.

Module 033: Structure Determination for Energy Calculations

The greater the interactions between the amino acids, the more stable a protein becomes. We can calculate energy of a folded protein based on the number and types of atomic interactions.

How to find the number of interactions

- To determine the number of each type of interaction within a protein, we need to find its inter-atomic distances.
- Based on specific atomic distances, we can guess the type of atomic interaction.
- By looking up at the bond/energy table, we can compute the overall energy.

Techniques for structure determination

- X-Ray Crystallography
- Nuclear Magnetic Resonance (NMR) Spectroscopy

We need to know the structure of the protein to calculate atomic distances. Atomic distances tell us about atomic interactions with neighboring atoms. To determine the structure, we use X-Ray or NMR.

Module 034: Review of Experimental Structure Determination

The greater the interactions between the amino acids, the more stable a protein becomes. We can calculate energy of a folded protein based on the number and types of atomic interactions.

The structure also dictates which functions a protein can perform via the positioning of hydrophilic & polar amino acids. For determining stability, structure & function, we need to find the amino acid interactions. Several experimental methods exist for structure determination.

- X-Ray Crystallography
- Nuclear Magnetic Resonance (NMR) Spectroscopy

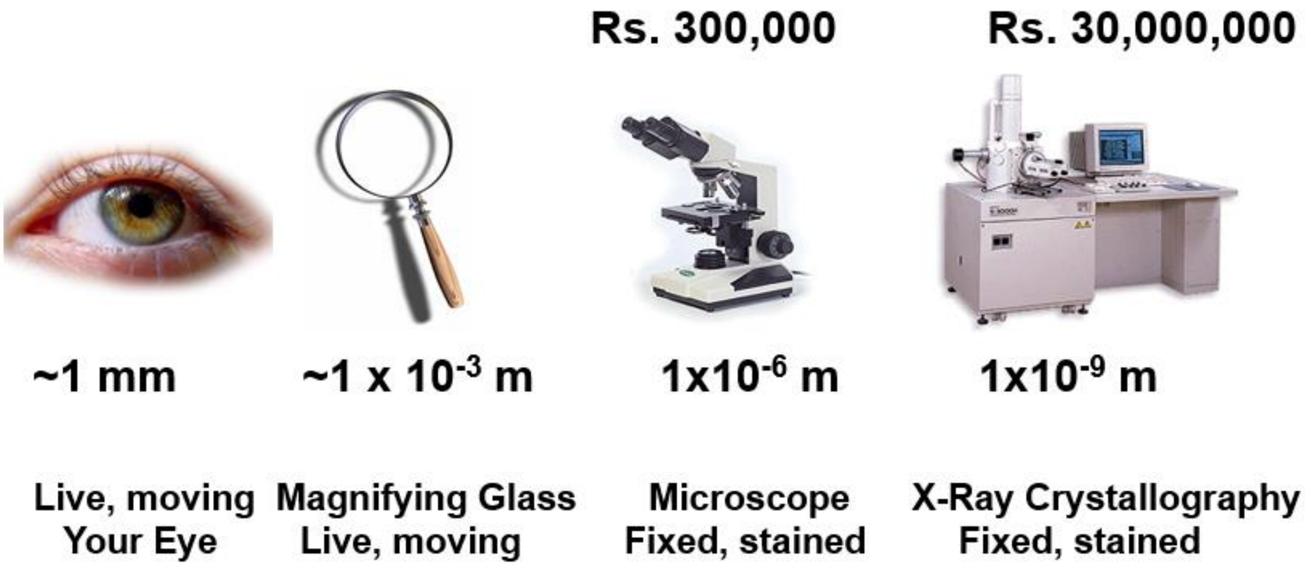


Figure 55 to measure a bond/interaction, we must first see atoms

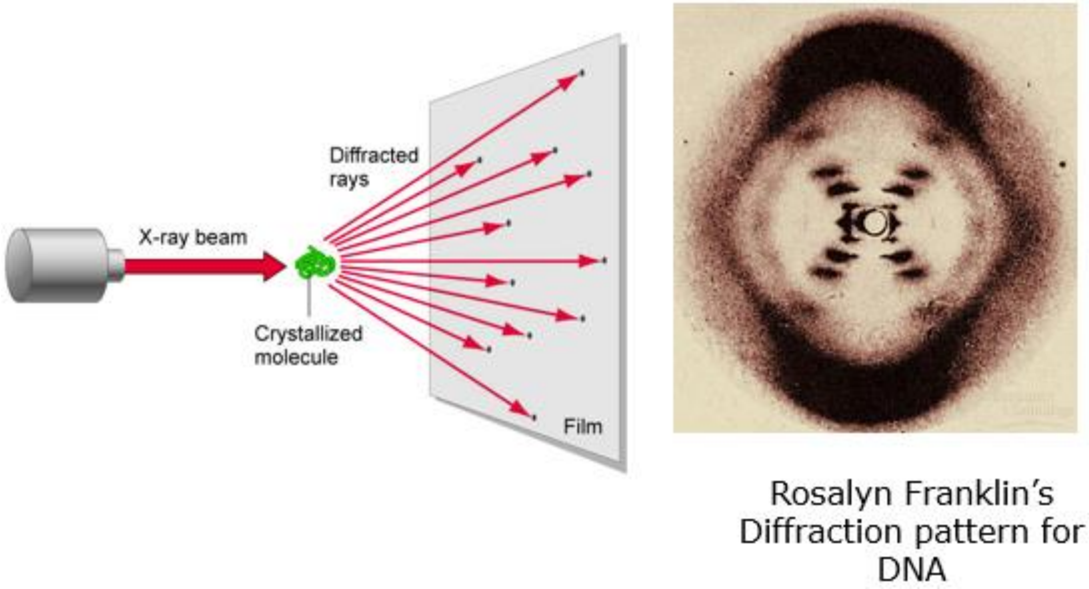


Figure 56 Principle of X-Ray Crystallography

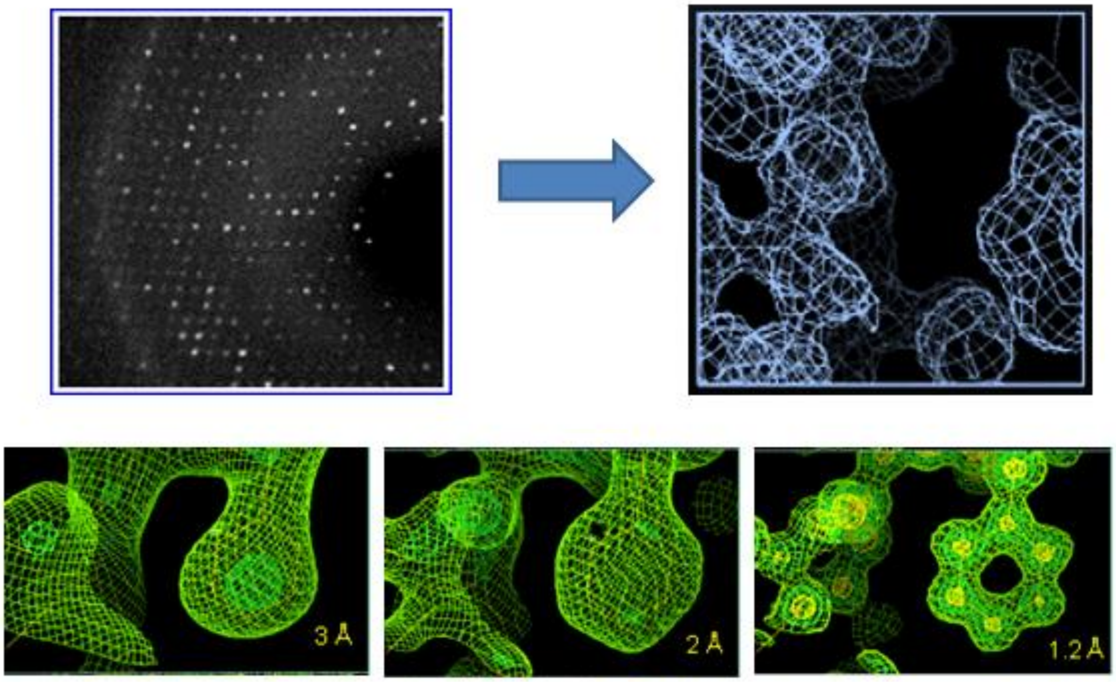


Figure 57 from Diffraction Patterns to Atomic Positions

Upon establishing the atomic positions and distances, we can then check for possible interaction between the different atoms. Atomic distances can help us classify interaction types e.g. hydrogen bonds, electrostatic & polar.

Module 035: Alpha Helices - I

Atomic distances can tell us about their existential interactions. Different types of interactions may occur between atoms. E.g. Hydrogen Bonds, Polar etc. If two atoms are participating in a covalent bond, their distance is ~0.96A. In case of hydrogen bond formation between atoms, the inter-atomic distance is ~1.97A. X-Ray data should have a minimum of 1.97A resolution.

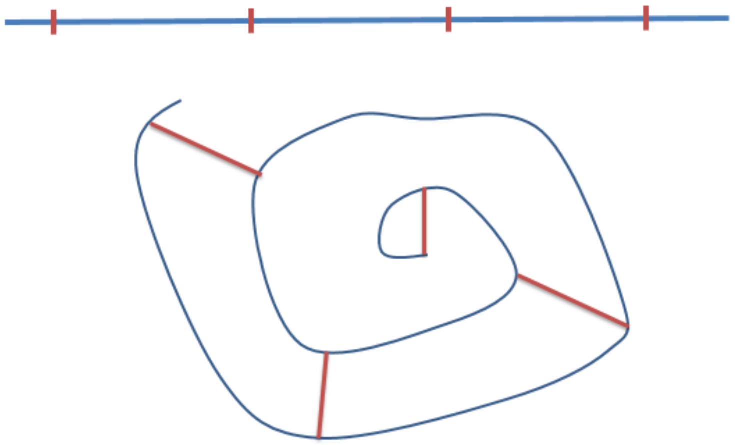


Figure 58 Hydrogen Bonds to Fold an Amino Acid Chain

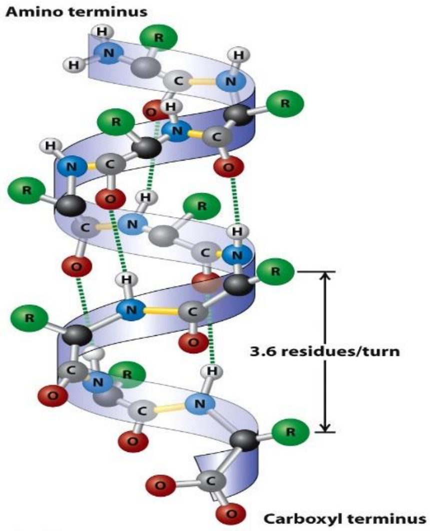


Figure 3-4
Molecular Cell Biology, Sixth Edition
© 2008 W. H. Freeman and Company

X-Ray Crystallography data shows that Hydrogen atoms of N-Term may come together with Oxygen atoms of C-term amino acid at 4th neighboring position. Their atomic distance is ~1.9A and hence are considered to be in a hydrogen bonds.

Module 036: Alpha Helices – II

X-Ray Crystallography of protein shows that Hydrogen atoms of N-Term come together with Oxygen atoms of C-term amino acid at 4th neighboring position to make Hydrogen bonds.

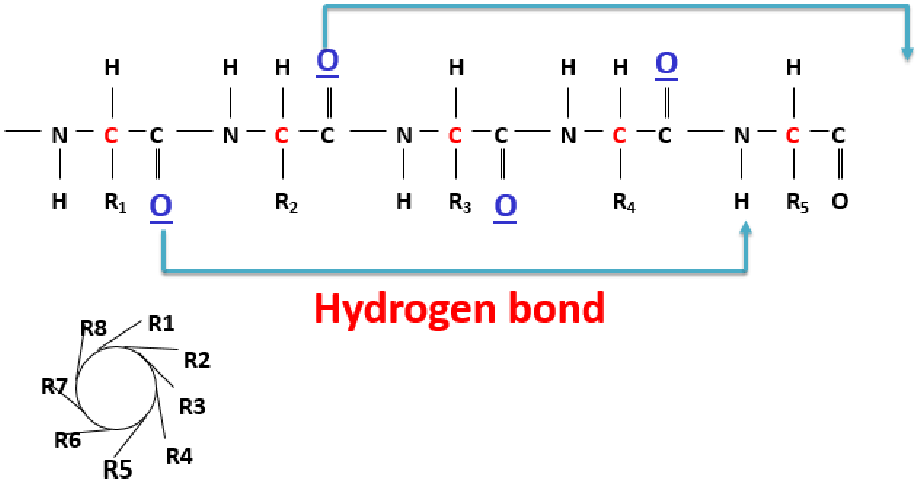


Figure 59 Forming Alpha Helix

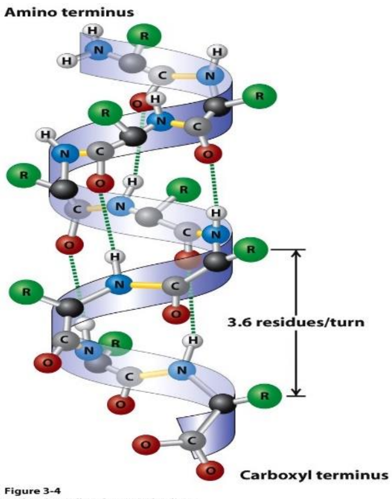


Figure 3-4
Molecular Cell Biology, Sixth Edition
© 2008 W. H. Freeman and Company

Every Oxygen bound to 4th neighboring Amino Group' Hydrogen.

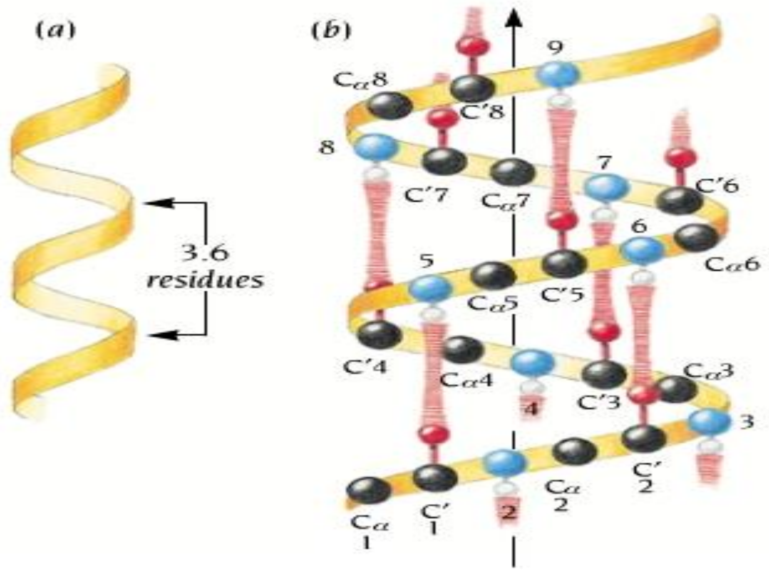
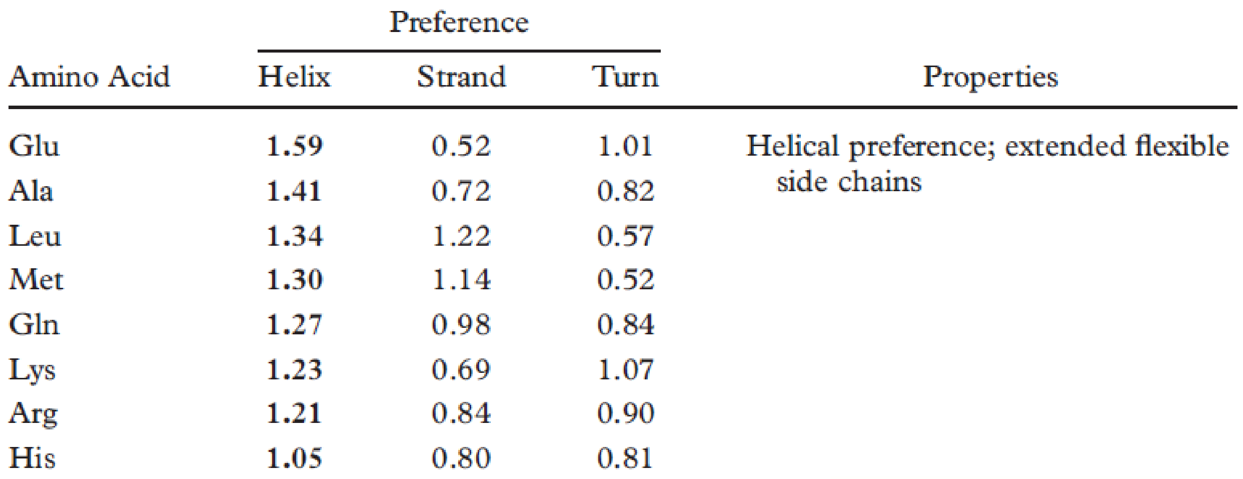


Figure 60 Carbons (Black) & Nitrogen's (Blue): 1-5, 2-6, 3-7…



	Preference			
Amino Acid	Helix	Strand	Turn	Properties
Glu	**1.59**	0.52	1.01	Helical preference; extended flexible side chains
Ala	**1.41**	0.72	0.82	
Leu	**1.34**	1.22	0.57	
Met	**1.30**	1.14	0.52	
Gln	**1.27**	0.98	0.84	
Lys	**1.23**	0.69	1.07	
Arg	**1.21**	0.84	0.90	
His	**1.05**	0.80	0.81	

Figure 61 Preference of Amino Acids for making Alpha Helices

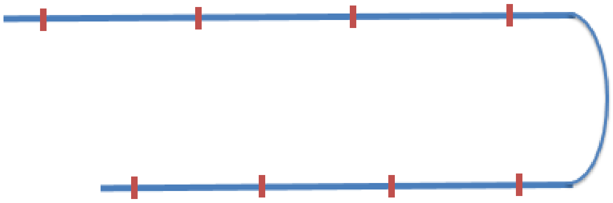
Helix Formers

From 20 amino acids, anyone can be present in the backbone. Is there a variable preference in amino acids to form helix? Yes, "Helix Formers" are generally hydrophobic amino acids (M, A, L…). Alpha Helices are formed by hydrogen bonding (O-H) between C_i and N_{i+4} atoms in the protein backbone.

Module 037: Beta Sheets - I

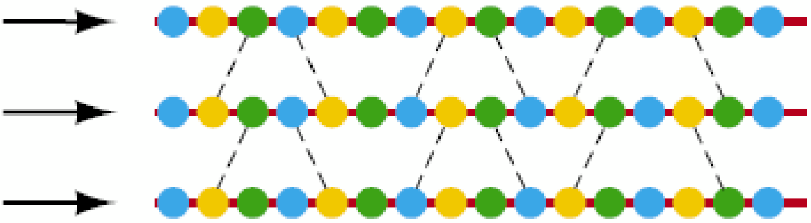
Alpha Helices are formed by hydrogen bonding (O-H) between C_i and N_{i+4} atoms in the protein backbone. Beta Sheets are another common secondary structure. They are constituted by several Beta Strands which come together. 5 to 10 resides are needed to make a Beta Strand, typically.

Hydrogen Bonds to make in Beta Strands

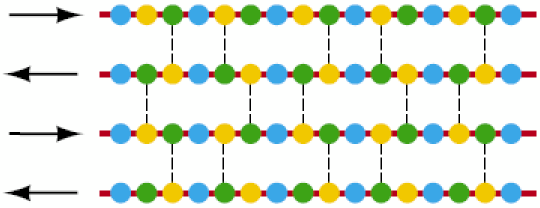


The Beta Sheet is made up of several Beta Strands

C-Alpha atoms and the CO and NH groups are shown in blue, yellow, and green, respectively.



This is called a parallel beta sheet.



This is called an anti-parallel beta sheet.

Beta Sheets are another secondary structure that can be formed as a result of hydrogen bonding between the protein back bones. Some amino acids have a preference for making Beta Sheets.

Module 038: Beta Sheets - II

Beta strands can make hydrogen bonds with each other and organize as beta sheets.

Beta Sheets have different Properties:

- Beta Strand
- Beta Sheet
- Beta Barrel
- Beta Sandwiches

Beta Barrels

Beta Barrel is made of a single beta sheet that twists and coils upon itself. The first strand in the beta sheet makes a hydrogen bonds with the last strand. A beta barrel is a large beta-sheet that twists and coils to form a closed structure in which the first strand is hydrogen bonded to the last. Beta-strands in beta-barrels are typically arranged in an antiparallel fashion.

https://en.wikipedia.org/wiki/**Beta_barrel**

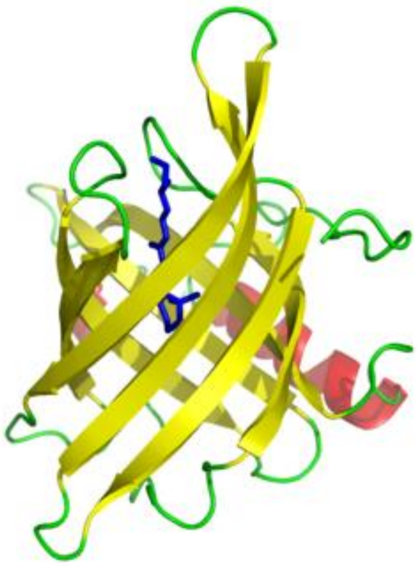


Figure 62 beta barrel

Beta Sandwiches

Beta Sandwiches are made of two beta sheets which are usually twisted and packed so their strands are aligned.

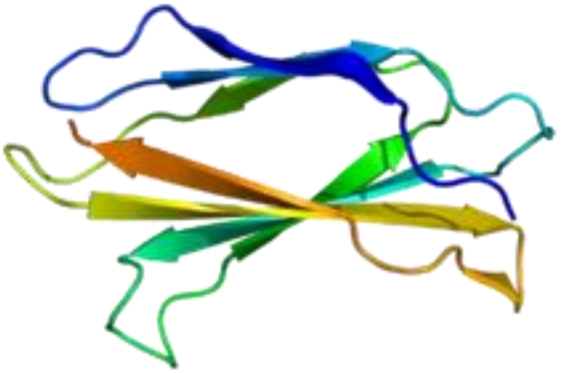
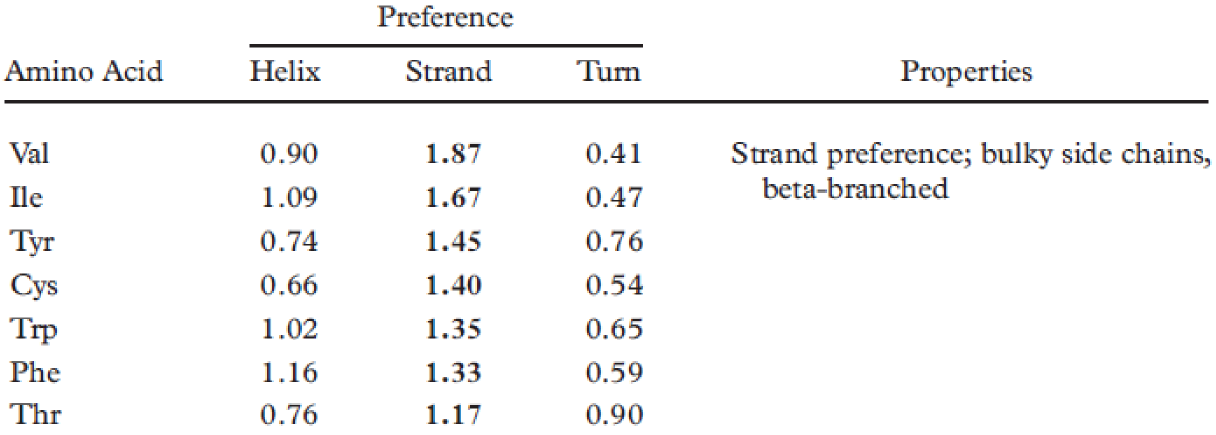


Figure 63 Illustration of the β-sandwich from Tenascin C (PDB entry: 1TEN).



	Preference			
Amino Acid	Helix	Strand	Turn	Properties
Val	0.90	**1.87**	0.41	Strand preference; bulky side chains, beta-branched
Ile	1.09	**1.67**	0.47	
Tyr	0.74	**1.45**	0.76	
Cys	0.66	**1.40**	0.54	
Trp	1.02	**1.35**	0.65	
Phe	1.16	**1.33**	0.59	
Thr	0.76	**1.17**	0.90	

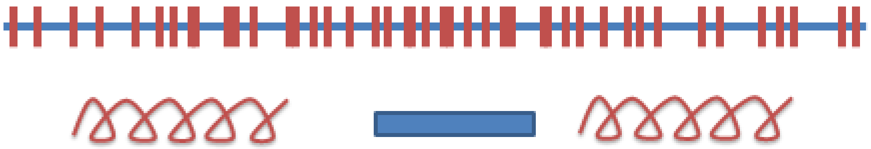
Figure 64 Preference of Amino Acids for making Beta strands

Beta Sheets are formed by H bonds between of 5–10 consecutive amino acids in one portion of the backbone with another 5–10 farther down the backbone. Beta strands may be adjacent (with a loop in between) or far with other structures in between.

Module 039: LOOPS-I

Alpha Helices and Beta Sheets are secondary structures formed as a result of hydrogen bonding in between protein backbone atoms.

Protein Backbone and Secondary Structures



Loops are formed by amino acids present in the middle of the Alpha Helices and Beta Sheets in a protein backbone.

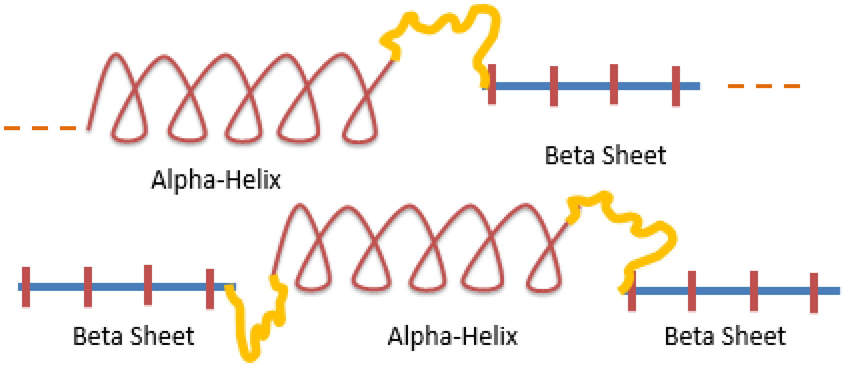
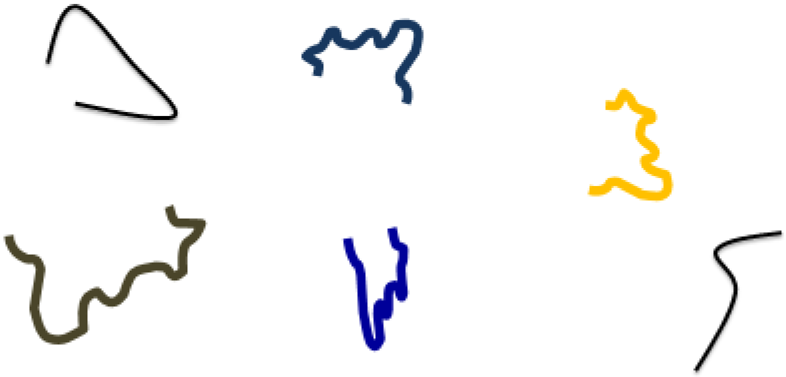


Figure 65 Joining Alpha Helices and Beta Sheets in a Protein Backbone

Variability in length and conformation allows loops to join Alpha Helices and Beta Sheets in a variety of ways. Loops are variable in length and 3-D conformations.



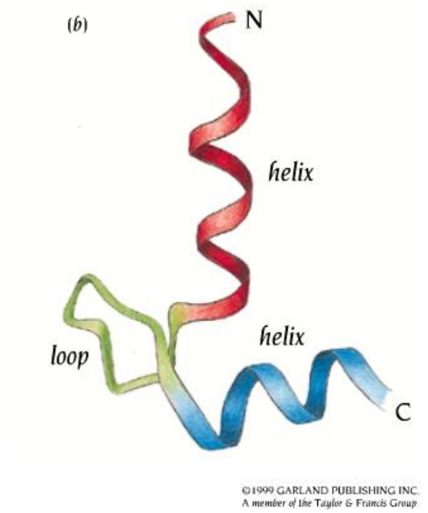
Characteristics

- Loops are mostly located on the surface of protein structure
- Mutate in sequence at a much faster rate than Alpha Helices and Beta Sheets
- Loops are flexible and can adopt multiple conformations

Loops dictate the overall structure of protein as they couple Alpha helices and beta sheets

Module 040: LOOPS-II

Loops dictate the overall structure of protein as they couple Alpha helices and Beta sheets. Loops are flexible and have variable lengths so as to successfully bridge between secondary structures.



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

Figure 66 Loops in 3D Conformation

Loop Properties

- ➢ Loops are mostly comprised of charged and polar amino acids
- ➢ Loops frequently participate as components of active sites

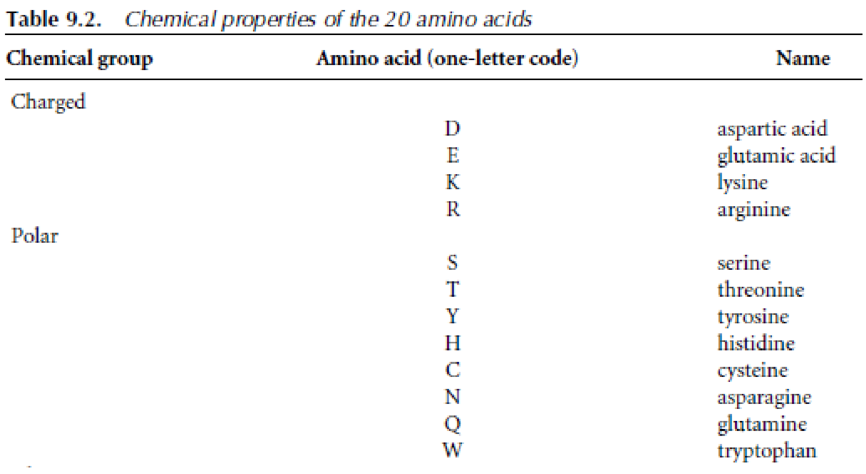


Table 9.2. *Chemical properties of the 20 amino acids*

Chemical group	Amino acid (one-letter code)	Name
Charged		
	D	aspartic acid
	E	glutamic acid
	K	lysine
	R	arginine
Polar		
	S	serine
	T	threonine
	Y	tyrosine
	H	histidine
	C	cysteine
	N	asparagine
	Q	glutamine
	W	tryptophan

Figure 67 Preference of Amino Acids for making Loops

Types of Loops

- Hairpin loops are two amino acids long and join anti-parallel Beta strands
- Other Loops may be 3 to 4 amino acids long
- Loops fall into various families

Loops are the third type of secondary structure after Alpha helices and Beta sheets. Loops are unique in that they are flexible and variable length. Loops constitute active sites.

Module 041: COILS

Alpha helices and beta sheets are the regular secondary structures. Loops are flexible secondary structures &connect alpha helices and beta sheets. Coils are another secondary structure. Coils are unstructured and unlike loops. Essentially, a secondary structure which is not a helix, sheet or loop is a coil.

Functional Aspects of Coils

- Coils are apparently disordered regions
- They are oriented randomly while being bonded to adjacent amino acids
- However, coils also appear to play important functional roles

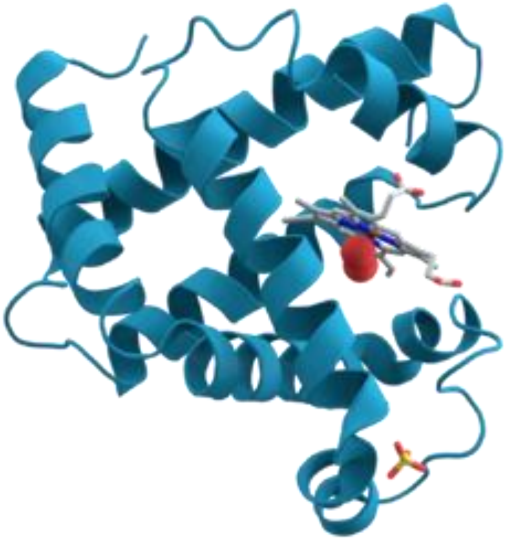


Figure 68 Coils in Myoglobin

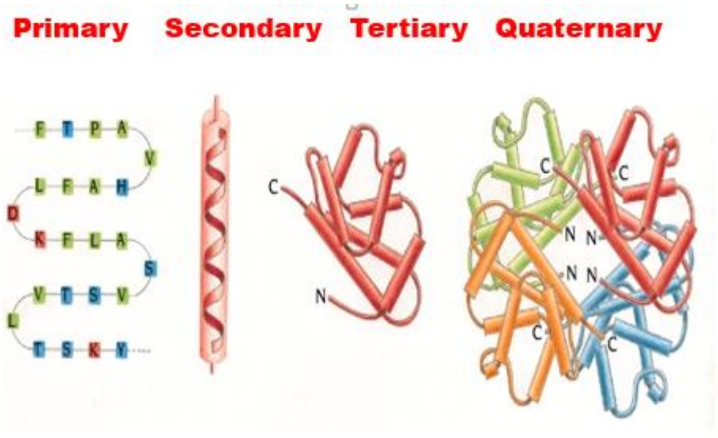
Coils are those secondary structure formed by the protein backbone which are neither helices, sheets nor loops. In fact, coils do not have a consistent classifiable structure. Hence, coils are random structure and random length.

Module 042: Structure Classification - I

Proteins have primary, secondary, tertiary & quaternary structures. Each level of protein structure organization is known to impart specific characteristics to the protein.

Review of the 4 structure levels

- Primary Structures
- Secondary Structures
- Tertiary Structures
- Quaternary Structures



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

Structural artifacts tend to be more conserved as compared to their sequences. Therefore, it may be useful to look at the secondary/tertiary structures for conservation study.

Classification

- The evolution of protein structures and their hierarchy is not systematized
- Hence, we need to classify the function of protein by examining their secondary and tertiary structures

Motifs (Non-functional Combinations of 2’ structures)

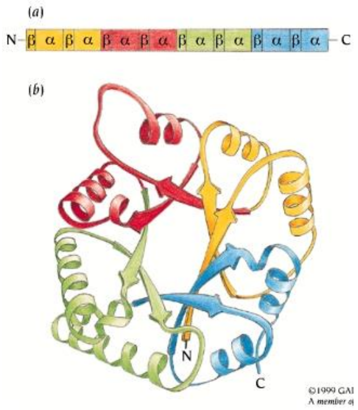


Figure 69 Domain (Functionally Complete)

Domains are semi-independent functional structures in a protein. Have a stable structure. Over ~40 residues. Protein may contain multiple domains.

Module 043: Structure Classification - II

Domains are semi-independent functional structures in a protein. Protein may contain multiple domains. Hence, we can try to classify proteins by their domains. Locally Compact – Domains interact (H-bonds) more internally than externally. Domains have a hydrophobic core. Domains are contiguous (min. chain breaks).

Domains have a minimal contact with rest of the peptide. Solvent area in contact with each domain should not vary significantly upon separating two separate domains.

Types of Domains

- Alpha Domains
- Beta Domains
- Alpha/Beta Domains
- Alpha + Beta Domains
- Alpha & Beta Multi-Domains
- Membrane & cell-surface proteins

So, by looking at proteins, we can list the domains present in each protein. Once domains in each protein are listed, we can classify whole proteins into various types and classes.

Module 044: Examples of Protein Domains

There are many domains for protein structure prediction.

- Alpha Domains
- Beta Domains
- Alpha/Beta Domains
- Alpha + Beta Domains
- Alpha & Beta Multi-Domains
- Membrane & cell-surface proteins

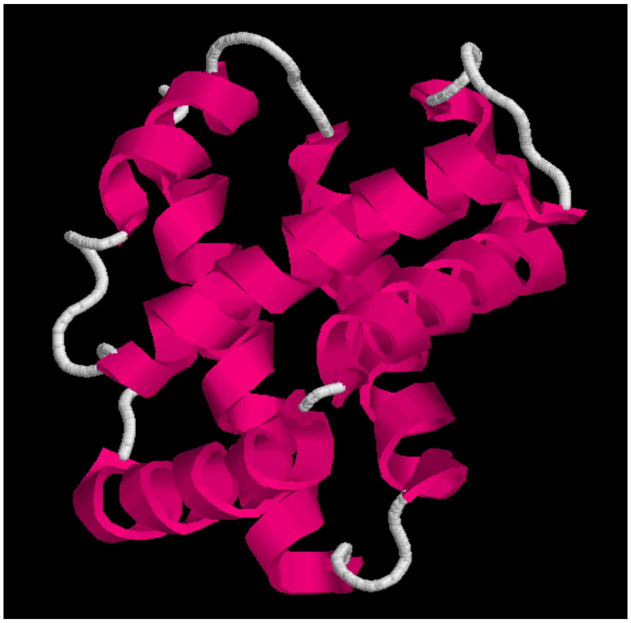


Figure 70 Alpha Domain: Hemoglobin (1bab)

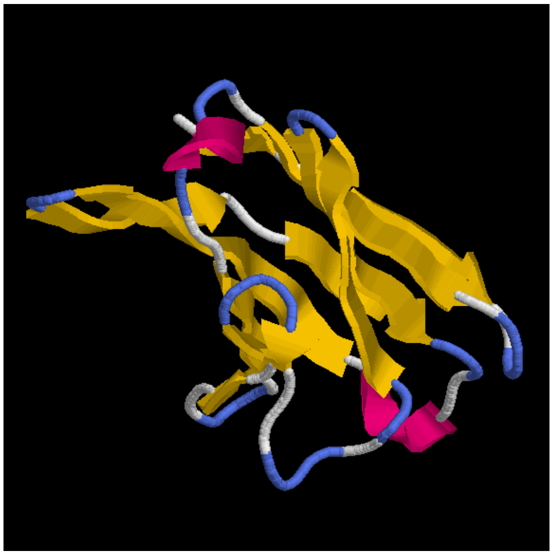


Figure 71 Immunoglobulin (8fab)

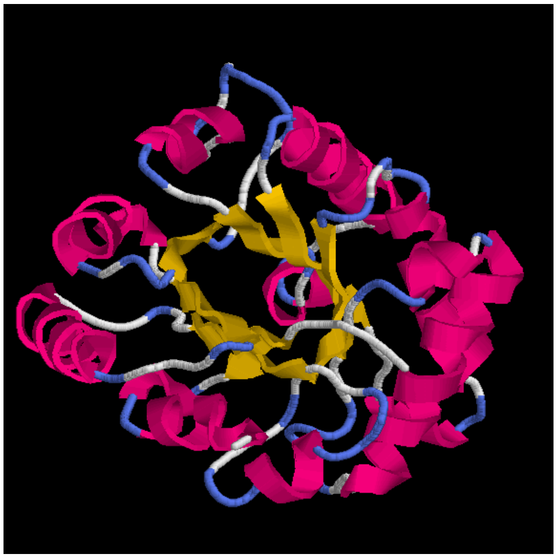


Figure 72 Alpha / Beta: Triosephosphate isomerase (1hti)

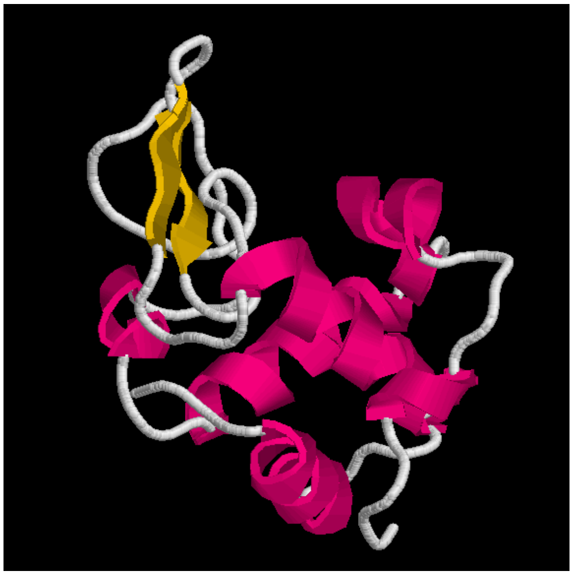


Figure 73 Alpha + Beta: Lysozyme (1jsf)

Various types of domain architectures exist in proteins. Such architectures can be classified into general structural classes. Databases can be made from classes.

Module 045: CATH Classification

Domains can be classified into structural classes. Classes can be further classified into Architecture and Topologies. Let’s see how it is done in CATH.

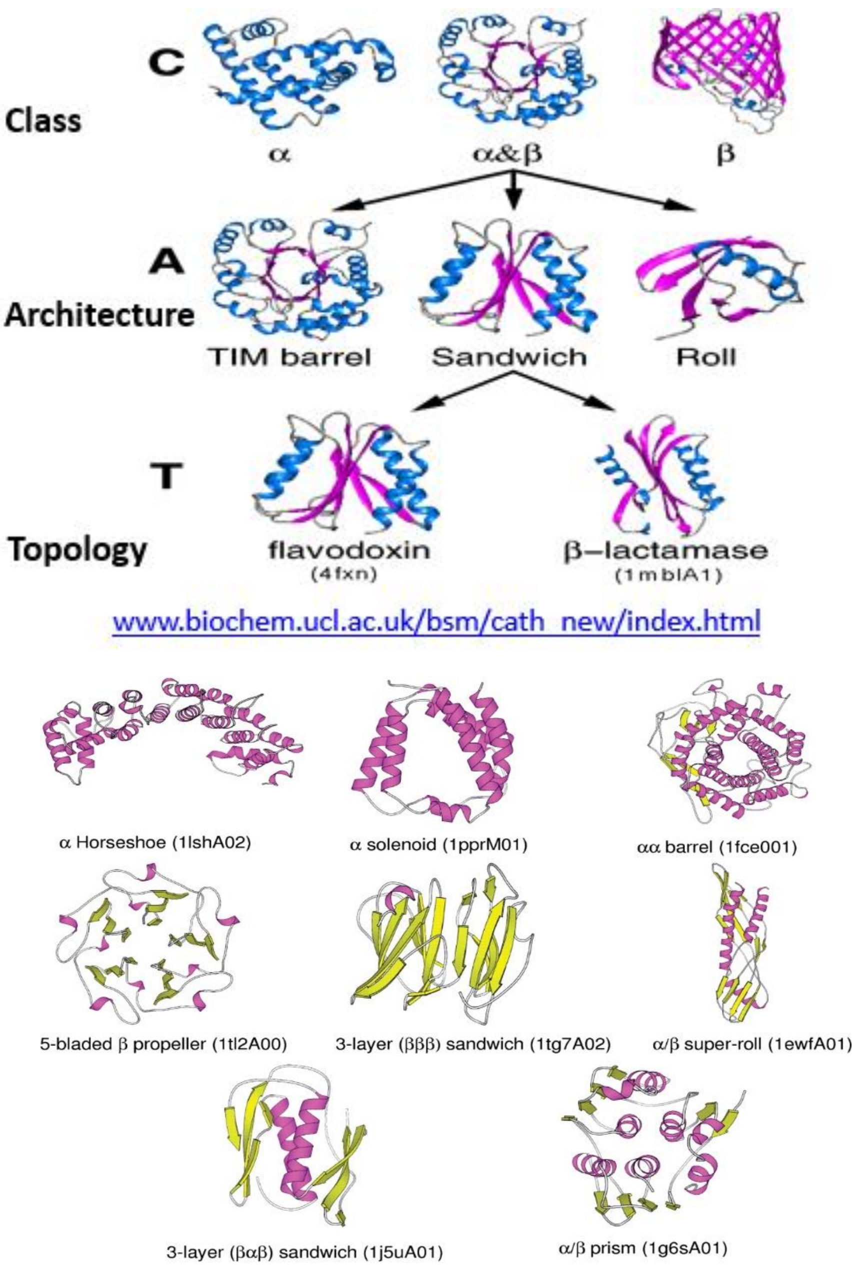


Figure 74 Structural Classes

Class

- Similar secondary structure content
- All α, all β, alternating α/β etc.

Architecture

- Also called FOLD
- Major structural similarity
- SSE's in similar arrangement

Topology

- Super Family
- Probable common ancestry
- Family membership

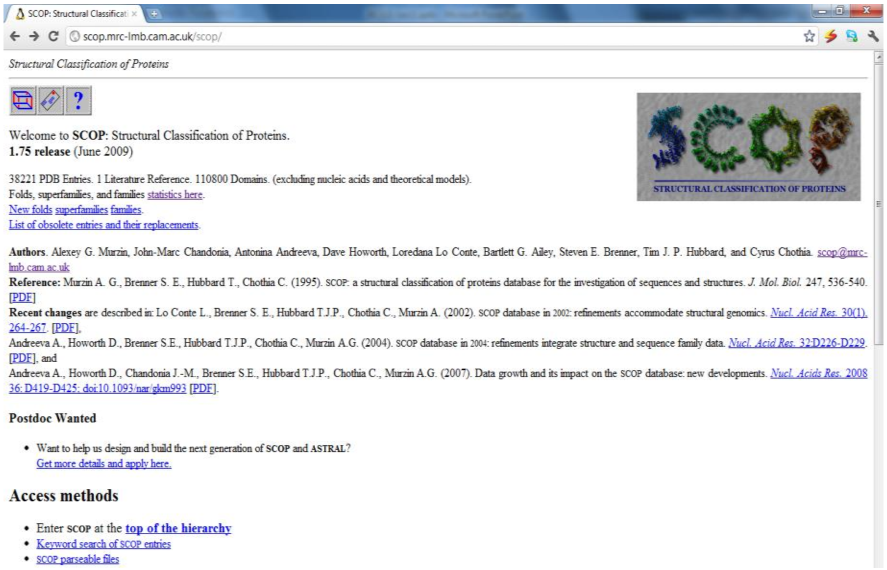
Homology

- Same Family
- Clear evolutionary relationship
- Pairwise sequence similarity > 30%

CATH classifies proteins by their structural similarity. It also considers the internal organization of the structural components in proteins.

Module 046: Classification Databases

Proteins are classified into various structural classes. CATH is one such system in which proteins are organized into classes, architecture, topology and homology.


SCOP: Structural Classificat
scop.mrc-lmb.cam.ac.uk/scop/
STRUCTURAL CLASSIFICATION OF PROTEINS

Structural Classification of Proteins

Welcome to **SCOP**: Structural Classification of Proteins.
1.75 release (June 2009)

38221 PDB Entries. 1 Literature Reference. 110800 Domains. (excluding nucleic acids and theoretical models).
Folds, superfamilies, and families statistics here.
New folds superfamilies families.
List of obsolete entries and their replacements.

Authors. Alexey G. Murzin, John-Marc Chandonia, Antonina Andreeva, Dave Howorth, Loredana Lo Conte, Bartlett G. Ailey, Steven E. Brenner, Tim J. P. Hubbard, and Cyrus Chothia. scop@mrc-lmb.cam.ac.uk
Reference: Murzin A. G., Brenner S. E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536-540. [PDF]
Recent changes are described in: Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. (2002). SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.* 30(1), 264-267. [PDF],
Andreeva A., Howorth D., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2004). SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acid Res.* 32:D226-D229. [PDF], and
Andreeva A., Howorth D., Chandonia J.-M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G. (2007). Data growth and its impact on the SCOP database: new developments. *Nucl. Acids Res.* 2008 36: D419-D425; doi:10.1093/nar/gkm993 [PDF].

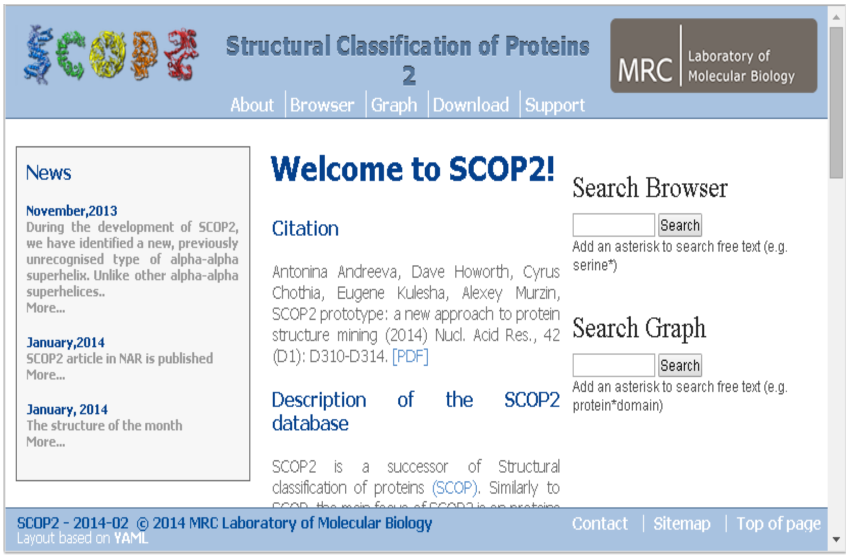
Postdoc Wanted

- Want to help us design and build the next generation of **SCOP** and **ASTRAL**?
  Get more details and apply here.

Access methods

- Enter SCOP at the **top of the hierarchy**
- Keyword search of SCOP entries
- SCOP parseable files

http://scop.mrc-lmb.cam.ac.uk/scop/


Structural Classification of Proteins 2
About | Browser | Graph | Download | Support
MRC | Laboratory of Molecular Biology

News

November,2013
During the development of SCOP2, we have identified a new, previously unrecognised type of alpha-alpha superhelix. Unlike other alpha-alpha superhelices..
More...

January,2014
SCOP2 article in NAR is published
More...

January, 2014
The structure of the month
More...

Welcome to SCOP2!

Citation

Antonina Andreeva, Dave Howorth, Cyrus Chothia, Eugene Kulesha, Alexey Murzin, SCOP2 prototype: a new approach to protein structure mining (2014) Nucl. Acid Res., 42 (D1): D310-D314. [PDF]

Description of the SCOP2 database

SCOP2 is a successor of Structural classification of proteins (SCOP). Similarly to SCOP, the main focus of SCOP2 is on proteins

Search Browser

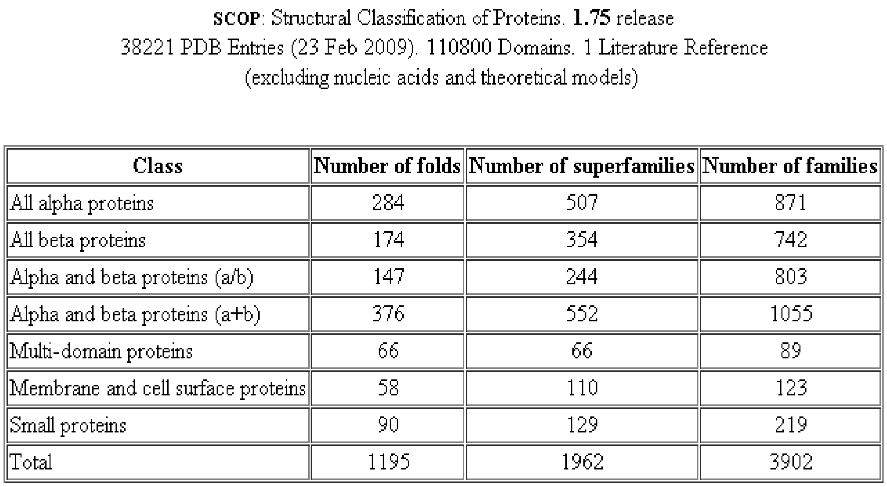
Search
Add an asterisk to search free text (e.g. serine*)

Search Graph

Search
Add an asterisk to search free text (e.g. protein*domain)

SCOP2 - 2014-02 © 2014 MRC Laboratory of Molecular Biology
Layout based on YAML

Contact | Sitemap | Top of page



SCOP: Structural Classification of Proteins. **1.75** release
38221 PDB Entries (23 Feb 2009). 110800 Domains. 1 Literature Reference
(excluding nucleic acids and theoretical models)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	284	507	871
All beta proteins	174	354	742
Alpha and beta proteins (a/b)	147	244	803
Alpha and beta proteins (a+b)	376	552	1055
Multi-domain proteins	66	66	89
Membrane and cell surface proteins	58	110	123
Small proteins	90	129	219
Total	1195	1962	3902

Figure 75 SCOP Classification Statistics

http://scop.mrc-lmb.cam.ac.uk/scop/count.html

FSSP - Family of Structurally Similar Proteins, based on the DALI algorithm. Pclass - Protein Classification, based on the LOCK and 3Dsearch algorithms.

Module 047: Algorithms for Structure Classification

Several algorithms exist for classifying protein structures.

Intra-Molecular Distance Algorithms.

- Proteins are considered as rigid bodies.
- They are placed in a 3D Cartesian coordinate system.
- Structural alignment in 3D.
- E.g. VAST, LOCK

Inter-Molecular Distance Algorithms

- Proteins are considered as rigid bodies.
- They are placed in 2D.
- Structural alignment using internal distances and angles.

The basic idea is to capture internal geometry of protein structures. E.g. DALI, and SSAP.

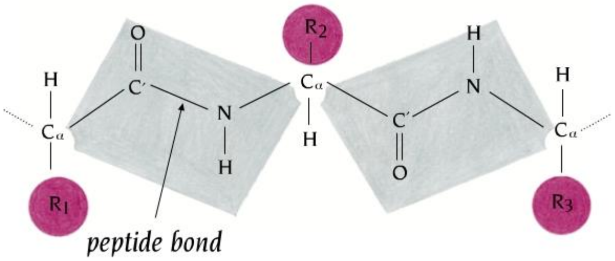
Such algorithms are also very useful to compare whole protein structures. They can help determine evolutionary relationship. Also, functional similarity can be estimated.

Module 048: Protein Structure Comparison

Proteins are assembled into primary (1'), secondary (2'), tertiary (3') and quaternary (4') structures. Protein sequence is less conserved than its structure. Protein structure determines function. Since protein structure dictates function, comparing two structures can help us evaluate if the proteins do the same or similar function.

Comparing Whole Protein Structures

Proteins contain multiple structural subunits e.g. secondary structures, motifs and domains. Structures of all such subunits are to be considered as one and compared. We know that domains are functionally independent components of the protein structure. Proteins may have multiple domains. So for two different proteins, sharing the same domain, we may want to compare only a portion of the overall structure i.e. a domain. For comparing the complete or partial protein structures, the position of Alpha Carbon atoms can be used. The (x, y, z) positions of Alpha Carbon atoms can be obtained from the PDB.



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

Figure 74 C-Alpha atoms in backbone

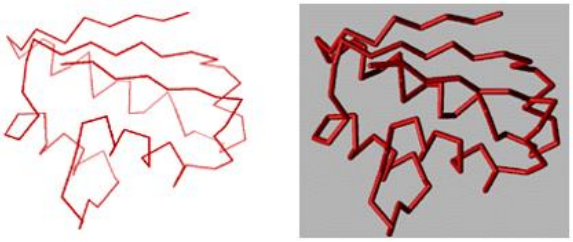


Figure 76 Tracing and Visualizing C-Alpha Backbone

http://www.danforthcenter.org/smith/MolView/Over/overview.html

PDB coordinates of Alpha Carbons in the protein back bone can be used for comparison. In this way, whole protein structure or domains etc. can be compared.

Module 049: Strategies for Structure Comparison - I

PDB coordinates of Alpha Carbons in the protein back bone can be used for comparison. Thus, two whole protein structures or domains within each structure can be compared.

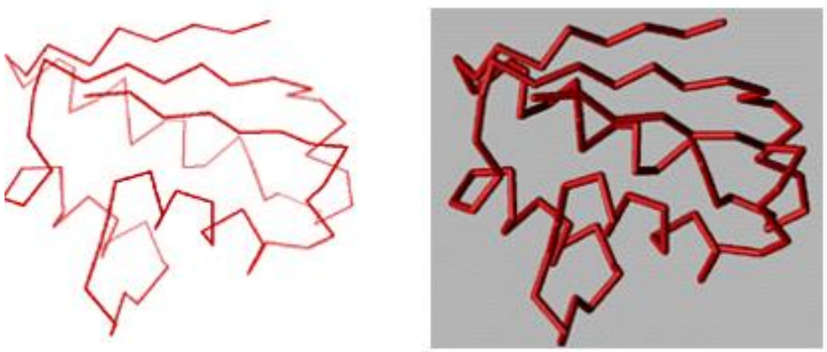


Figure 77 Tracing and Visualizing C-Alpha Backbone

http://www.danforthcenter.org/smith/MolView/Over/overview.html

Strategy # 1 – Whole Protein Structure Comparison by Intermolecular distances

- Two protein sequences are pair-wise aligned with each other
- Corresponding Alpha Carbons are identified
- Coordinates of corresponding Alpha Carbons are retrieved from PDB
- Their individual differences calculated
- Root Mean Square Distance is computed to assess the similarity

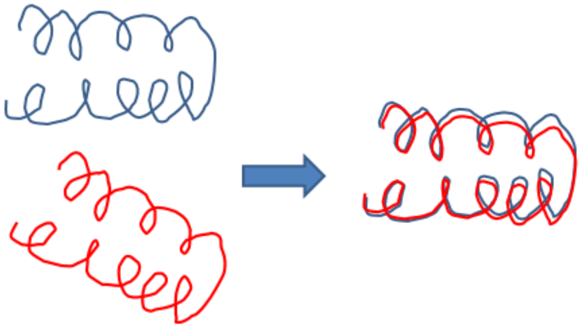
Whole protein structures can be compared by calculating the root mean squared difference (RMSD) between their Alpha Carbons positions. The lower the RMSD, the similar are the proteins.

Module 050: Strategies for Structure Comparison - II

Full protein structures can be compared and ranked by the overall differences in positions between their Alpha Carbons. But proteins are 3D and in various conformations.

Full Protein Comparison

(Translation -> Rotation -> RMSD)



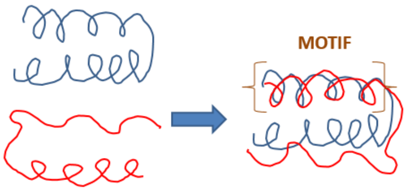
Calculating RMSD – An Example

$$RMS(A,B) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d(a_i, b_i)^2}$$

where $d(a_i - b_i)^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2$

Domain or Motif Comparison

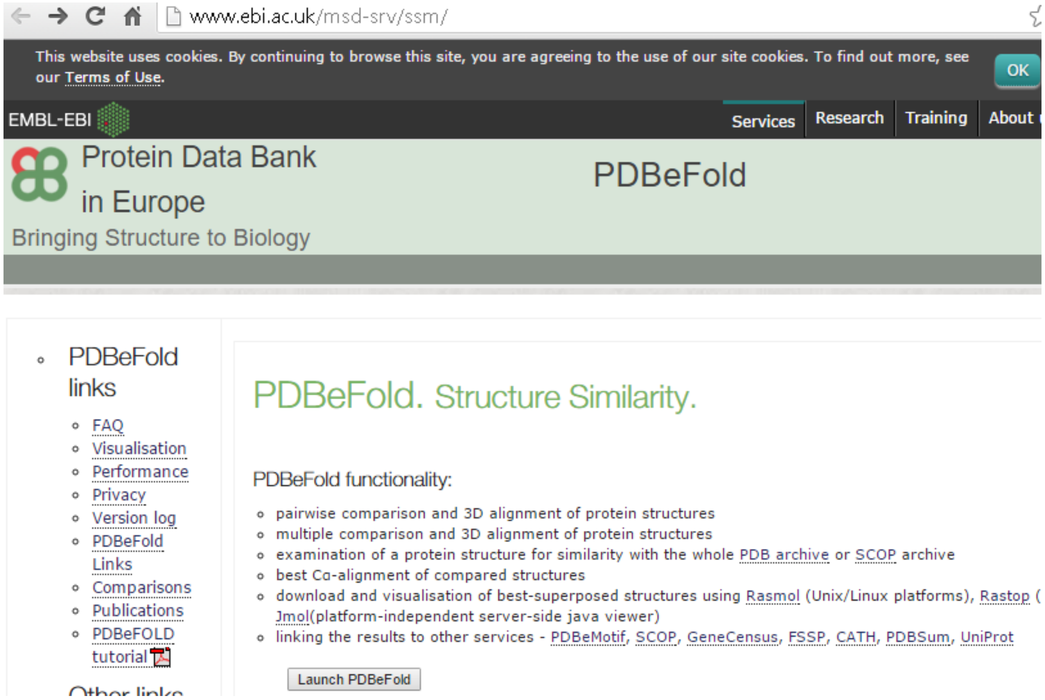
(Region Selection ->Translation -> Rotation -> RMSD)



Motifs, Domains and Full Proteins can be compared by using the rigid body super-positioning. Depending on the RMSD, proteins, their motifs and domains can be selectively compared.

Module 051: Online Resources for Structure Comparison

Multiple types of comparison can be performed between Proteins, Motifs, and Domains by rigid body super-positioning. RMSD tells us about the quality of the matches.



www.ebi.ac.uk/msd-srv/ssm/

This website uses cookies. By continuing to browse this site, you are agreeing to the use of our site cookies. To find out more, see our Terms of Use.

PDBeFold links

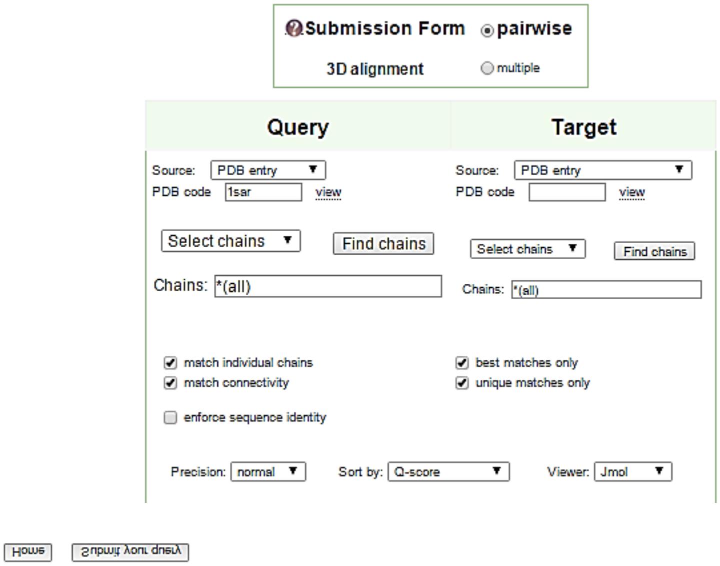
- FAQ
- Visualisation
- Performance
- Privacy
- Version log
- PDBeFold Links
- Comparisons
- Publications
- PDBeFOLD tutorial

Other links

PDBeFold. Structure Similarity.

PDBeFold functionality:

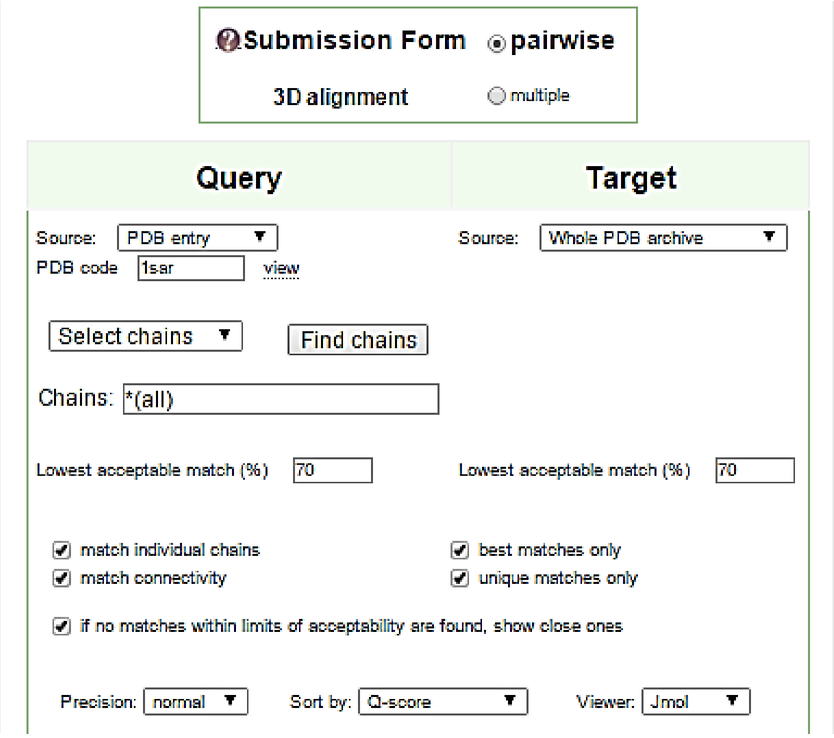
- pairwise comparison and 3D alignment of protein structures
- multiple comparison and 3D alignment of protein structures
- examination of a protein structure for similarity with the whole PDB archive or SCOP archive
- best Cα-alignment of compared structures
- download and visualisation of best-superposed structures using Rasmol (Unix/Linux platforms), Rastop (Jmol(platform-independent server-side java viewer)
- linking the results to other services - PDBeMotif, SCOP, GeneCensus, FSSP, CATH, PDBSum, UniProt



Submission Form ● pairwise 3D alignment ○ multiple

Query	Target
Source: PDB entry	Source: PDB entry
PDB code 1sar view	PDB code view
Select chains Find chains	Select chains Find chains
Chains: *(all)	Chains: *(all)
☑ match individual chains	☑ best matches only
☑ match connectivity	☑ unique matches only
□ enforce sequence identity	

Precision: normal Sort by: Q-score Viewer: Jmol



Submission Form ● pairwise

3D alignment ○ multiple

Query	Target
Source: PDB entry ▼	Source: Whole PDB archive ▼
PDB code 1sar view	
Select chains ▼ Find chains	
Chains: *(all)	
Lowest acceptable match (%) 70	Lowest acceptable match (%) 70
☑ match individual chains	☑ best matches only
☑ match connectivity	☑ unique matches only

☑ if no matches within limits of acceptability are found, show close ones

Precision: normal ▼ Sort by: Q-score ▼ Viewer: Jmol ▼

Protein structures can be compared in multiple ways. Till now, we can compare proteins by their motifs, domains and full structures. There are several advanced techniques for this as well.

Module 052: Protein Structure Prediction

Complex protein structures enable proteins to perform complex functions. We know over a million protein sequences but only about 100,000 protein structures. Estimating exact protein structures is very difficult. It's difficult to crystallize proteins. Even if we manage to get protein's X-Ray, to reconstruct the structure is extremely complex.

Since we know so many sequences, they can be used for predicting protein structures. This indeed is possible and helpful.

The Basic Idea

- Amino acids determine the protein structure
- We have a large protein sequence dataset (uniprot)

Hence, we can fold protein sequences and predict their structures

Why predict and why not exact solutions?

A deterministic solution of protein folding is a major unsolved problem in molecular biology. Proteins fold spontaneously or with the help of enzymes or chaperones. To computationally predict protein structures, we need to copy or mimic the natural folding.

To fold we must learn the steps

Step 1: "Collapse"- leading to burial of hydrophobic AA's

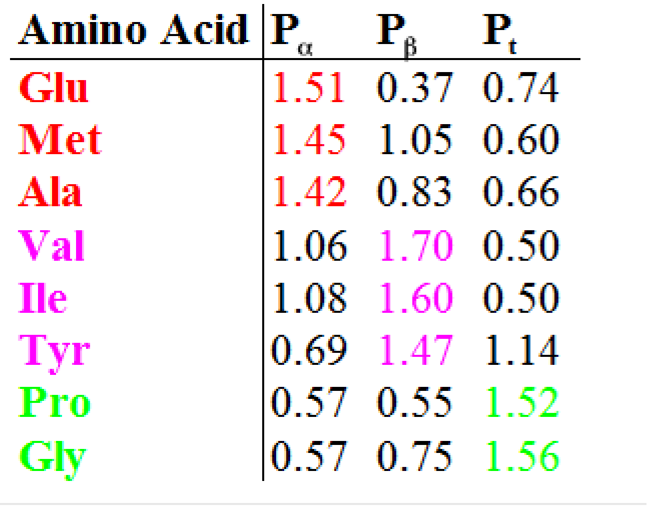
Step 2: Fluid globule - helices & sheets form, but are unorganized

Step 3: Compaction, and rearrangement of 2‘structures

Protein structure prediction involves learning how the amino acids in primary sequence fold. Using this information, upon getting a protein sequence, we can try to predict how it folds

Module 053: Predicting Secondary Structures

By looking at the structures in PDB, we know that Alanine mostly found in Alpha Helices. So if we have several Alanines in the sequence, then we can anticipate that a helix may be formed by them. What if we survey the entire PDB and check the presence of each amino in each type of secondary structure. If we know which amino acid is found in which specific secondary structure, then we can use it for prediction.



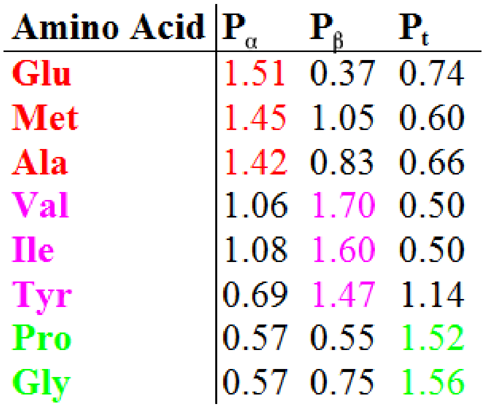
Amino Acid	P_α	P_β	P_t
Glu	1.51	0.37	0.74
Met	1.45	1.05	0.60
Ala	1.42	0.83	0.66
Val	1.06	1.70	0.50
Ile	1.08	1.60	0.50
Tyr	0.69	1.47	1.14
Pro	0.57	0.55	1.52
Gly	0.57	0.75	1.56

Figure 78 Chou & Fasman (1974 & 1978)

Several algorithms have been designed to predict 2' given an amino acid sequence. The first such algorithm was the Chou-Fasman Algorithm. We will see it in the upcoming modules.

Module 054: Introduction to Chou Fasman Algorithm

3D Structure of proteins is determined by their Amino Acid sequence. Note that we only know 100,000 3D protein structures, but 10 times more sequences. For those proteins whose structure is already known, can we evaluate their amino acid sequence?



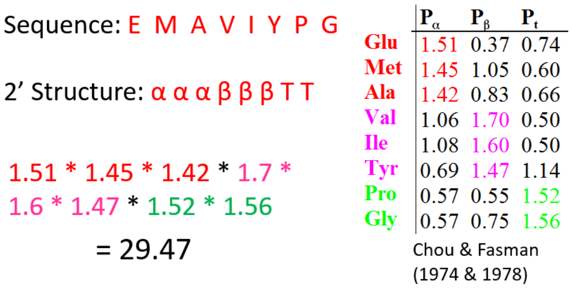
Amino Acid	P_α	P_β	P_t
Glu	1.51	0.37	0.74
Met	1.45	1.05	0.60
Ala	1.42	0.83	0.66
Val	1.06	1.70	0.50
Ile	1.08	1.60	0.50
Tyr	0.69	1.47	1.14
Pro	0.57	0.55	1.52
Gly	0.57	0.75	1.56

Chou & Fasman (1974 & 1978)

Figure 79 Propensity Table

Predicting the 2' structures

Now, let's consider that if we are given an amino acid sequence, we can simply look up the propensity table and assign the tentative secondary structure.



Sequence: E M A V I Y P G

2' Structure: α α α β β β T T

1.51 * 1.45 * 1.42 * 1.7 * 1.6 * 1.47 * 1.52 * 1.56

= 29.47

	P_α	P_β	P_t
Glu	1.51	0.37	0.74
Met	1.45	1.05	0.60
Ala	1.42	0.83	0.66
Val	1.06	1.70	0.50
Ile	1.08	1.60	0.50
Tyr	0.69	1.47	1.14
Pro	0.57	0.55	1.52
Gly	0.57	0.75	1.56

Chou & Fasman (1974 & 1978)

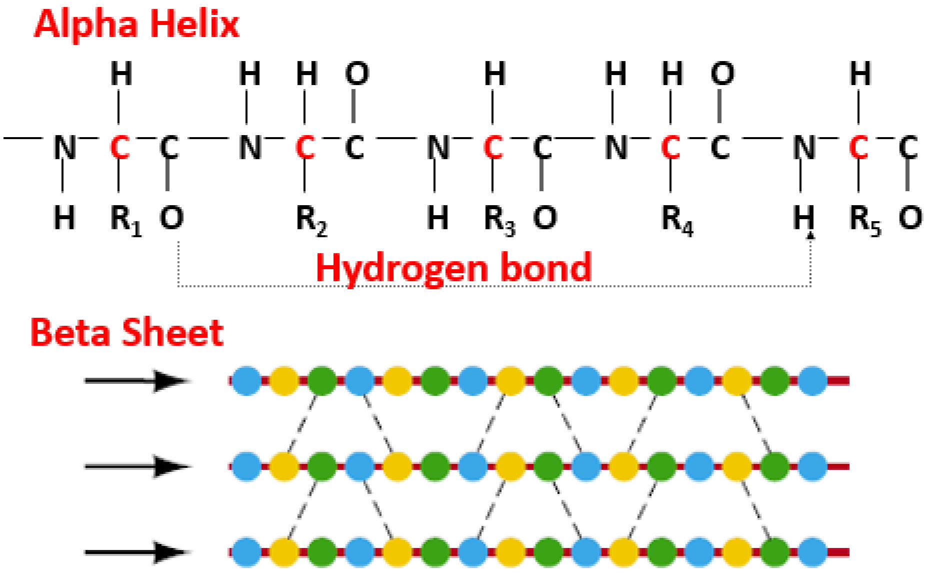
The propensity for forming "ααα β β β T T" is = 29.47

Given an amino acid sequence, look up the propensity table for each amino acid's propensity for various 2' structures. Product of these propensity values will give you the overall propensity for formation of each 2' structure.

Module 055: 2’ Structures in Chou Fasman Algorithm

For a primary sequence, and a tentative 2’ structure, propensity table can help us compute the overall propensity. Product of propensity values is computed for overall propensity for each 2’ structure. An important point to note here is that 2’ structures are formed due to hydrogen bonding between amino acids.

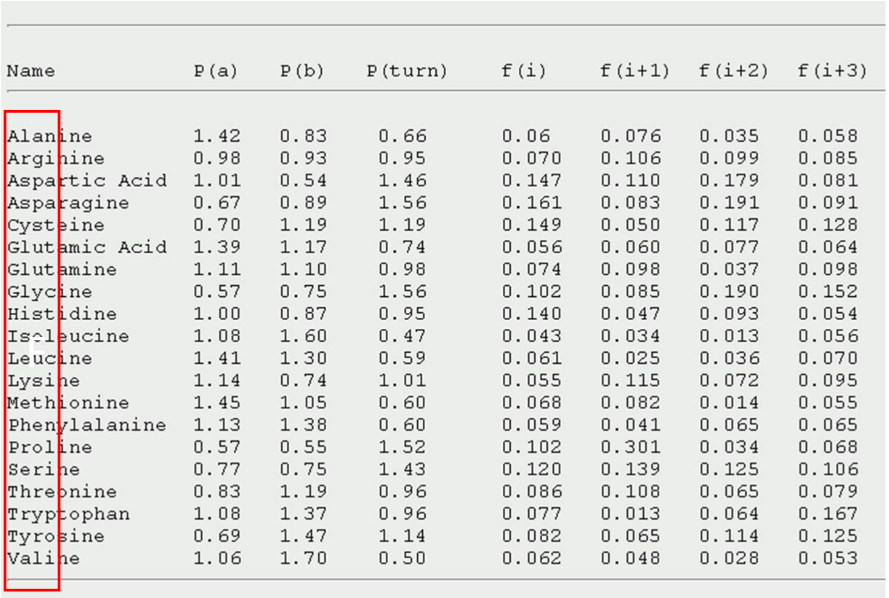
So, we need to consider the neighboring amino acids as well.



You only need to compute propensities for a small number 2’ structures. The highest net propensity will be the most probably secondary structure that will be formed.

Module 056: Chou Fasman Algorithm - I

Only a small number of combinations of secondary structures are possible due to their individual properties. Such as 4 amino acids are needed to start an Alpha Helix and 5 amino acids for Beta Sheet. Note that besides the alpha helix and beta sheets, LOOPS are another secondary structure. Loops are small ~ 3-4 amino acids.



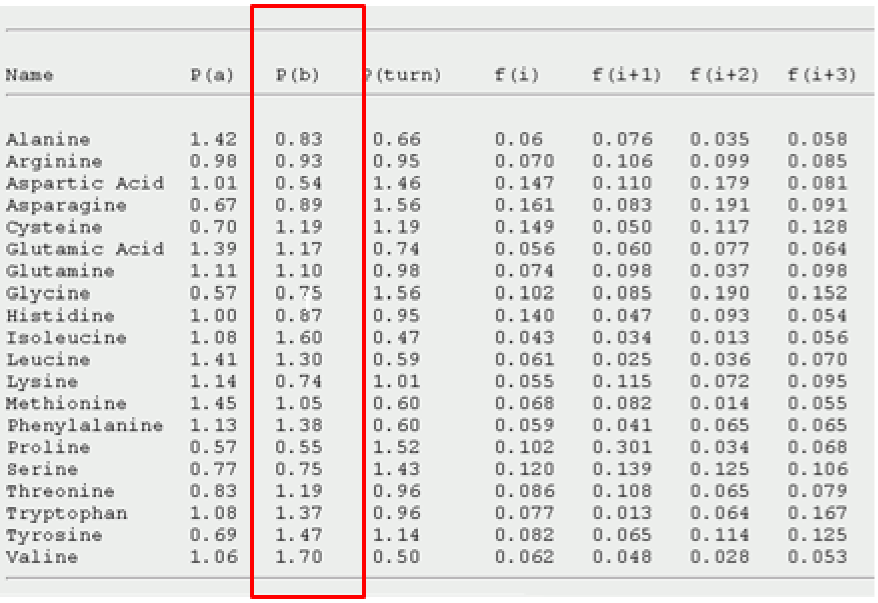
Name	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	1.42	0.83	0.66	0.06	0.076	0.035	0.058
Arginine	0.98	0.93	0.95	0.070	0.106	0.099	0.085
Aspartic Acid	1.01	0.54	1.46	0.147	0.110	0.179	0.081
Asparagine	0.67	0.89	1.56	0.161	0.083	0.191	0.091
Cysteine	0.70	1.19	1.19	0.149	0.050	0.117	0.128
Glutamic Acid	1.39	1.17	0.74	0.056	0.060	0.077	0.064
Glutamine	1.11	1.10	0.98	0.074	0.098	0.037	0.098
Glycine	0.57	0.75	1.56	0.102	0.085	0.190	0.152
Histidine	1.00	0.87	0.95	0.140	0.047	0.093	0.054
Isoleucine	1.08	1.60	0.47	0.043	0.034	0.013	0.056
Leucine	1.41	1.30	0.59	0.061	0.025	0.036	0.070
Lysine	1.14	0.74	1.01	0.055	0.115	0.072	0.095
Methionine	1.45	1.05	0.60	0.068	0.082	0.014	0.055
Phenylalanine	1.13	1.38	0.60	0.059	0.041	0.065	0.065
Proline	0.57	0.55	1.52	0.102	0.301	0.034	0.068
Serine	0.77	0.75	1.43	0.120	0.139	0.125	0.106
Threonine	0.83	1.19	0.96	0.086	0.108	0.065	0.079
Tryptophan	1.08	1.37	0.96	0.077	0.013	0.064	0.167
Tyrosine	0.69	1.47	1.14	0.082	0.065	0.114	0.125
Valine	1.06	1.70	0.50	0.062	0.048	0.028	0.053

1. Scan through the sequence : E M A V I Y P G
2. Identify sequence regions where:
   - 4 out of 6 contiguous residues give a P(α) > 1.0
   - That region is declared as alpha-helix
   - Extend helix to both sides until 4 out of 6 contiguous residues give a P(α) < 1.0

That is declared end of the helix. For Alpha Helices, 4 contiguous amino acids are required. Their Alpha-Helix propensity should be more than 1.0. Once this propensity falls below 1.0, Alpha-Helix stops.

Module 057: Chou Fasman Algorithm - II

Alpha Helices are formed from 4 contiguous amino acids having an Alpha-Helix propensity over 1.0. The Alpha-Helix stops if this propensity falls below 1.0. Once Alpha Helices are constructed, and concluded, the remaining amino acids can be evaluated for Beta sheets and turns etc. Let’s see how Beta sheets are evaluated using Chou Fasman Algorithm.



Name	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	1.42	0.83	0.66	0.06	0.076	0.035	0.058
Arginine	0.98	0.93	0.95	0.070	0.106	0.099	0.085
Aspartic Acid	1.01	0.54	1.46	0.147	0.110	0.179	0.081
Asparagine	0.67	0.89	1.56	0.161	0.083	0.191	0.091
Cysteine	0.70	1.19	1.19	0.149	0.050	0.117	0.128
Glutamic Acid	1.39	1.17	0.74	0.056	0.060	0.077	0.064
Glutamine	1.11	1.10	0.98	0.074	0.098	0.037	0.098
Glycine	0.57	0.75	1.56	0.102	0.085	0.190	0.152
Histidine	1.00	0.87	0.95	0.140	0.047	0.093	0.054
Isoleucine	1.08	1.60	0.47	0.043	0.034	0.013	0.056
Leucine	1.41	1.30	0.59	0.061	0.025	0.036	0.070
Lysine	1.14	0.74	1.01	0.055	0.115	0.072	0.095
Methionine	1.45	1.05	0.60	0.068	0.082	0.014	0.055
Phenylalanine	1.13	1.38	0.60	0.059	0.041	0.065	0.065
Proline	0.57	0.55	1.52	0.102	0.301	0.034	0.068
Serine	0.77	0.75	1.43	0.120	0.139	0.125	0.106
Threonine	0.83	1.19	0.96	0.086	0.108	0.065	0.079
Tryptophan	1.08	1.37	0.96	0.077	0.013	0.064	0.167
Tyrosine	0.69	1.47	1.14	0.082	0.065	0.114	0.125
Valine	1.06	1.70	0.50	0.062	0.048	0.028	0.053

1. Compute P(β) for contiguous regions of 5 Amino Acids
2. From these regions, identify regions where:
3. 5 contiguous residues have P(α) > P(β)

That region is finalized as alpha-helix.Repeat this step for the full amino acid sequence to finalize all possible alpha helical regions in the sequence.

Alpha Helices can be finalized if their propensity is higher than the propensity for Beta Sheets in regions of 5 amino acids. For those regions where that is not the case, further evaluation is required.

Module 058: Chou Fasman Algorithm - III

Alpha Helices are formed from 4 contiguous amino acids having an Alpha-Helix propensity over 1.0. The Alpha-Helix stops if this propensity falls below 1.0. Alpha Helices were finalized if their propensity was higher than the propensity for Beta Sheets in regions of 5 amino acids.

We can evaluate such regions for Beta Sheets. Let us see step by stop how to find a beta sheet and how to differentiate them from alpha helices.

Scan the sequence to identify regions where:

- 3 out of 5 amino acids have P(β) > 1.0
- That region is declared as beta sheet
- Extend beta sheet to both sides until
  4 contiguous residues average P(β) < 1.0
- That is declared end of the beta sheet
- Those regions are finalized as beta-sheets which have average P(β) > 1.05 and the average P(β) > P(α) for that region.

Regions where overlapping alpha-helices and beta-sheets occur are declared helices if

- the average P(a-helix) > P(b-sheet) for that region

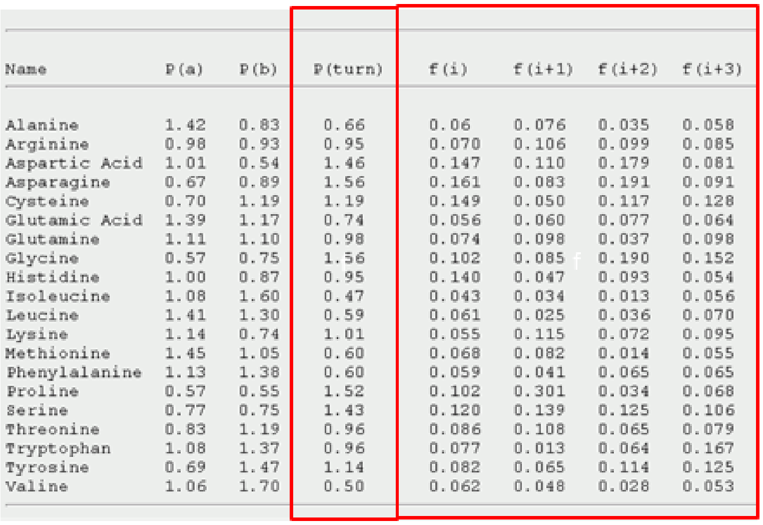
Else, a beta sheet is declared if

- average P(b-sheet) > P(a-helix) for that region

Using the strategy of higher propensity, alpha helices and beta sheets can be completely resolved. Assignments for each beta sheet and alpha helix can be finalized.

Module 059: Chou Fasman Algorithm - IV

After computing the propensity of alpha helices and beta sheets, we need to settle for loops. Let’s see how we can find out the loops using Chou Fasman Algorithm.

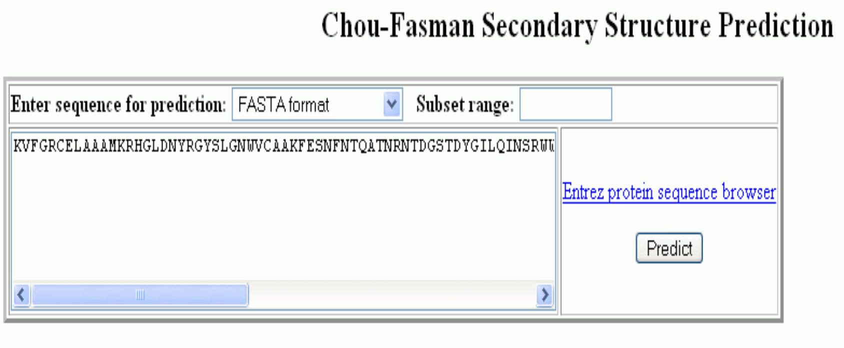


Name	P(a)	P(b)	P(turn)	f(i)	f(i+1)	f(i+2)	f(i+3)
Alanine	1.42	0.83	0.66	0.06	0.076	0.035	0.058
Arginine	0.98	0.93	0.95	0.070	0.106	0.099	0.085
Aspartic Acid	1.01	0.54	1.46	0.147	0.110	0.179	0.081
Asparagine	0.67	0.89	1.56	0.161	0.083	0.191	0.091
Cysteine	0.70	1.19	1.19	0.149	0.050	0.117	0.128
Glutamic Acid	1.39	1.17	0.74	0.056	0.060	0.077	0.064
Glutamine	1.11	1.10	0.98	0.074	0.098	0.037	0.098
Glycine	0.57	0.75	1.56	0.102	0.085	0.190	0.152
Histidine	1.00	0.87	0.95	0.140	0.047	0.093	0.054
Isoleucine	1.08	1.60	0.47	0.043	0.034	0.013	0.056
Leucine	1.41	1.30	0.59	0.061	0.025	0.036	0.070
Lysine	1.14	0.74	1.01	0.055	0.115	0.072	0.095
Methionine	1.45	1.05	0.60	0.068	0.082	0.014	0.055
Phenylalanine	1.13	1.38	0.60	0.059	0.041	0.065	0.065
Proline	0.57	0.55	1.52	0.102	0.301	0.034	0.068
Serine	0.77	0.75	1.43	0.120	0.139	0.125	0.106
Threonine	0.83	1.19	0.96	0.086	0.108	0.065	0.079
Tryptophan	1.08	1.37	0.96	0.077	0.013	0.064	0.167
Tyrosine	0.69	1.47	1.14	0.082	0.065	0.114	0.125
Valine	1.06	1.70	0.50	0.062	0.048	0.028	0.053

For any *jth* residue in sequence, we calculate
f (Total) = f(j) f(j+1) f(j+2) f(j+3) (tetrapeptide)

If

- f(Total) > 0.000075
- the average value for P(turn) > 1.00 in the tetra peptide
  - the averages for the tetra peptide are such P(a-helix) < P(turn) > P(b-sheet)



Chou-Fasman Secondary Structure Prediction

Enter sequence for prediction: FASTA format Subset range:

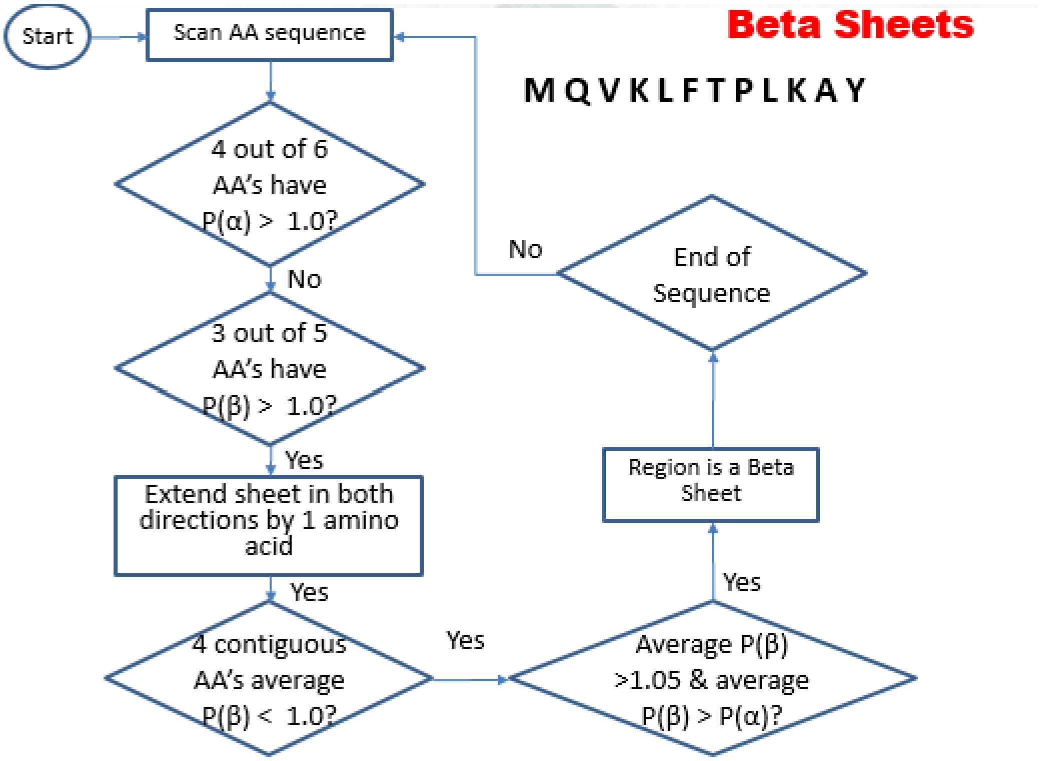
KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWW

http://fasta.bioch.virginia.edu/fasta_www/chofas.htm

Chou Fasman Algorithm helps predict Alpha Helices, Beta Sheets and Turns. The algorithm is based on statistical occurrence of Amino Acids in known structures.

Module 060: Chou Fasman Algorithm – Flowchart I

Chou Fasman Algorithm helps predict secondary structures such as Alpha Helices, Beta Sheets and Turns. Step by step flowchart of the entire algorithm.


Beta Sheets
MQVKLFTPLKAY
Start
Scan AA sequence
4 out of 6 AA's have P(α) > 1.0?
No
3 out of 5 AA's have P(β) > 1.0?
Yes
Extend sheet in both directions by 1 amino acid
Yes
4 contiguous AA's average P(β) < 1.0?
Yes
Average P(β) >1.05 & average P(β) > P(α)?
Yes
Region is a Beta Sheet
End of Sequence
No

Beta sheets can be predicted from primary amino acid sequences. Next, we will see the flowchart of Alpha Helices and Beta Turns.

Module 061: Chou Fasman Algorithm – Flowchart II

Chou Fasman Algorithm helps predict secondary structures such as Alpha Helices, Beta Sheets and Turns. Step by step flowchart of the entire algorithm.

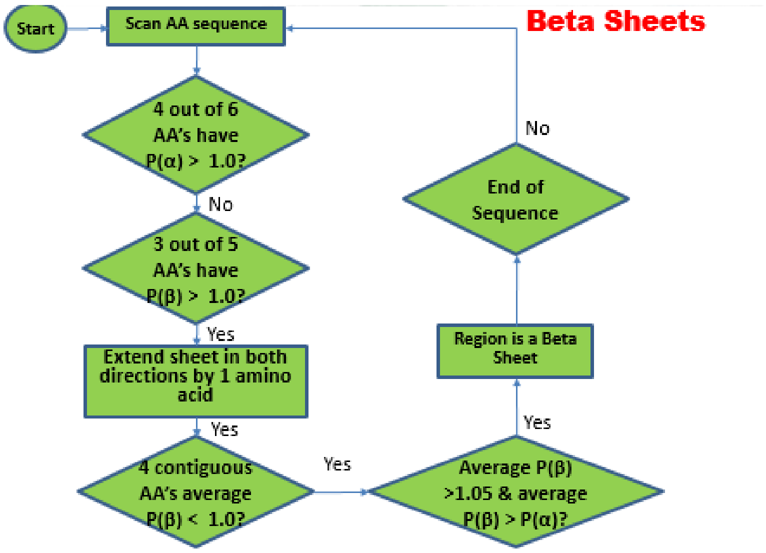

Beta Sheets
Start
Scan AA sequence
4 out of 6 AA's have P(α) > 1.0?
No
3 out of 5 AA's have P(β) > 1.0?
Yes
Extend sheet in both directions by 1 amino acid
Yes
4 contiguous AA's average P(β) < 1.0?
Yes
Average P(β) >1.05 & average P(β) > P(α)?
Yes
Region is a Beta Sheet
End of Sequence
No

Figure 80 Beta sheet flowchart

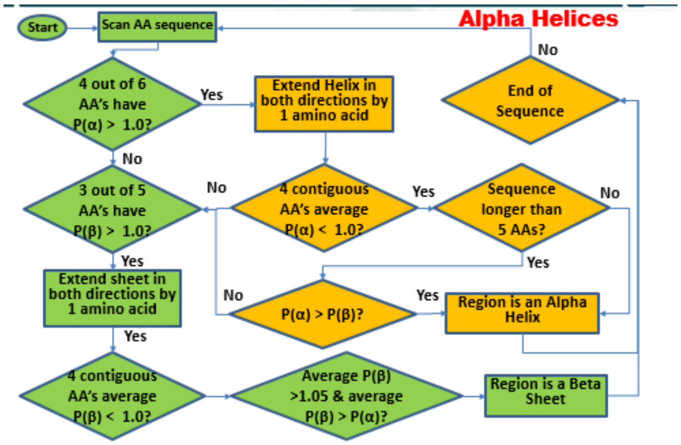

Alpha Helices
Start
Scan AA sequence
4 out of 6 AA's have P(α) > 1.0?
Yes
Extend Helix in both directions by 1 amino acid
No
3 out of 5 AA's have P(β) > 1.0?
No
4 contiguous AA's average P(α) < 1.0?
Yes
Sequence longer than 5 AAs?
No
Yes
Yes
Extend sheet in both directions by 1 amino acid
No
P(α) > P(β)?
Yes
Region is an Alpha Helix
Yes
4 contiguous AA's average P(β) < 1.0?
Average P(β) >1.05 & average P(β) > P(α)?
Region is a Beta Sheet
End of Sequence
No

Figure 81 Alpha helices flowchart

Now we have reviewed flowcharts for Alpha Helices and Beta Sheets. Next up is the flow chart for Beta Turns.

Module 062: Chou Fasman Algorithm – Flowchart III

Chou Fasman Algorithm helps predict secondary structures such as Alpha Helices, Beta Sheets and Turns. Step by step flowchart of the entire algorithm.

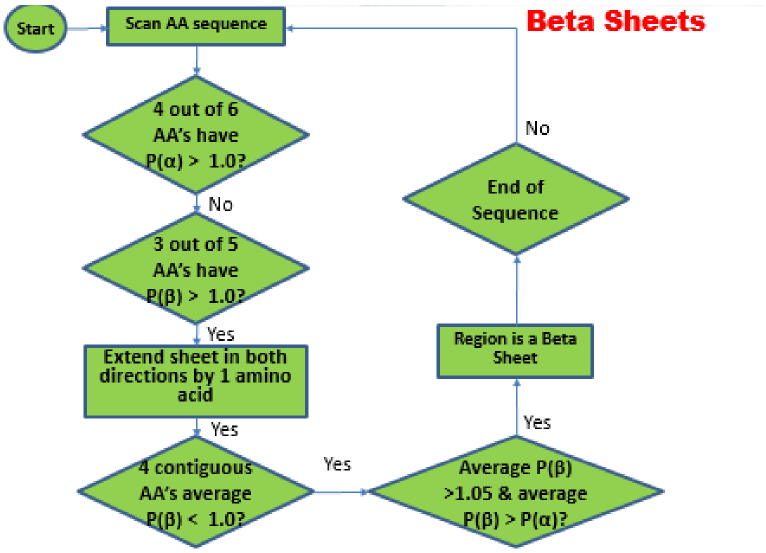

Beta Sheets
Start
Scan AA sequence
4 out of 6 AA's have P(α) > 1.0?
No
3 out of 5 AA's have P(β) > 1.0?
Yes
Extend sheet in both directions by 1 amino acid
Yes
4 contiguous AA's average P(β) < 1.0?
Yes
Average P(β) >1.05 & average P(β) > P(α)?
Yes
Region is a Beta Sheet
End of Sequence
No

Figure 82 Beta sheet flowchart

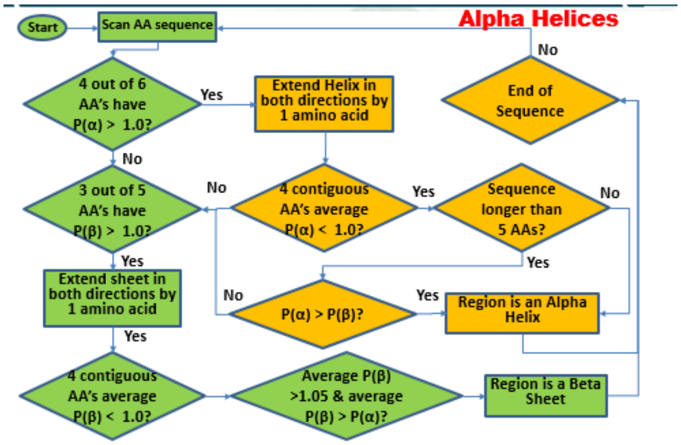

Alpha Helices
Start
Scan AA sequence
4 out of 6 AA's have P(α) > 1.0?
Yes
Extend Helix in both directions by 1 amino acid
End of Sequence
No
No
3 out of 5 AA's have P(β) > 1.0?
No
4 contiguous AA's average P(α) < 1.0?
Yes
Sequence longer than 5 AAs?
No
Yes
Yes
Extend sheet in both directions by 1 amino acid
No
P(α) > P(β)?
Yes
Region is an Alpha Helix
Yes
4 contiguous AA's average P(β) < 1.0?
Average P(β) >1.05 & average P(β) > P(α)?
Region is a Beta Sheet

Figure 83 Alpha helices flowchart

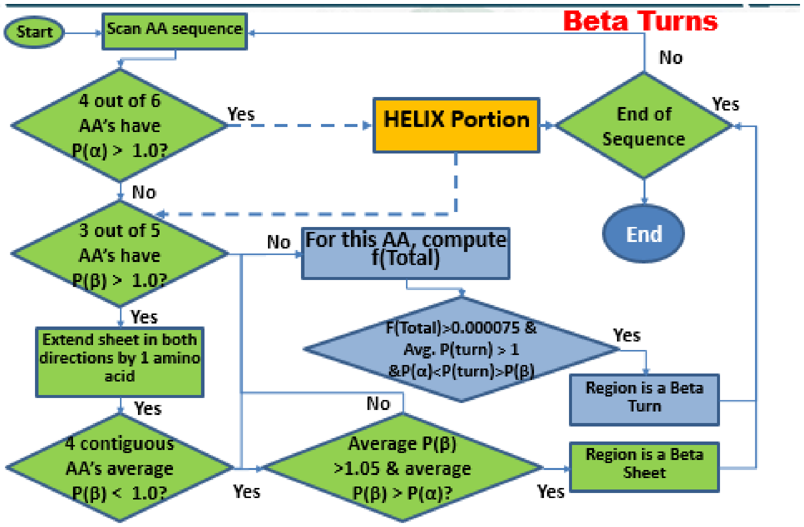

Beta Turns
Start
Scan AA sequence
4 out of 6 AA's have P(α) > 1.0?
Yes
No
HELIX Portion
End of Sequence
No
Yes
End
3 out of 5 AA's have P(β) > 1.0?
No
For this AA, compute f(Total)
Yes
Extend sheet in both directions by 1 amino acid
F(Total)>0.000075 & Avg. P(turn) > 1 &P(α)<P(turn)>P(β)
Yes
Region is a Beta Turn
No
Yes
4 contiguous AA's average P(β) < 1.0?
Yes
Average P(β) >1.05 & average P(β) > P(α)?
Yes
Region is a Beta Sheet

Figure 84 Beta turn

Alpha helices, beta sheets and turns can be predicted using Chou Fasman Algorithm. This algorithm is based on statistical analysis of amino acid occurrences in proteins.

Module 063: Chou Fasman Algorithm – Improvements

Alpha helices, beta sheets and turns can be predicted using Chou-Fasman Algorithm. The algorithm is based on statistical analysis of amino acid occurrences in proteins.

Secondary structure propensity values of alpha helix, beta sheet and turns should be recalculated with the latest protein data sets.

IMPROVEMENTS

Special consideration for:

- Nucleation regions
- Membrane proteins
- Hydrophobic domains
- Consider variable coil and loop sizes besides the from tetra peptide turns
- Consider local protein folding environments
- Solvent accessibility of residues
- Protein structural class
- Protein's organism

Chou Fasman can be improved to better predict secondary structures by incorporating biochemical factors and updated statistics!

Module 064: Summary of Visualization, Classification & Prediction

Structure Classification

- relationship between protein structure and function
- There is need to classify proteins
- Hierarchy of classification

Structure visualization, classification and prediction equip us to perform functional evaluation of proteins. This is important for understanding disease and designing drugs for treating them.