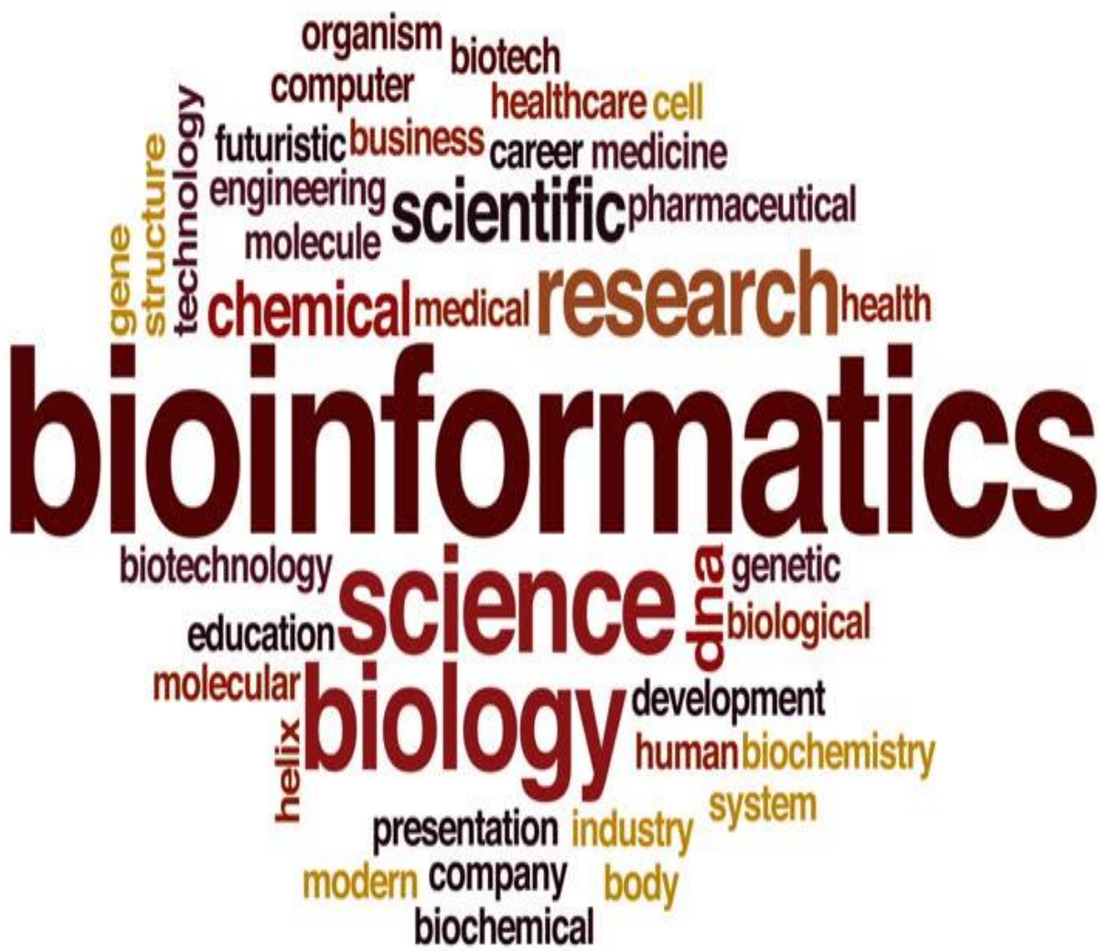
Virtual University of Pakistan

Federal Government University

World-Class Education at Your Doorstep

M.SC ZOOLOGY

BIF401-Bioinformatics-I



HANDOUTS TOPIC NO 1 TO 250

MUHAMMAD IMRAN

1 Background of 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.

2 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.

3 Need for Bioinformatics – I

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.

4 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.

5 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.

6 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.

7 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.

8 Frontiers 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.

9 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
Molecular Evolution	Molecular evolution and phylogeny
	Sequence Evolution
	Introduction to Unweighted Pair Group Method with Arithmetic Mean (UPGMA)
	Introduction to maximum parsimony

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

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!

10 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!

11 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

12 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

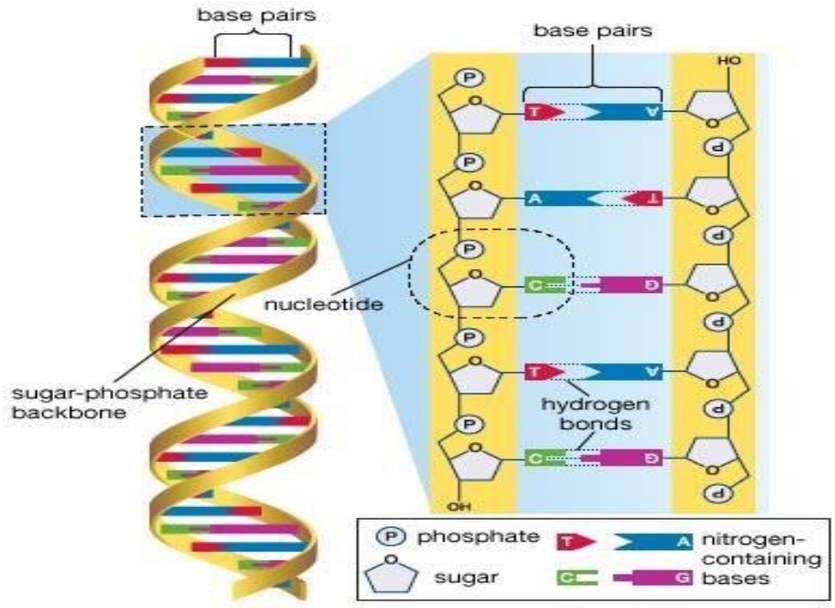


Figure 0.1 DNA Double helix

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.

13 Transcription

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

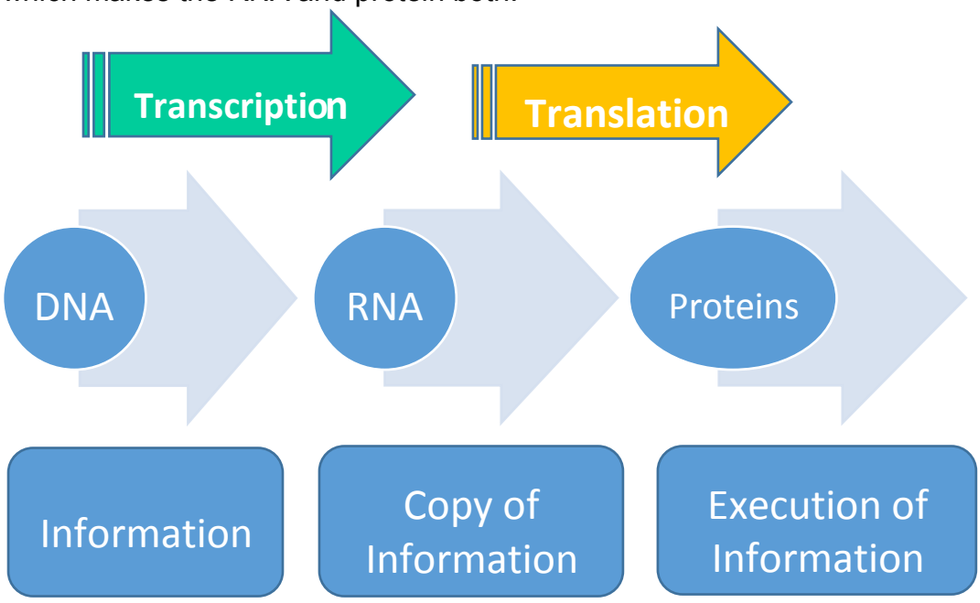


Figure 0.2 Flow of information from DNA to Proteins

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.

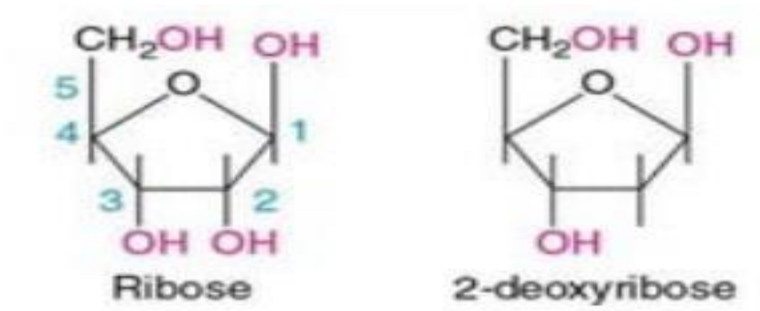
14 Nucleotides

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:

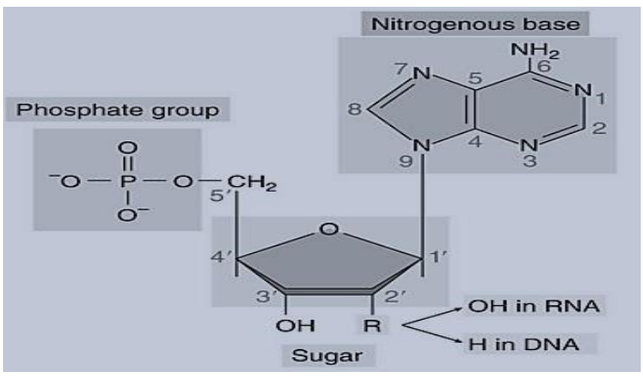


Difference between RNA and DNA sugar

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.

15 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

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

16 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.

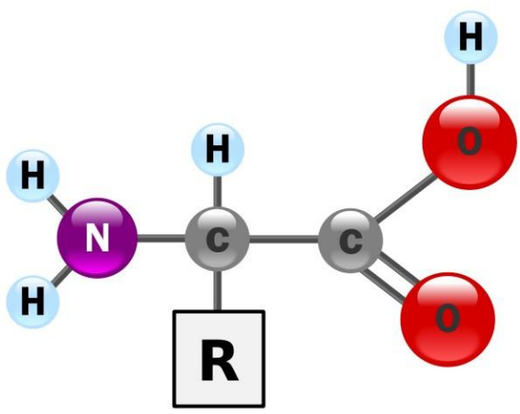
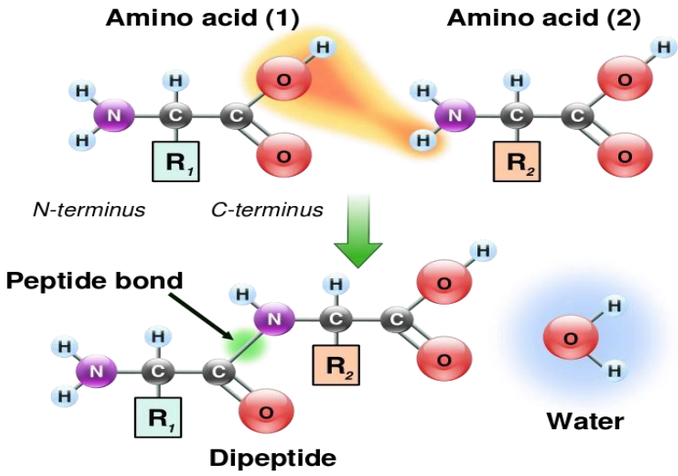


Figure 0.5 structure of amino acid

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.

17 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

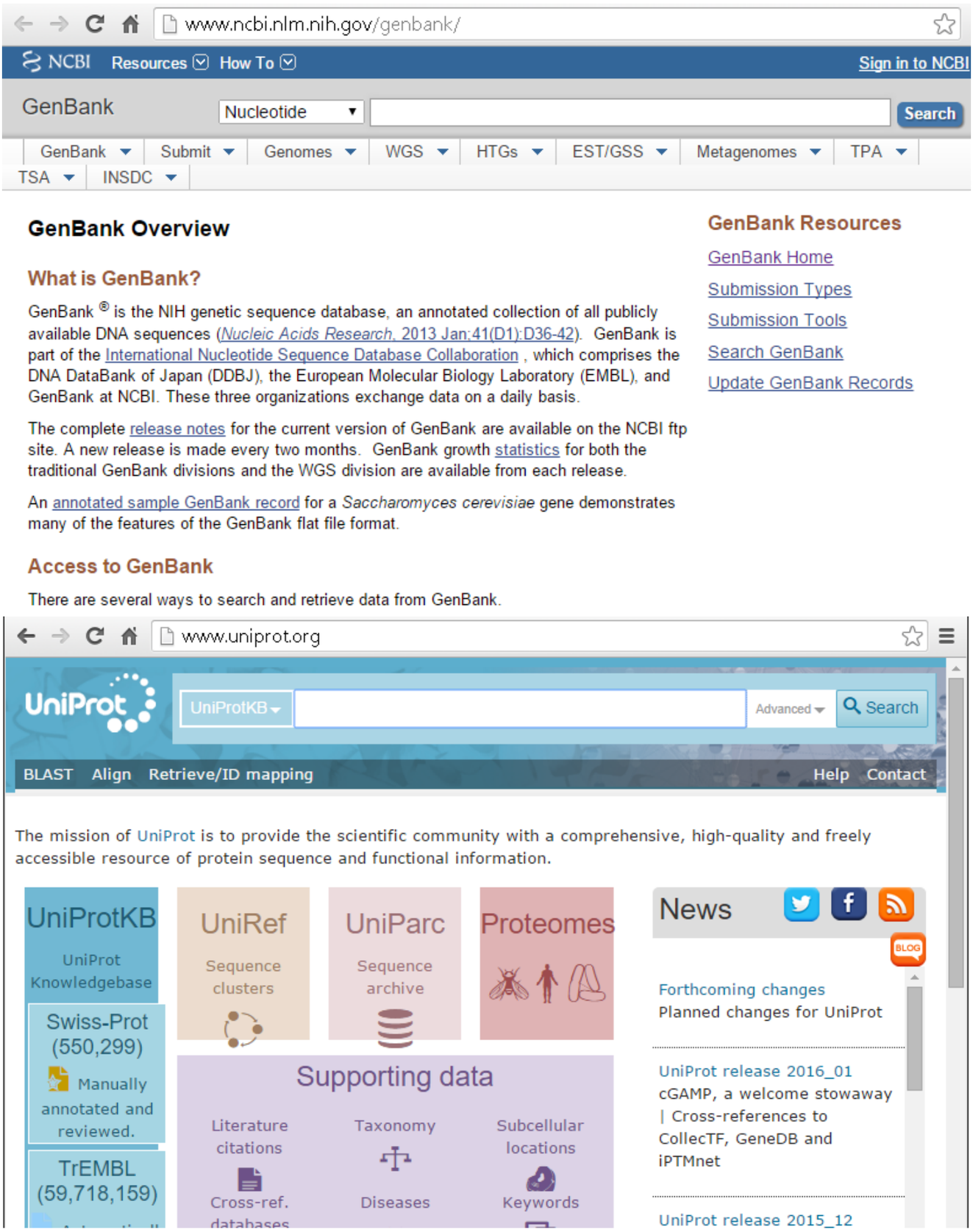
This large number of sequence or bases cannot be stored in a single computer that's why solution lies in public sequence data bases for DNA & RNA the public database is **GenBank (by NIH).**

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.

Storage of Biological Sequence Information

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



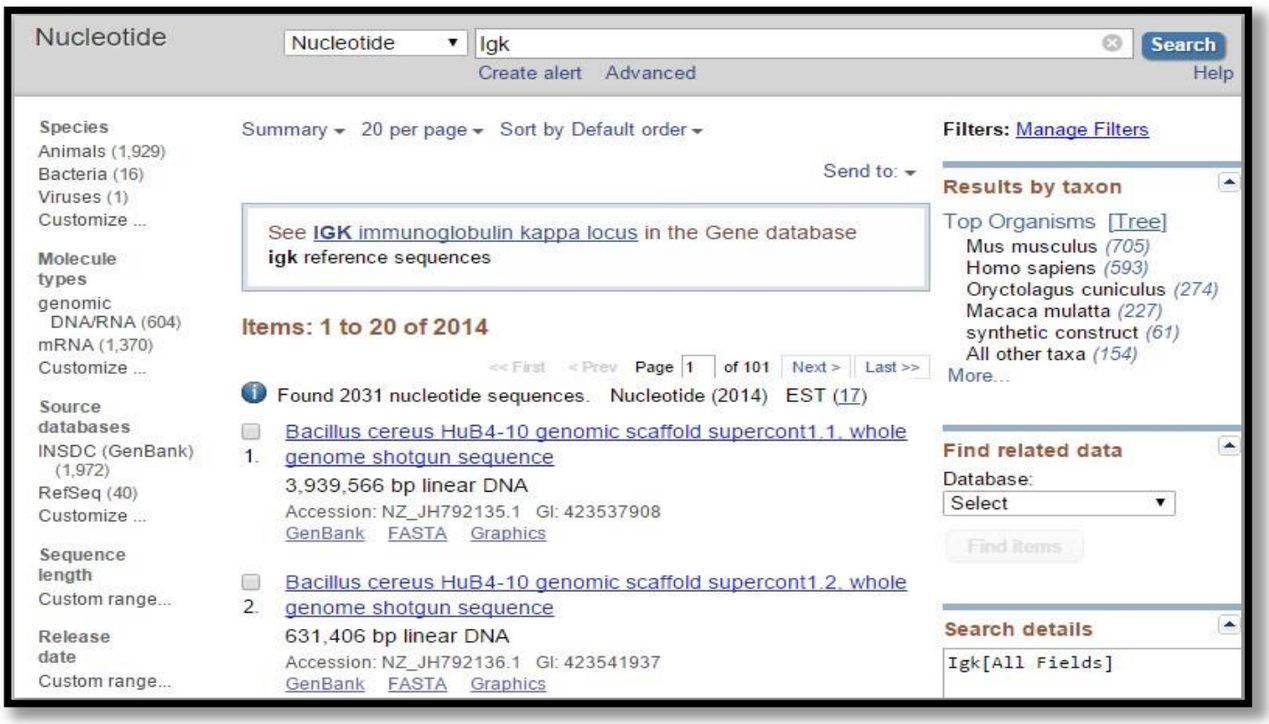
18 Using Entrez (GenBank)

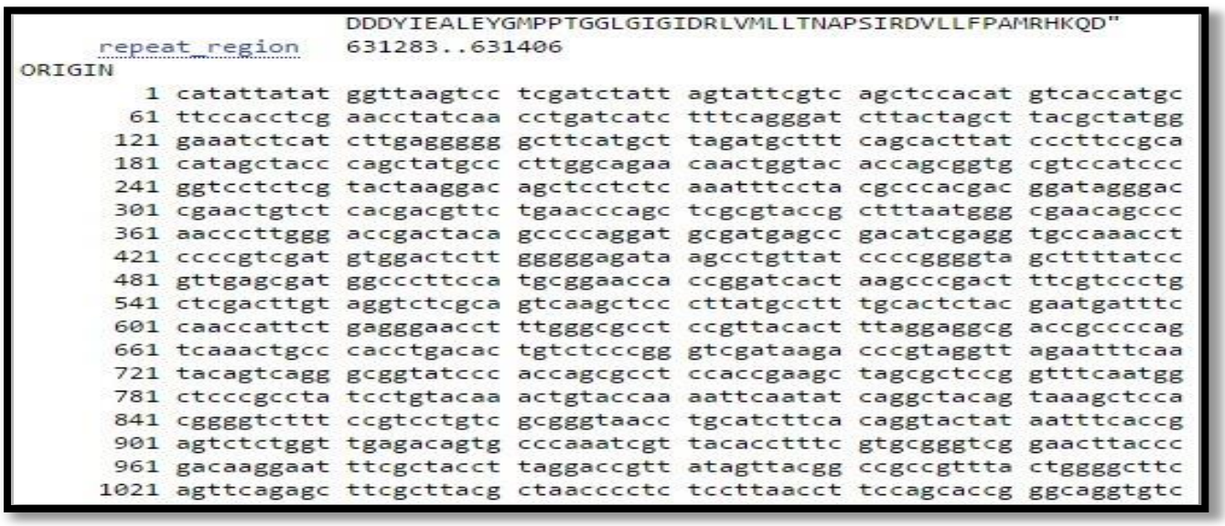
GenBank is online database where researcher can get access to the sequences of DNA, RNA and proteins.

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.





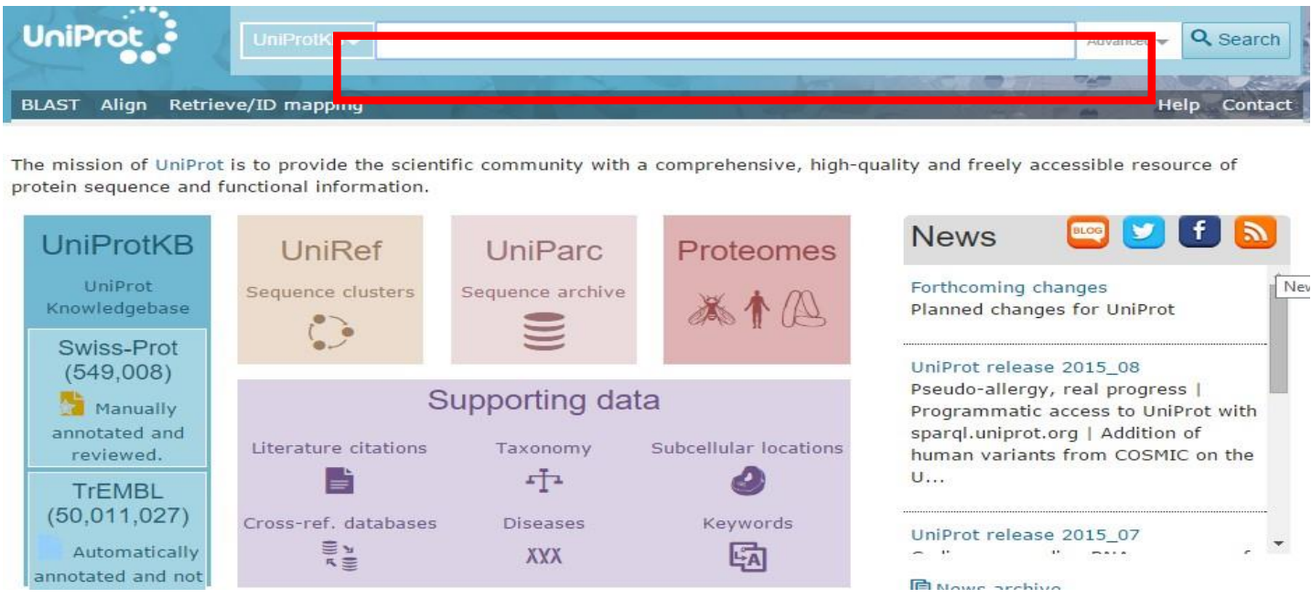
Sequences can be searched from GenBank by typing;

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

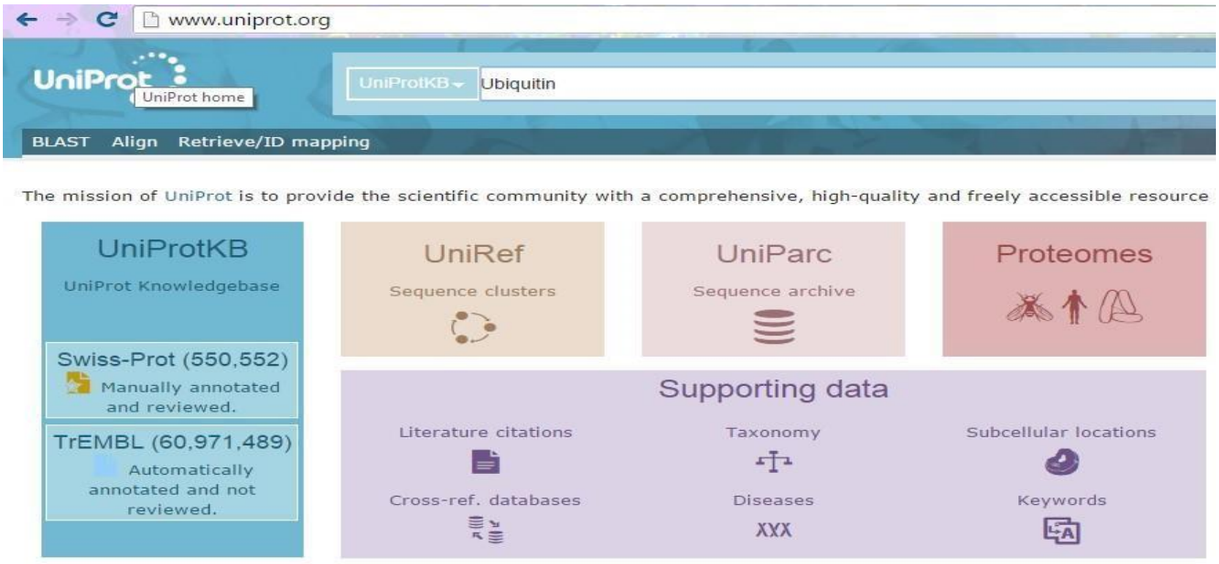
19 Using Uniprot

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

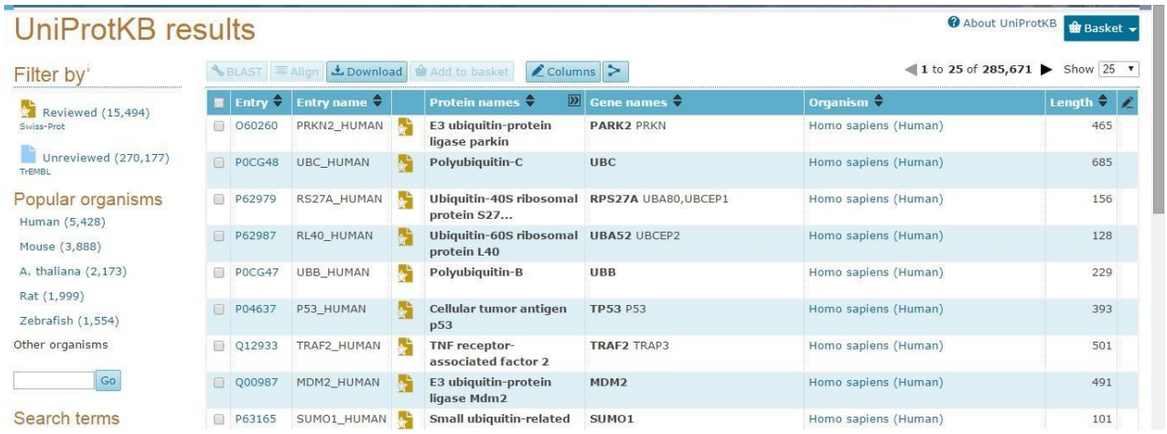
www.Uniprot.org



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.



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



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.

20 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

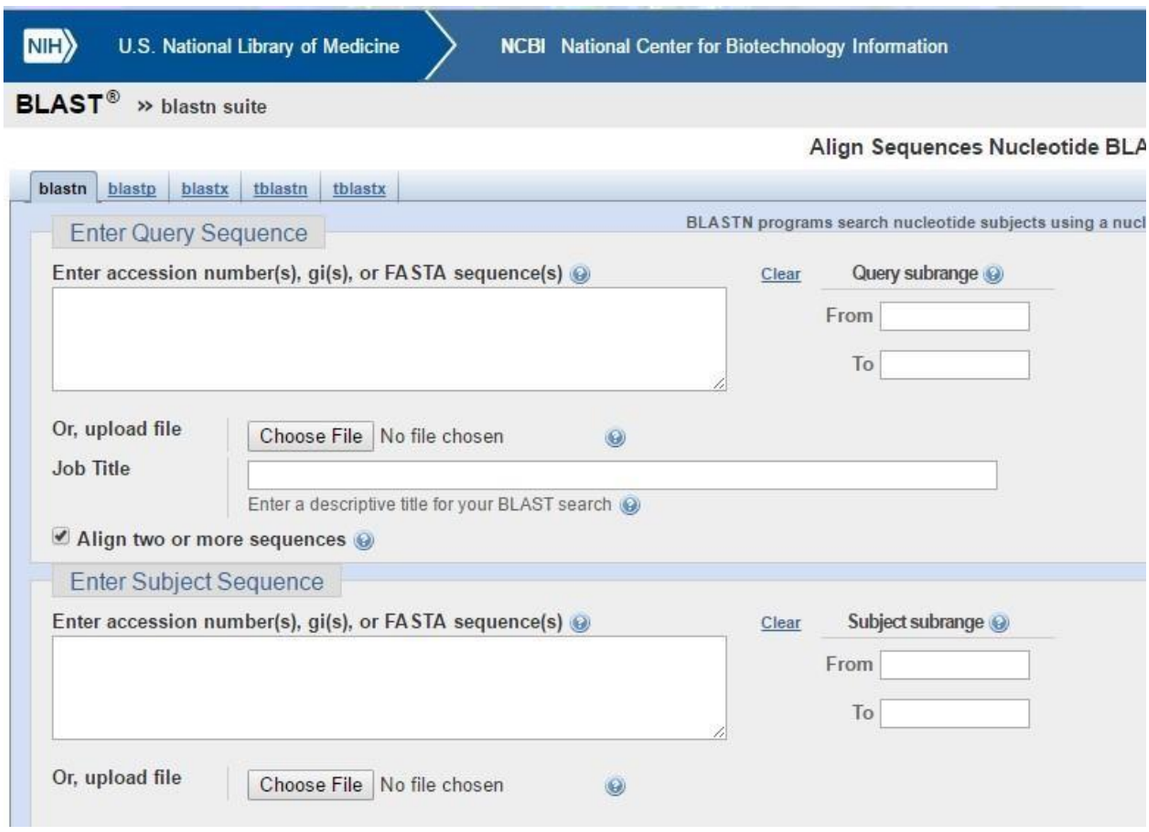


Figure 0.6 BLAST is used to compare the nucleotides sequences

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.

21 Similarities and 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.

Exact Matching

Not only should the two or more sequences being compared have the exact same number of each nucleotide (for DNA /RNA) or amino acid (for Proteins), but that they should be arranged in the exact same order!

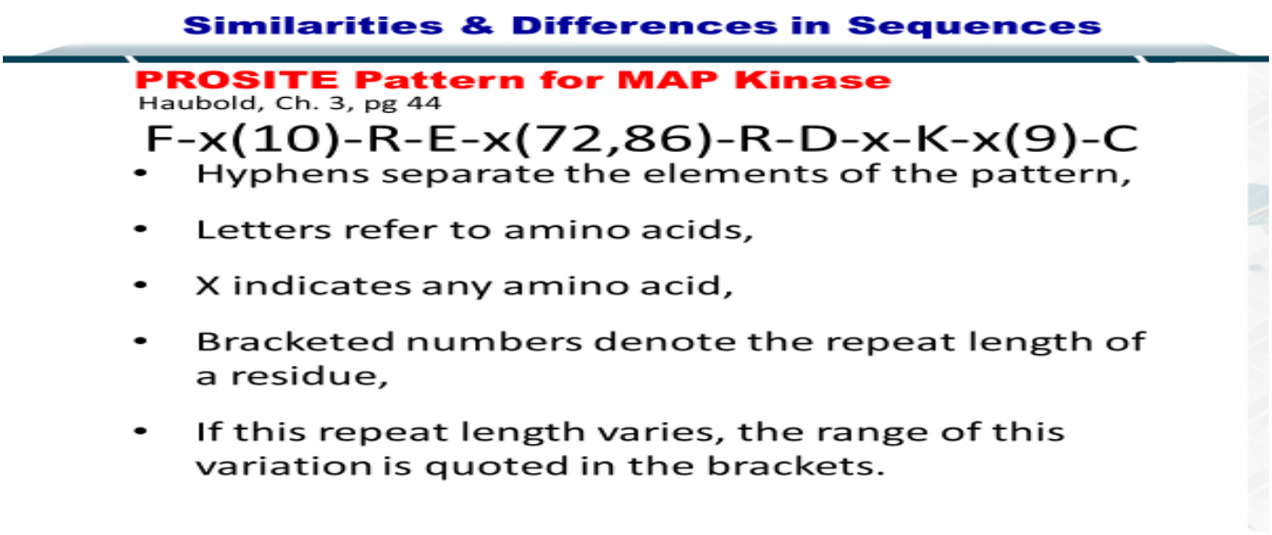
CONCLUSION

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/



22 Pairwise Sequence Alignment-I

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.

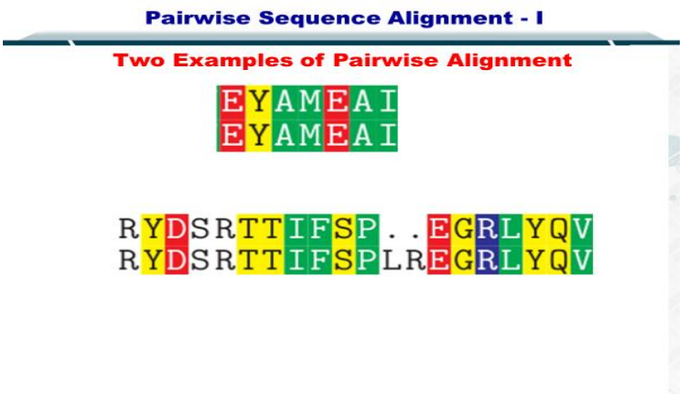
There are two types of pair alignments.

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).

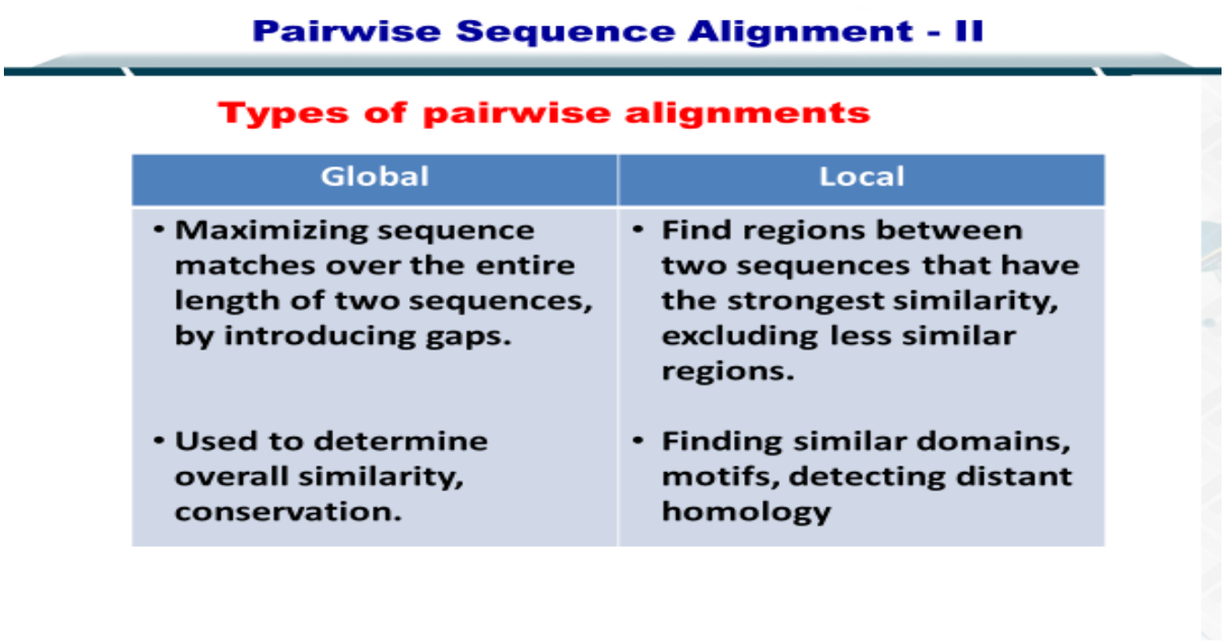


23 Pairwise Sequence Alignment-II

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

Salient Points

- **Sliding the sequences past each other**
- **Aim is to maximize matches**
- **Gaps ('.') could be inserted to account for insertions and deletions**
- **Gaps ('.') may carry a penalty for reducing scores from unreasonable alignments**
- **Without a gap penalty an exact match and match with gaps will get equal scores**



Global alignment - maximizes the number of matches between the query and source sequences along the entire length of both the sequences.

Local alignment - gives the highest scoring local match between both query and sequences.

Optimal alignment - one that exhibits the most correspondences between the query and the source sequences. It is the alignment with the **highest score**. Biologically meaningful?

24 Pairwise Sequence Alignment-III

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.

Pairwise Sequence Alignment - III

Optimal Alignment

Align 1 & 2

1: THIS IS A RATHER LONGER SENTENCE THAN THE NEXT.
2: THIS IS A SHORT SENTENCE.

1: THIS IS A RATHER LONGER SENTENCE THAN THE NEXT.
2: THIS IS A .....SHORT.....SENTENCE.....

OR

1: THIS IS A RATHER LONGER SENTENCE THAN THE NEXT.
2: THIS IS A ....SHORT.....SENT..EN.....CE.....

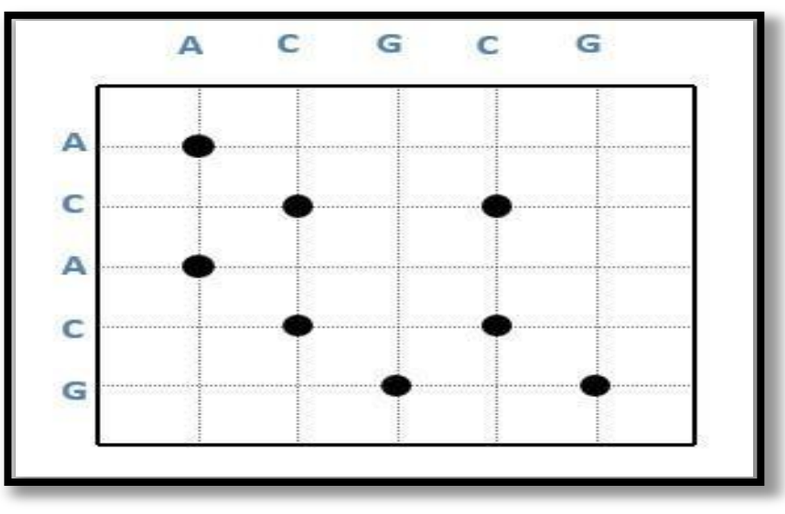
Which alignment is optimal in biological context?

25 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.

		A	C	G	C	G
A						
C						
A						
C						
G						

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



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

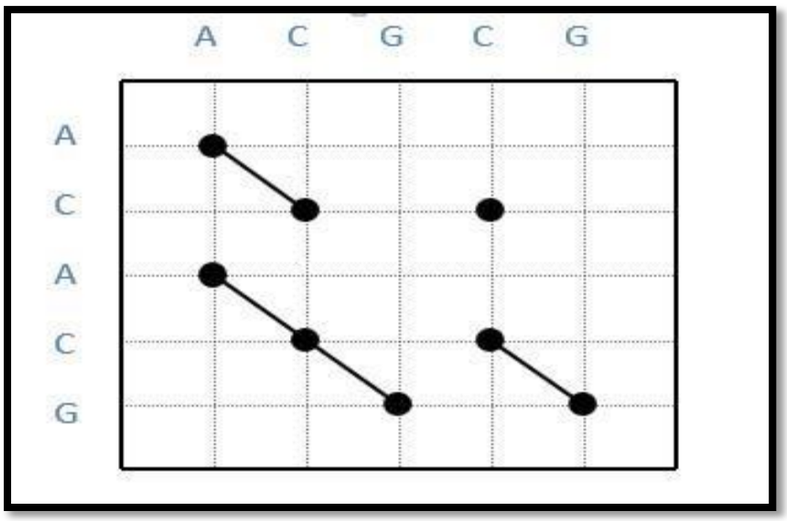


Figure 0.7 dot plot diagonal pattern

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

26 Example 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.

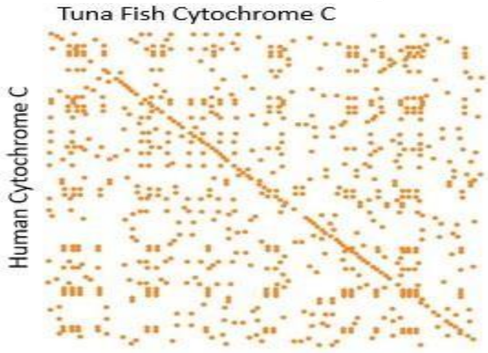


Figure 0.8 tuna fish vs Human

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.

27 Identity 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.

28 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.

29 Why local alignments?

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.

30 Aligning In-dels

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

31 Aligning Mutations in Sequences

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

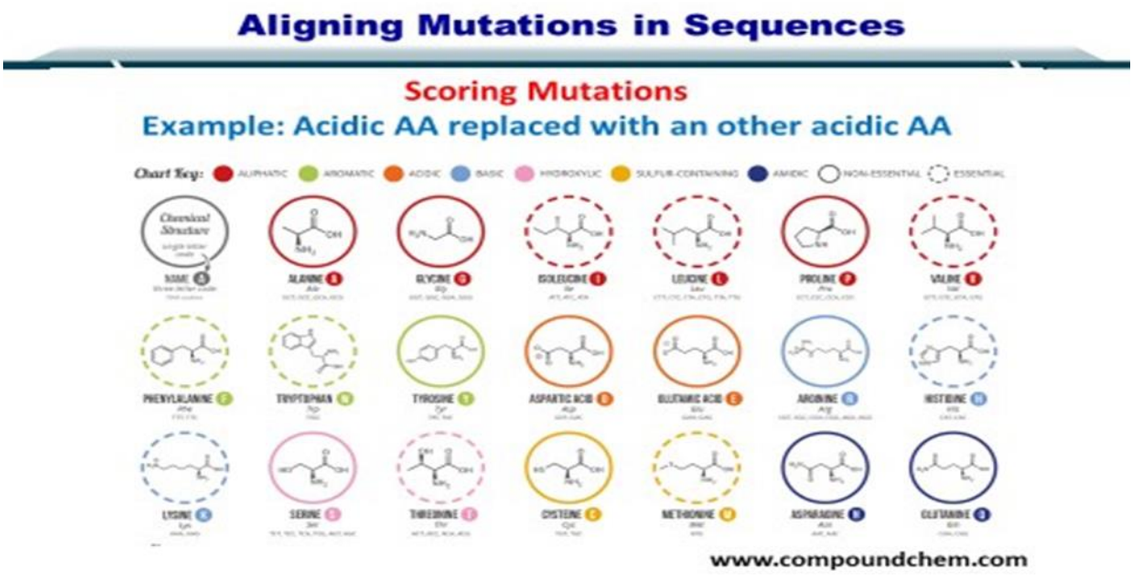
Mutation is totally different from Indels, because in Mutation we replace the amino acid with other amino acids and replace the nucleotides with other and we don't use Gaps is inserted in template or target for mutation.

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



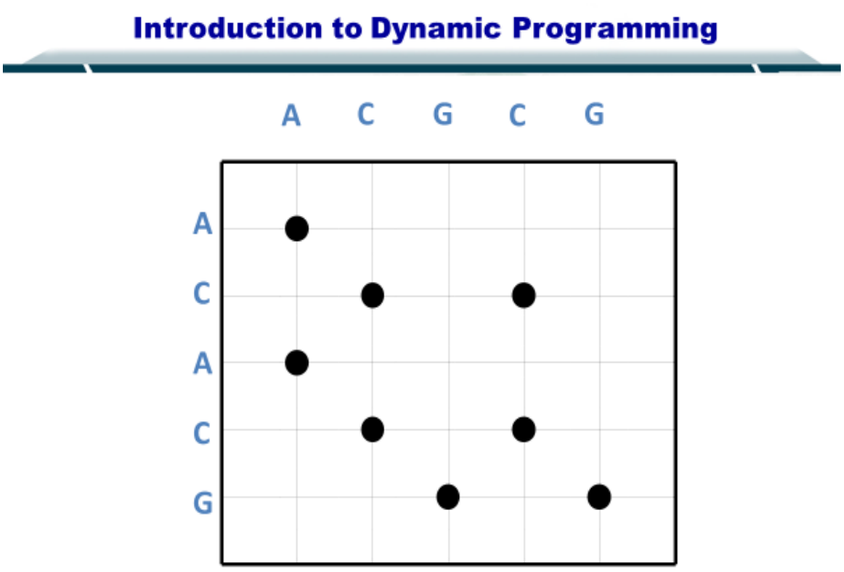
CONCLUSION

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

32 Introduction to Dynamic Programming

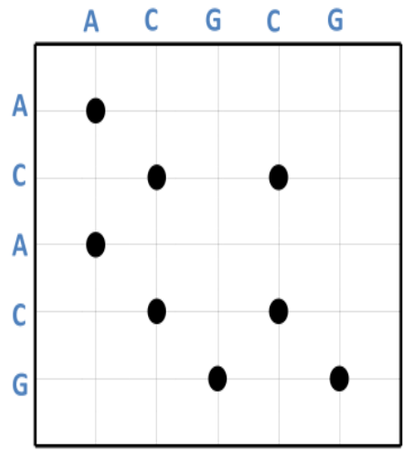
Introduction

- **To perform an alignment by sliding sequences across each other, we used dot plots to find matching nucleotides and amino acids**

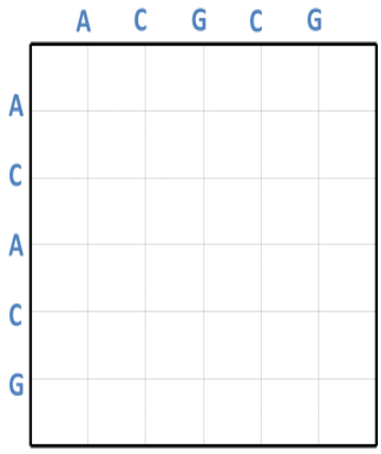


Where are the indels and gaps?

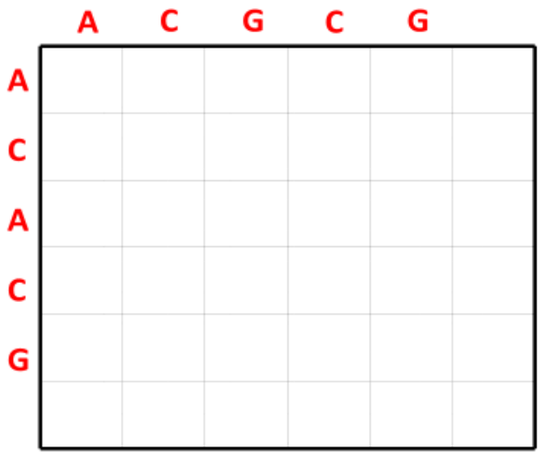
- **Dot plots cannot capture insertions, deletions and gap indications**
- **How can we deal with them?**
- **Modify the dot plot!**



Introduction to Dynamic Programming



Introduction to Dynamic Programming



Conclusion

- **Matches are labelled by +1 instead of a dot**
- **Indels (gaps) and substitutions (mutations) can be included in the dot plot as -1**

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.

33 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.

Compare "A" with "B"

Compare "AB" with "BC"

Compare "ABC" with "DEF"

Compare "ABCD" with "EFGH"

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.

34 Dynamic Programming Methodology

Introduction

- Dynamic programming (DP) helps reduce the large computational cost associated with sequence comparisons
- Let's learn the DP methodology

Dynamic Programming Methodology

DP uses a "scoring function" to deal with matches, mismatches and gaps!

An example Scoring Scheme:

Match:	**+a**
Mismatch:	**- b**
Gap:	**- c**

Score = #matches + #mismatches + #gaps

Dynamic Programming Methodology

DP uses a "scoring function" to deal with matches, mismatches and 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

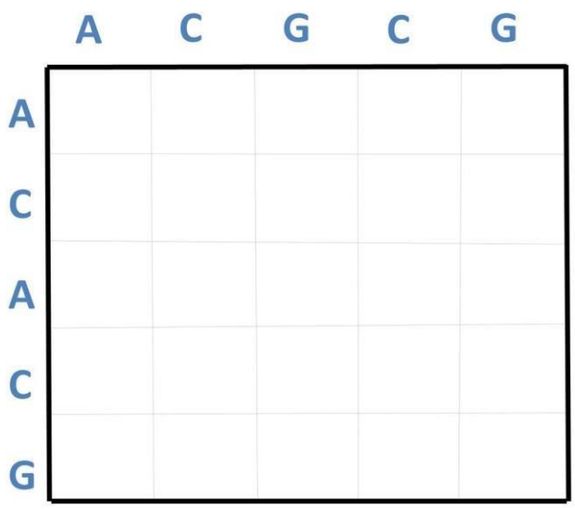
Dynamic Programming Methodology

Coming back to the dot plot version

	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

Dynamic Programming Methodology

Match Rewards: 10
Mismatch penalty: 2



Conclusion

- Alignments are represented by diagonals in the dot matrix plot
- Total score can be computed for each possible alignment
- The best alignment is then selected

35 Needleman Wunsh 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

Needleman-Wunsch

match = 1 mismatch = -1 gap = -1

		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

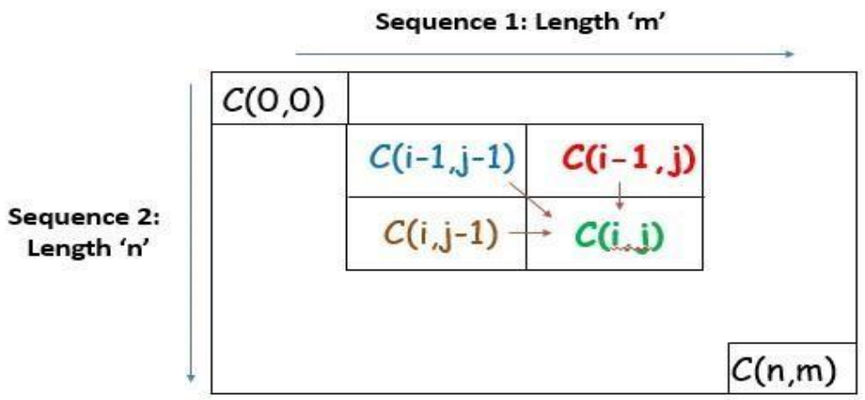


Figure 12 Needleman wunsch algorithm way of computation of nucleotides

36 Needleman Wunsh Algorithm-II

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

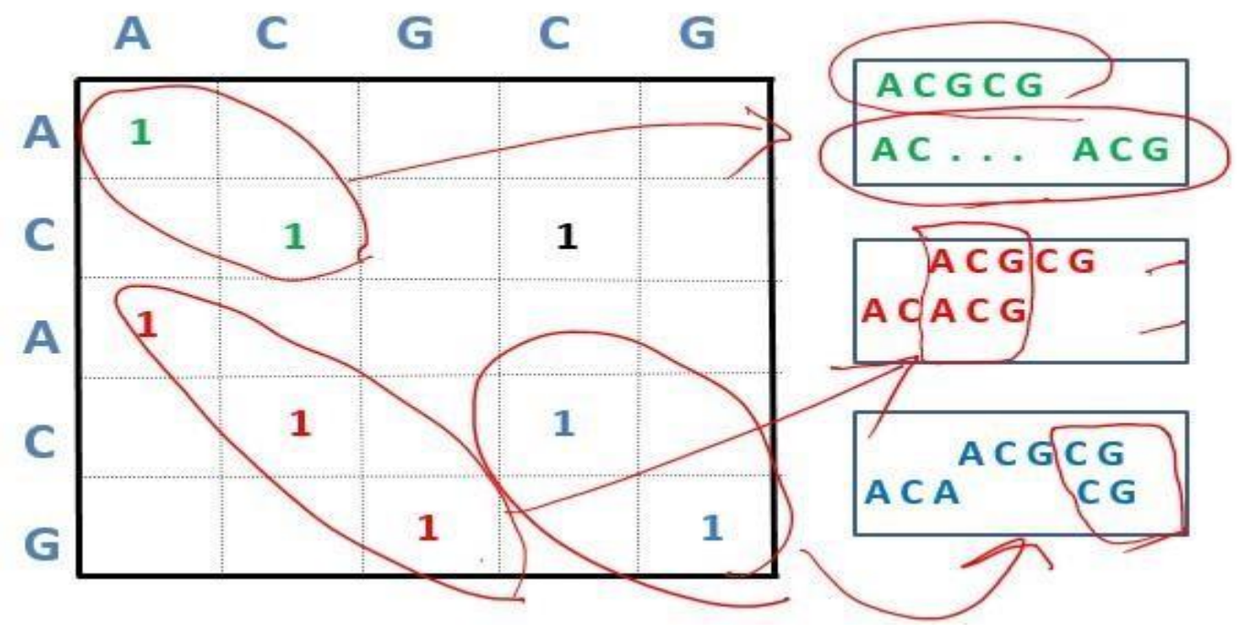


Figure 13 various combinations of sequences through dot plot

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.

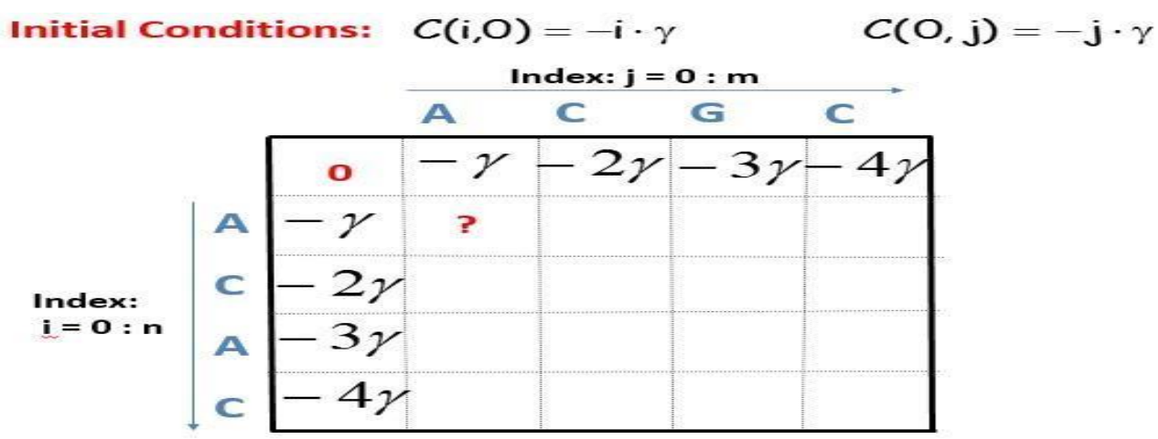


Figure 14 initial column and row are kept zero (0)

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

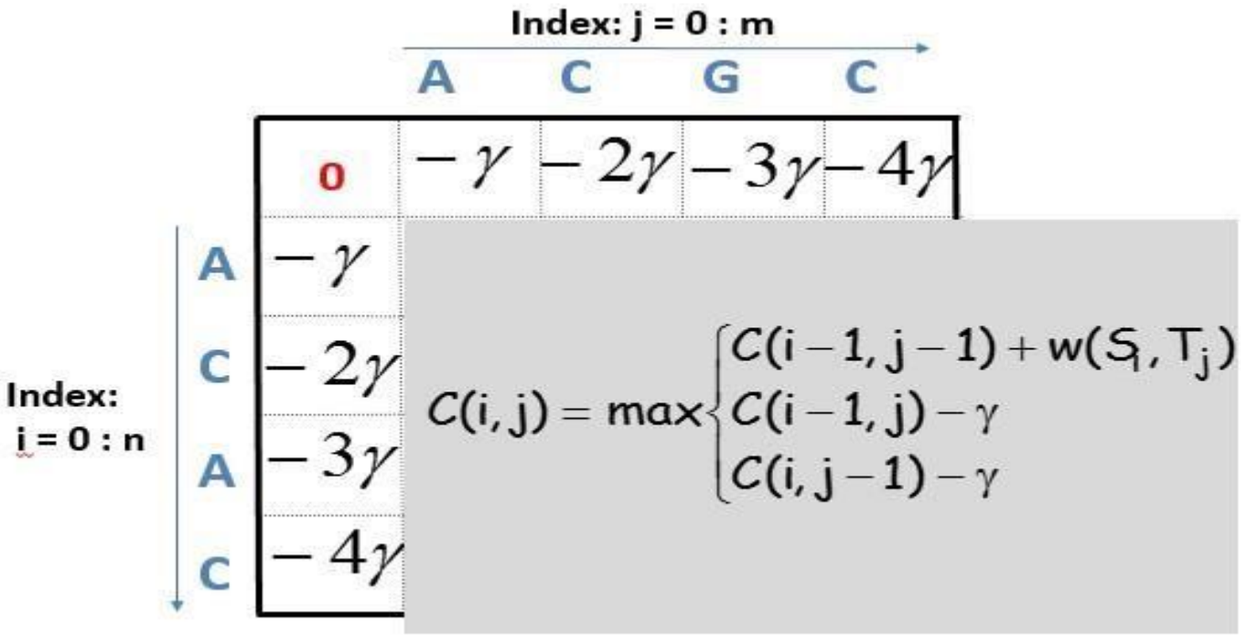


Figure 15 maximum score element is selected from all three sides comparison

The terms for match, mismatch are:

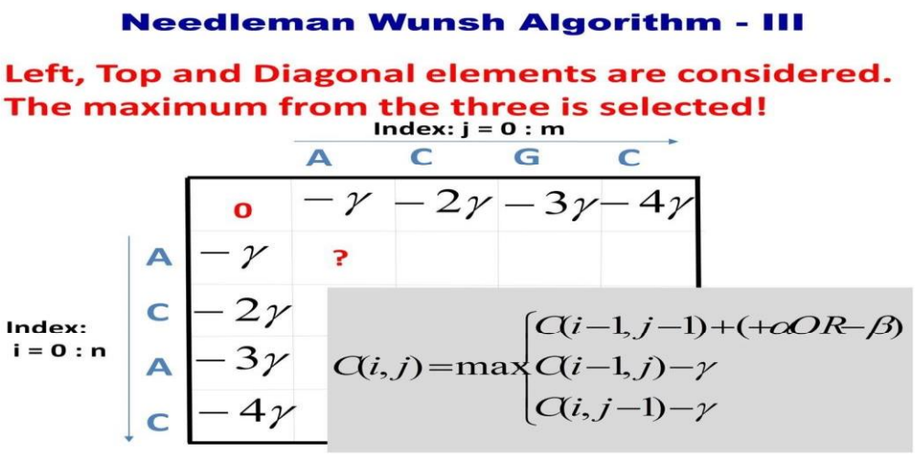
α Alpha = Match reward
β Beta = Mismatch penalty
γ Gamma = Gap penalty

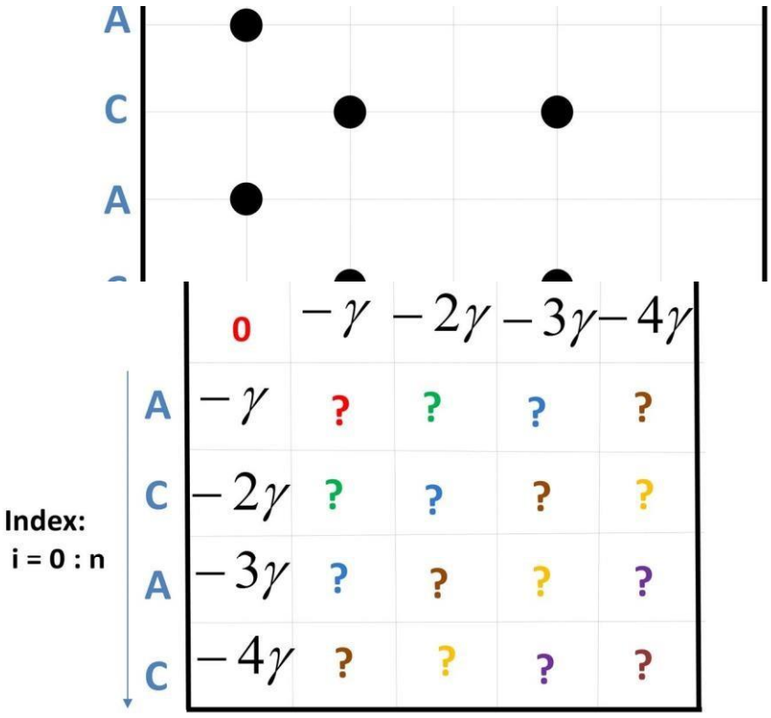
The matrix is computed progressively until the bottom right element

37 Needleman Wunsh Algorithm-III

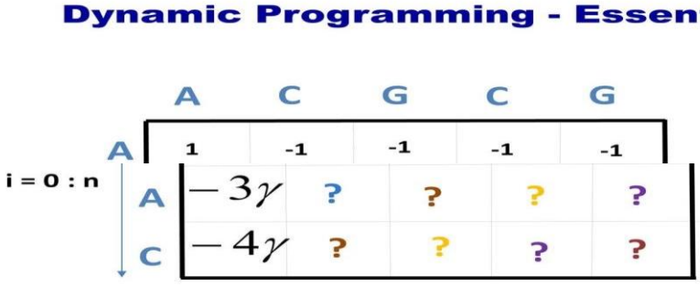
Background

- Top, Left and Diagonal elements are considered to calculate an element in the matrix
- The matrix is computed progressively until the bottom right element





Compare “ABCD” with “EFGH”



Conclusion

- We are trying to use the best score from Top, Left and Diagonal elements
- This strategy will be very useful in selecting the best combination of alignments!

38 Needleman Wunsh 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								

	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.

39 Backtracking Alignments

Background

- Top, Left and Diagonal elements are considered to calculate an element in the matrix
- Match, mismatch and gap penalty is used to compute score from each position!

Introduction

- How can we use the Needleman Wunsh algorithm to find the optimal alignment?

- The solution lies in a method called the “Traceback”
  - Let’s see how a traceback works

Backtracking Alignments

+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								

Backtracking Alignments

	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

Conclusion

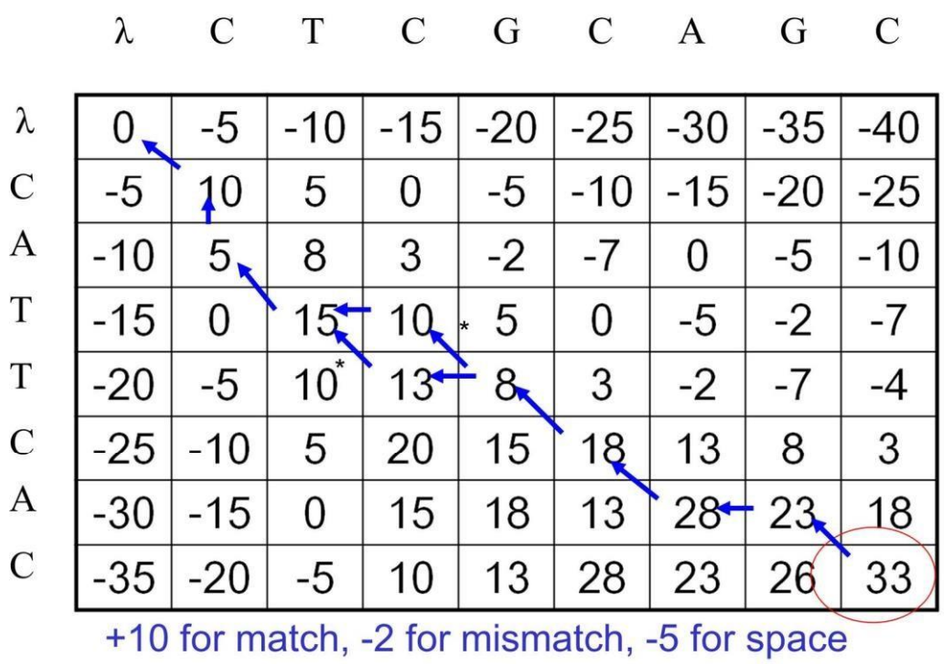
- After completely calculating the matrix, we need to do a “traceback”
- Traceback is such that we start from the bottom right and select the element from top, left or diagonal which led to the starting element

40 Global and Local Alignments

Background

- We have filled the alignment matrix using Needleman Wunsh Algorithm
- Traceback allows us to extract the optimal alignment

Revisiting Global and Local Alignments



- But, is it the local or global alignment?

The difference?

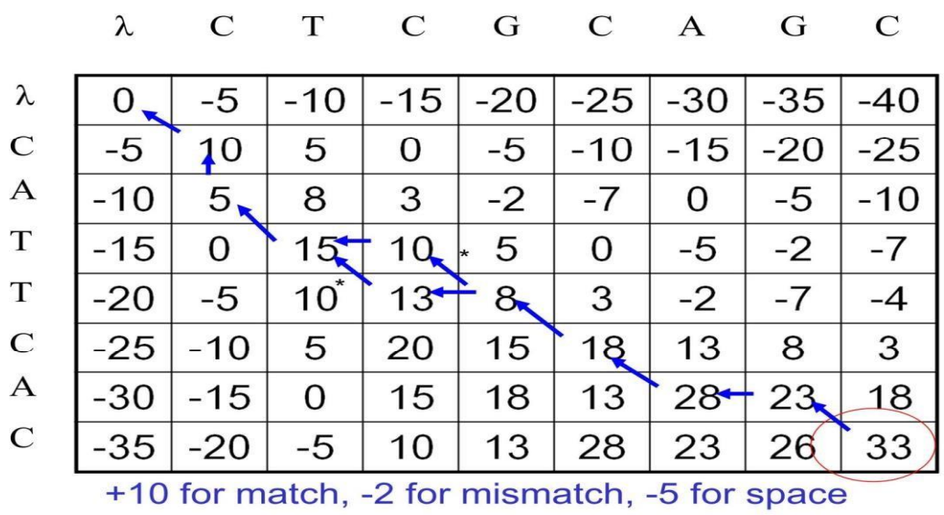
- Needleman Wunsh algorithm traceback begins from the bottom right element
- Proceeds progressively until we reach the top left
- This provides us with a global alignment

Skipping the gaps

- What if we have a different traceback strategy?

- Can we start from any position within the alignment matrix to avoid gap regions?

Revisiting Global and Local Alignments



- Yes, we can and that will give us a local alignment

Conclusion

- Traceback strategy allows us to differentiate between a local and a global alignment

- The Smith Waterman Algorithm allows us to elicit local alignments and we will look at it later.

41 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.

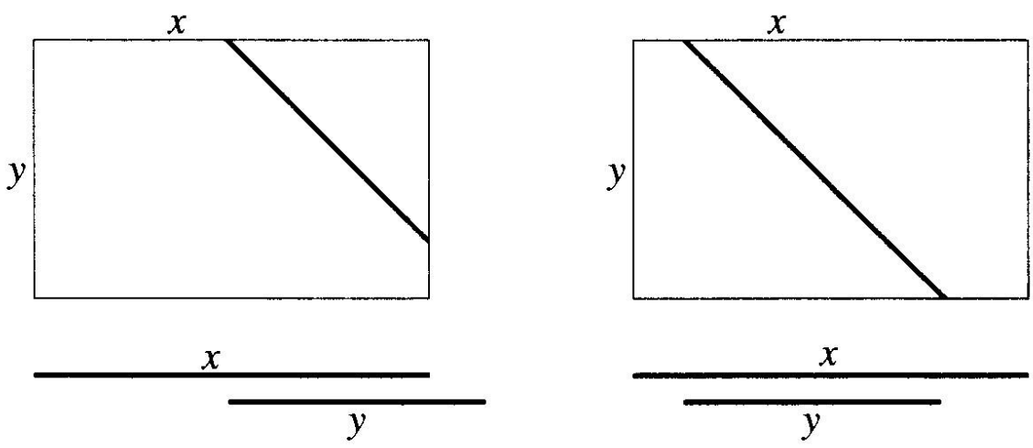


Figure 21 leading and trailing edge mismatches versus global alignment by gap-insertion (stretching) of sequences

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

42 Example of Overlap Matches

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.

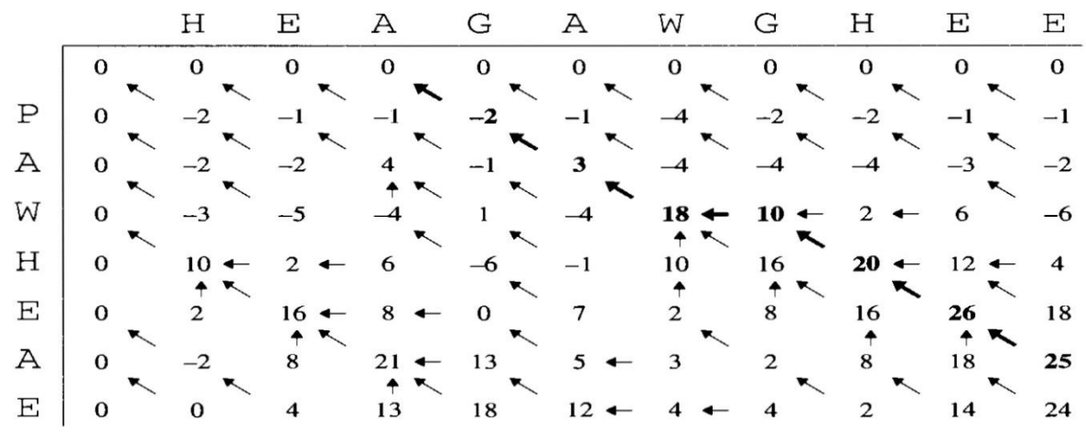


Figure 23 Traceback in amino acid sequence alignment

Scoring stagey is:

Match = +2. Mismatch = -1, Gap = -2 Sequences are:

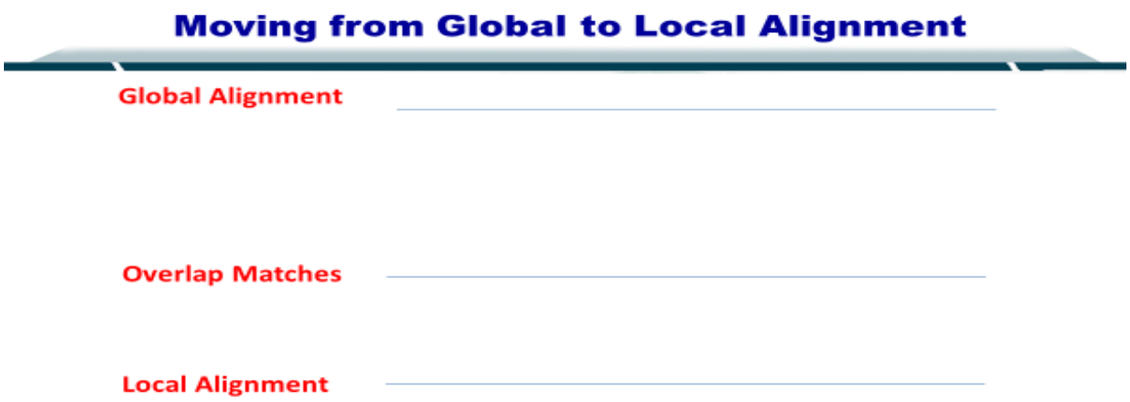
GAWGHEE

PAW-HEA

43 Moving From global to local alignment

Introduction

- **Local alignment may then need to be computed between two such sequences**
- **Let's take a look at why local alignments are useful in real life biology**



Conclusion

- Local Alignments can identify exons which are present in both sequences
- Proteins of different kind and of different species often exhibit local similarities
- Hence, local similarities may indicate "functional subunits"

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.

44 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.

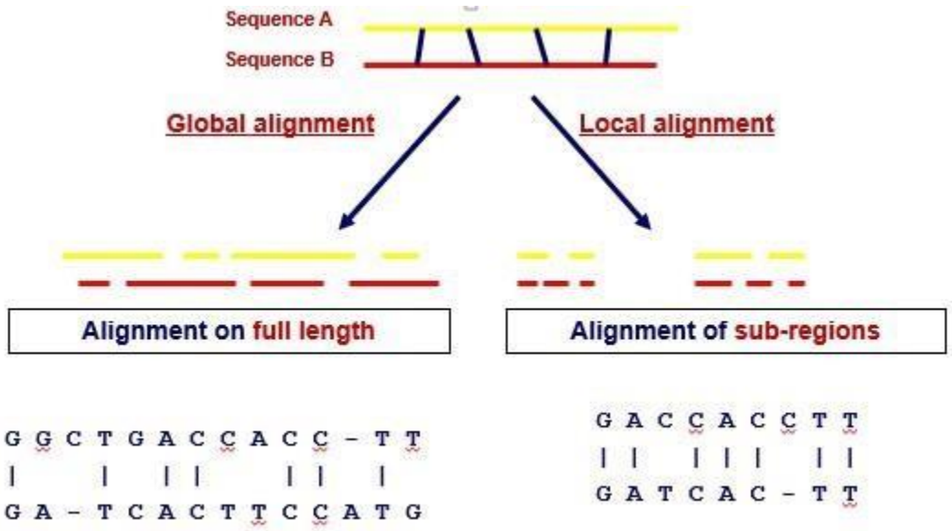


Figure 24 Global and local sequence alignment comparison.

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.

45 Example of Smith Waterman Algorithm

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

Example of Smith Waterman Algorithm

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

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

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

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

Example of 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

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

46 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.

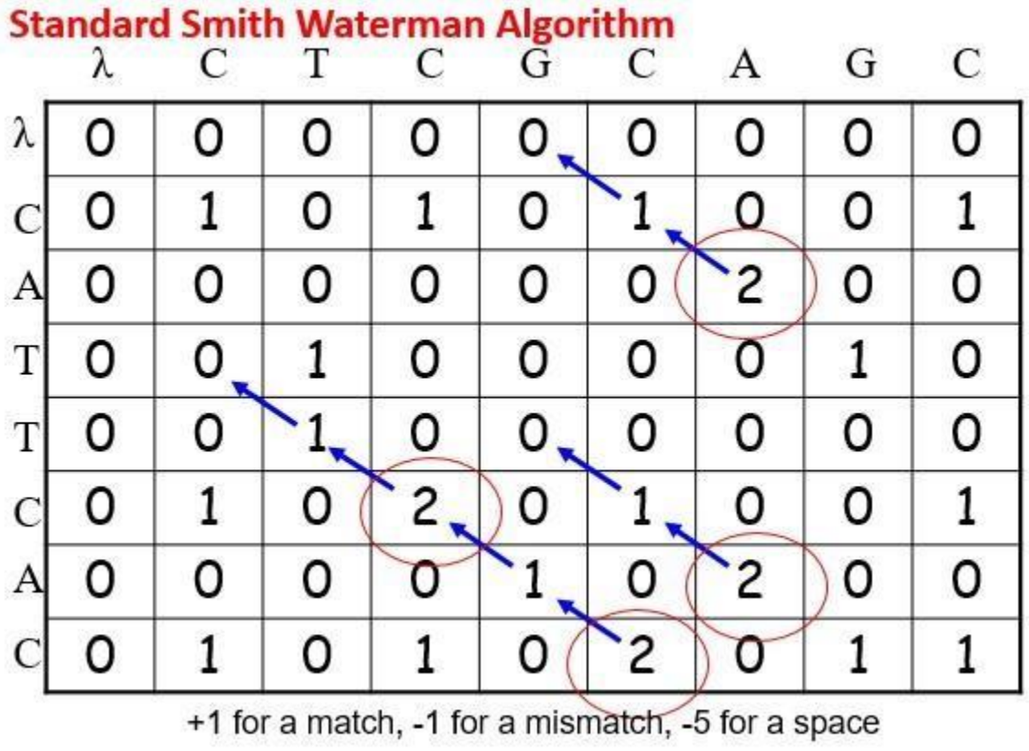


Figure 27 (-5) is threshold score in table

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

47 Exmaple of repeated alignment

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.

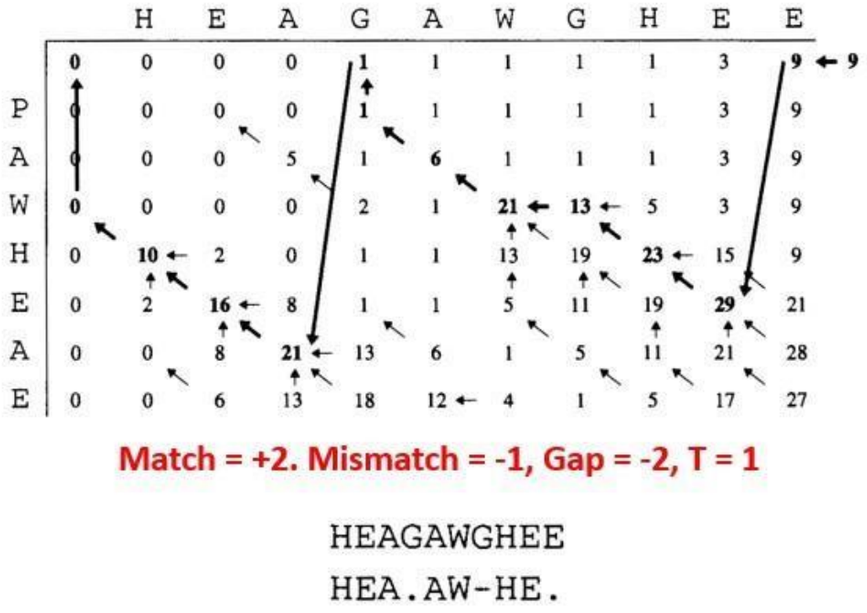


Figure 28 Trackback from different sides to find maximum or Threshold (T) Score.

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”

48 Review of Traceback Strategies for Global, Overlap, Local and Repeated Alignments

Background

- We have seen how biological sequences can be searched and compared using various recurrence relations and traceback strategies
- Let's review them!

Review of Traceback Strategies

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

Review of Traceback Strategies

Overlap Matches

		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

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

GAWGHEE

PAW-HEA

Review of Traceback Strategies

Repeated Alignments

		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	**1**	1	1	1	1	3	**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.

Conclusion

- Slight modification in the recurrence relation and changes to the traceback strategy can help compare sequences in a variety of ways
- DynamicPrograming solves this problem in quick time!

49 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.

50 Measuring alignments by scoring

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.

51 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.



Figure 0.12 Ubiquitin Protein where amino acids matching

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

52 Deriving scoring matrices

Each amino acid have different property.

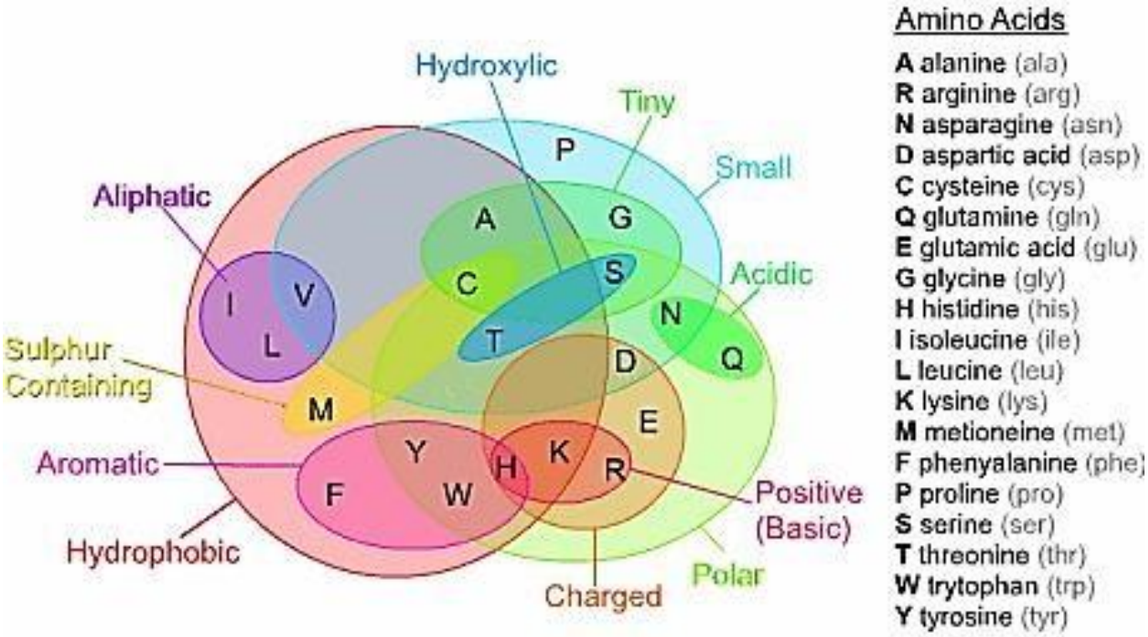


Figure 0.13 properties of amino acids (Image Esquivel et al. (2013)

And each amino acid have different frequency.

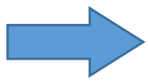
Table 7.1 The average amino acid frequencies reported by Robinson and Robinson in [173].

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.

53 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’= $(PAM1)^n$

54 BLOSSUM 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.

55 Introduction to Multiple Sequences 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.

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.

56 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

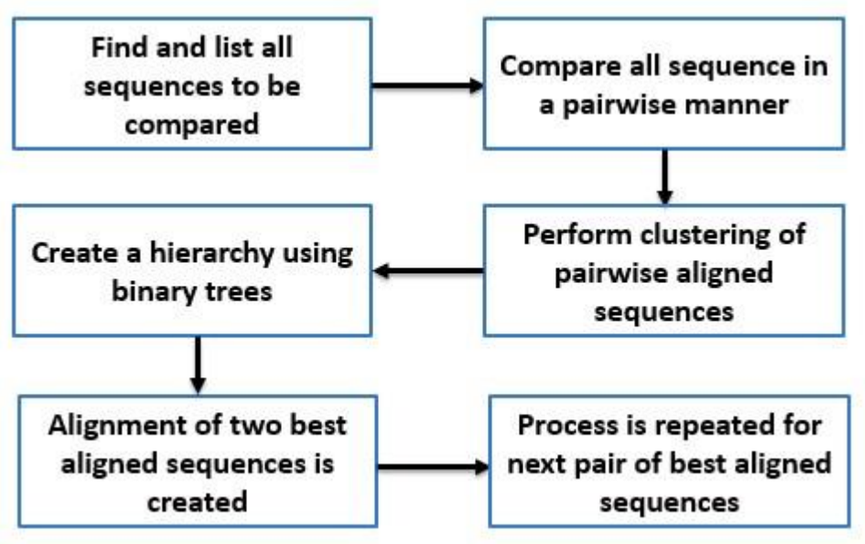


Figure 0.17 Methodology

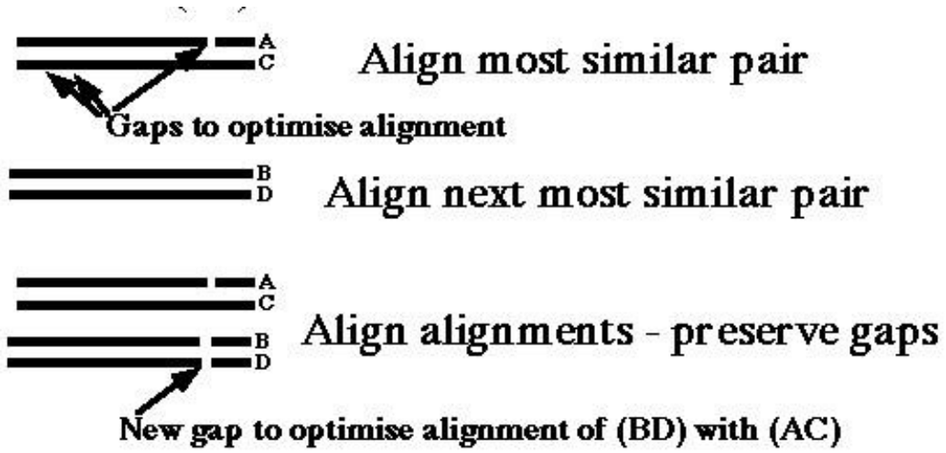


Figure 0.18 sequence alignments

CONCLUSION

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

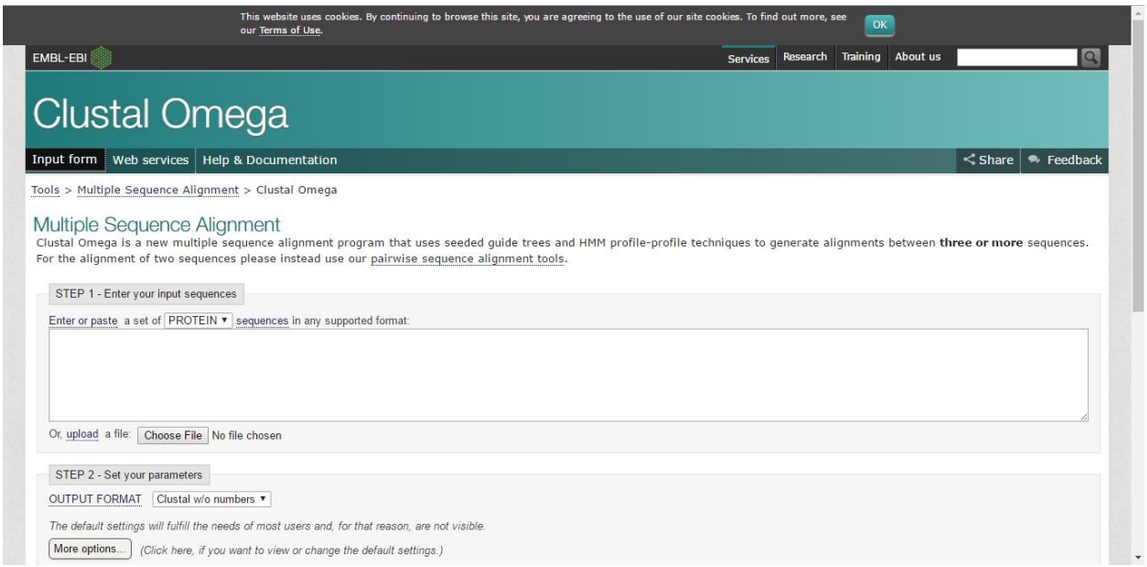


Figure 0.19 CLUSTAL – Online tool

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

57 Progressive Alignment for MSA

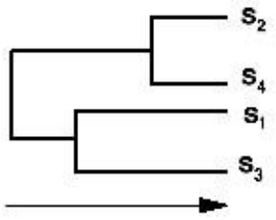
MSA involves progressive alignment of sequences. Doing so many progressive alignments can be slow.

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.*



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

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

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

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

	v_1	v_2	v_3	v_4
v_1	–			
v_2	.17	–		
v_3	.87	.28	–	
v_4	.59	.33	.62	–

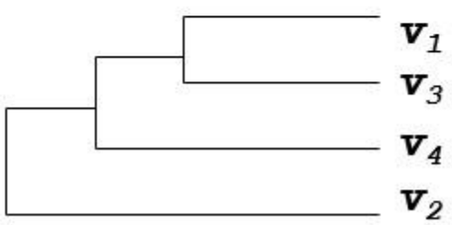


Figure 0.20 Similarity Matrix

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.

58 MSA Example

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

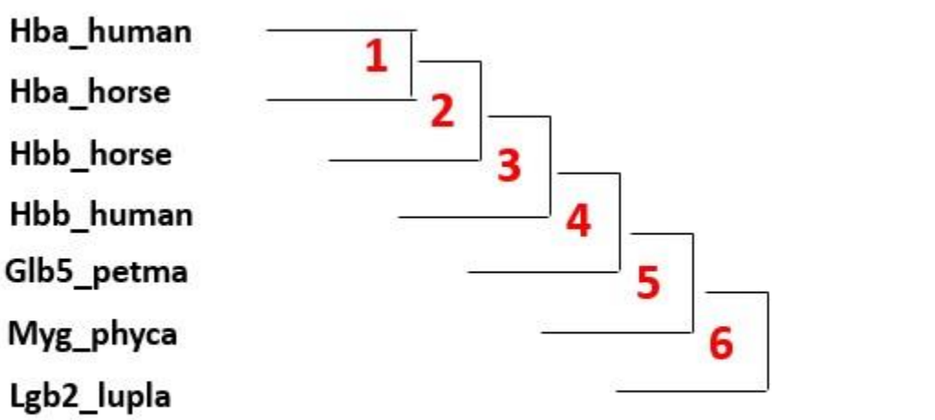
For example:

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

		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

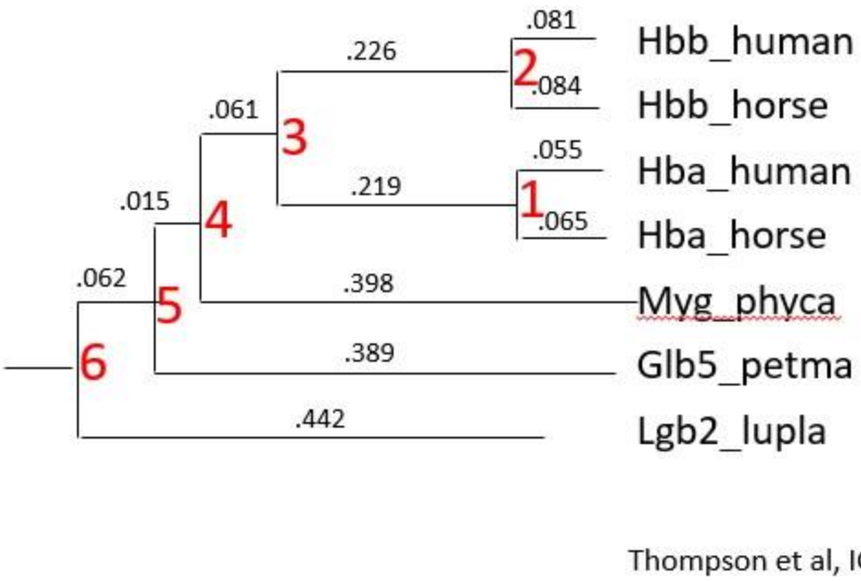


Figure 0.23 Progressive alignment following a guide tree

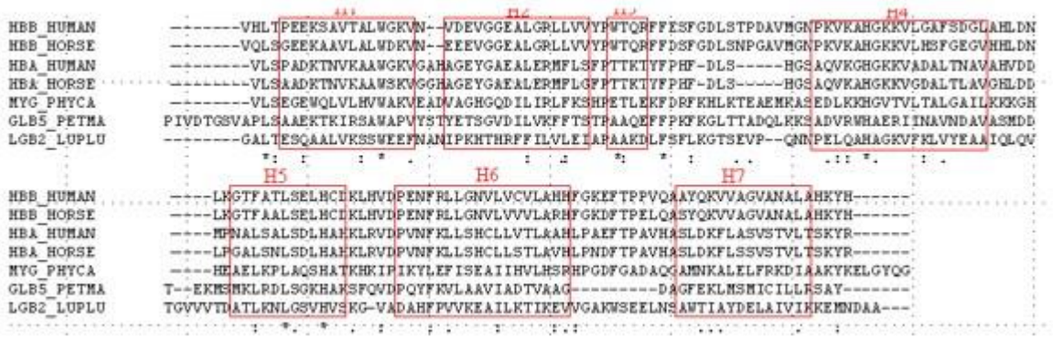


Figure 0.24 Alignment results

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!

59 CLUSTAL

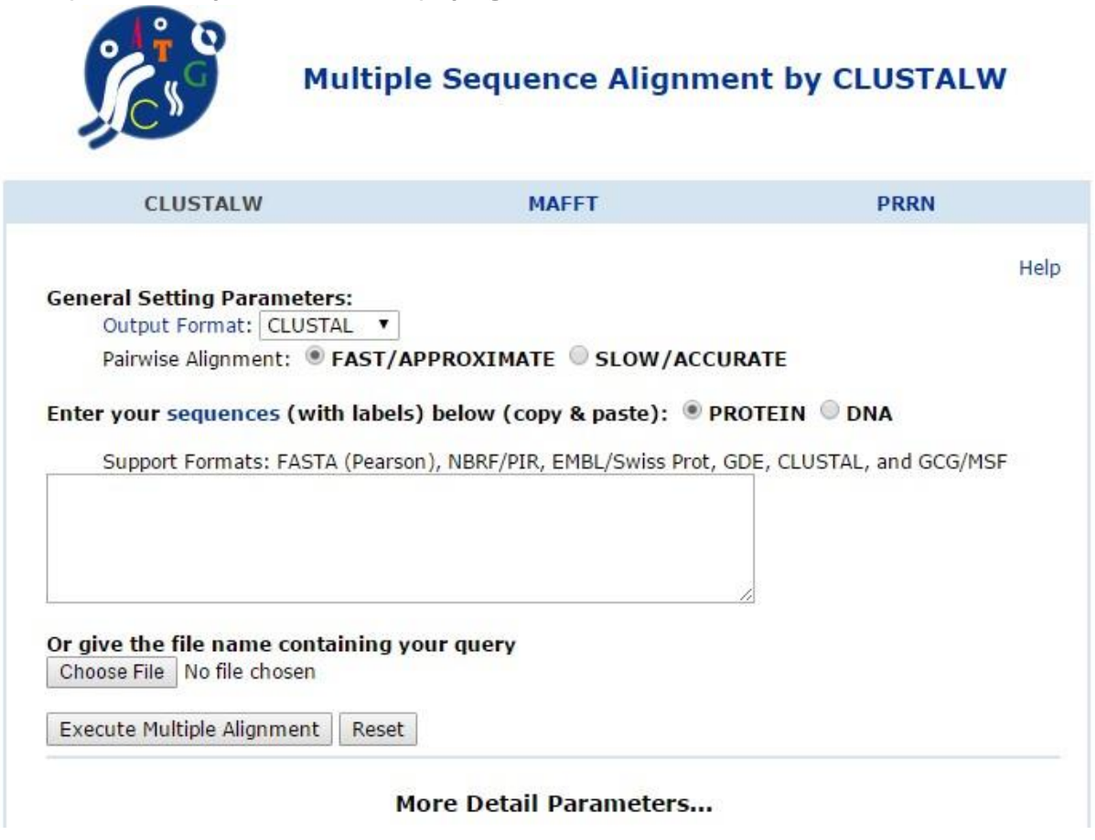
MSA involves progressive alignment of sequences. Doing so many progressive alignments can be slow. CLUSTALW is an online tool to perform MSA.

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



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

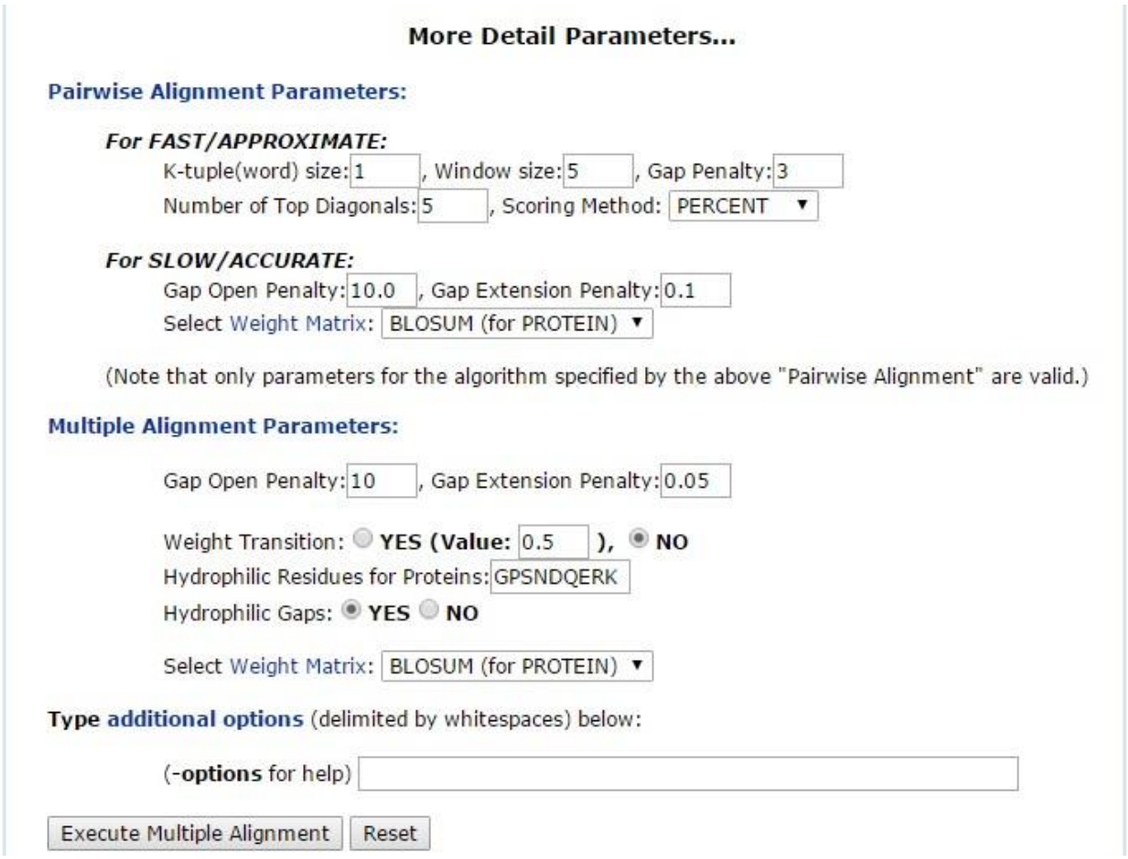


Table 1. Some recent and less recent available methods for MSAs.

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

60 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®

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 GO

- 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 | blastp | blastx | tblastn | tblastx

Enter Query Sequence

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

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

Query subrange

From

To

Or, upload file Choose 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: YouTube Create custom database

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

61 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



Figure 0.25Input to BLAST: Gene IDs

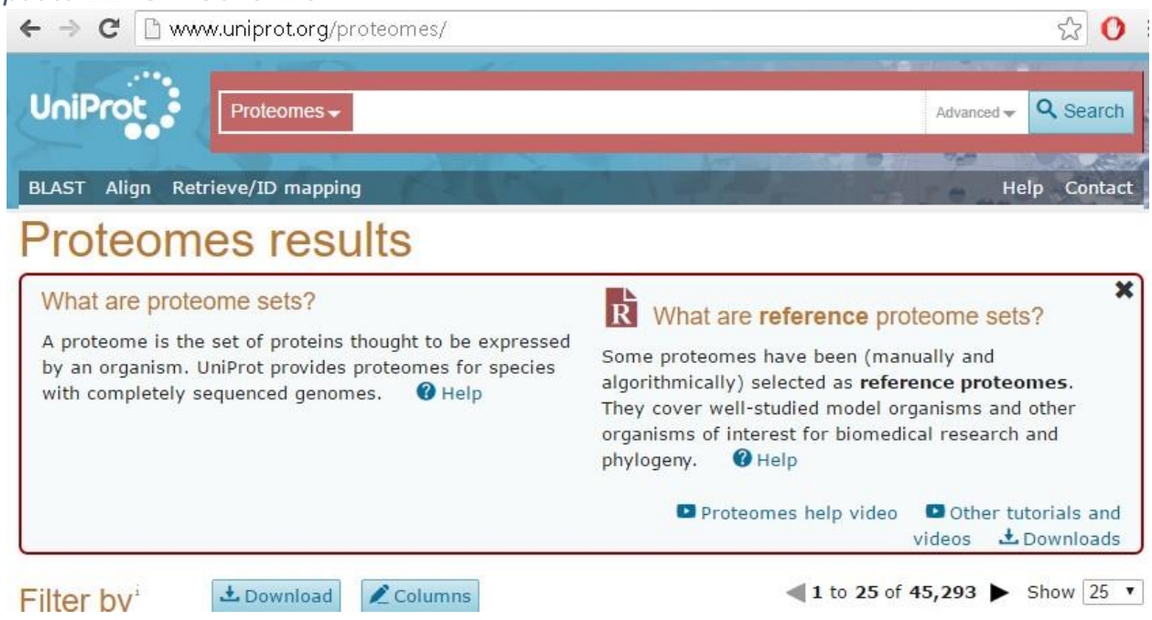


Figure 0.26 Input to BLAST: Protein IDs

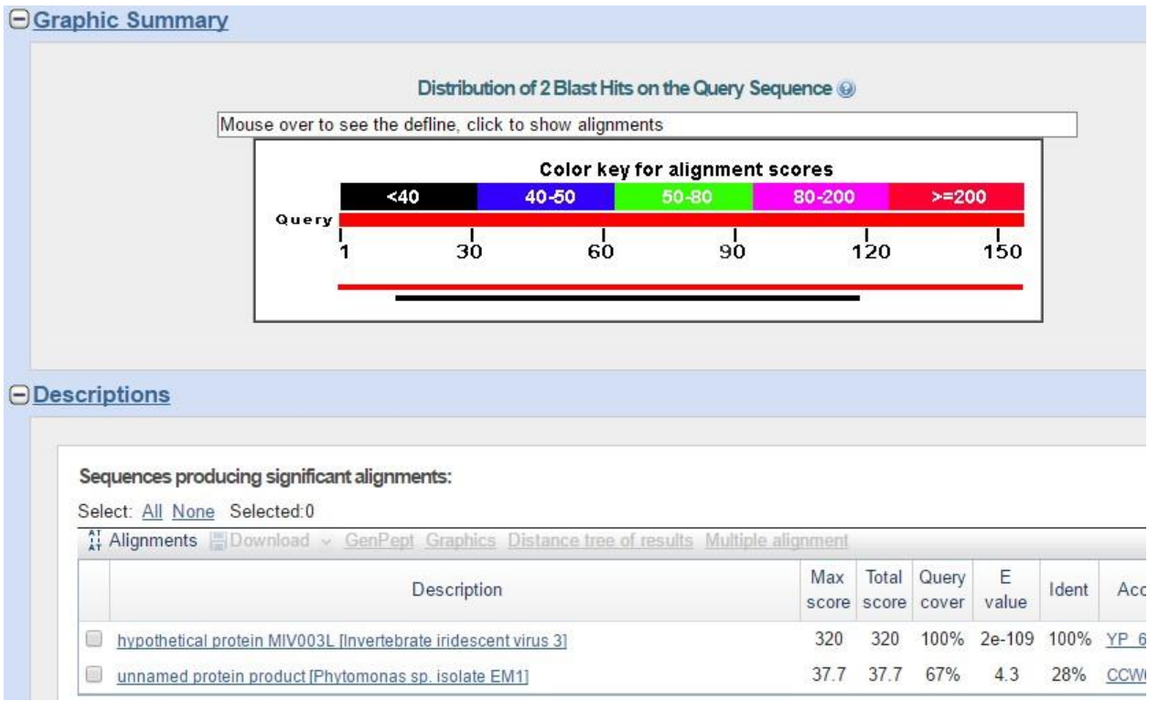


Figure 0.27 Results from BLAST

62 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

63 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 sixframe 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

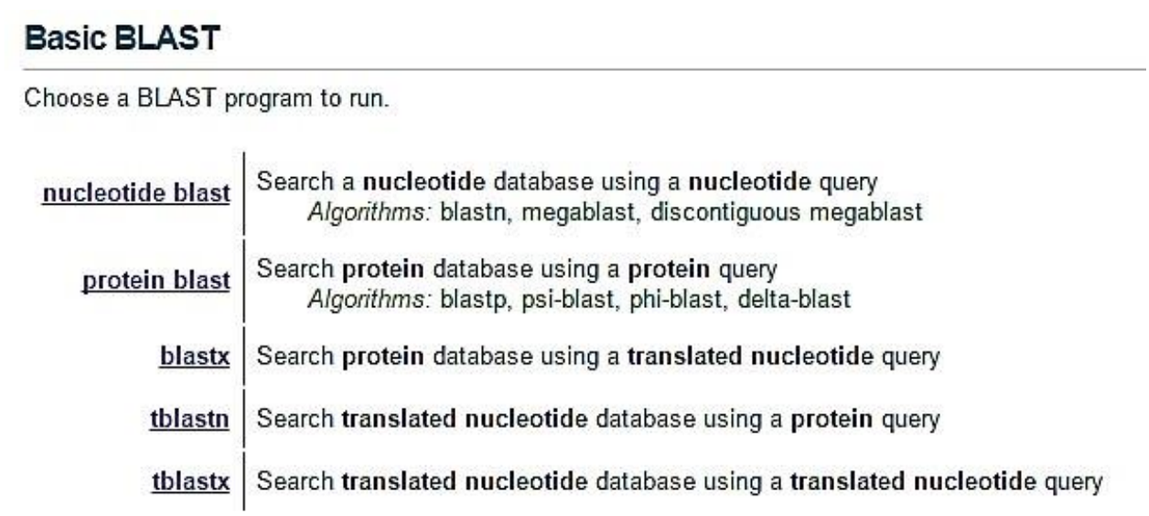
64 Summary of BLAST

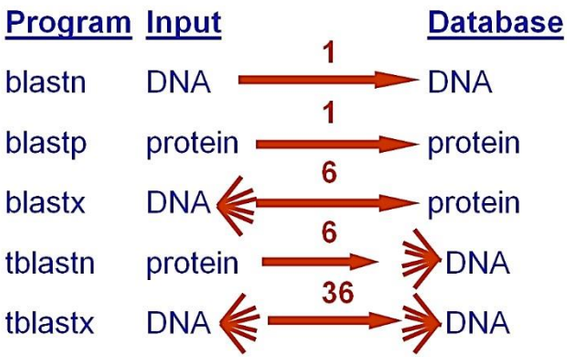
BLAST can search sequence databases and identify unknown sequences by comparing them to the known sequences. This can help identify the parent organism, function and evolutionary history.

Step1: obtain a query of sequence

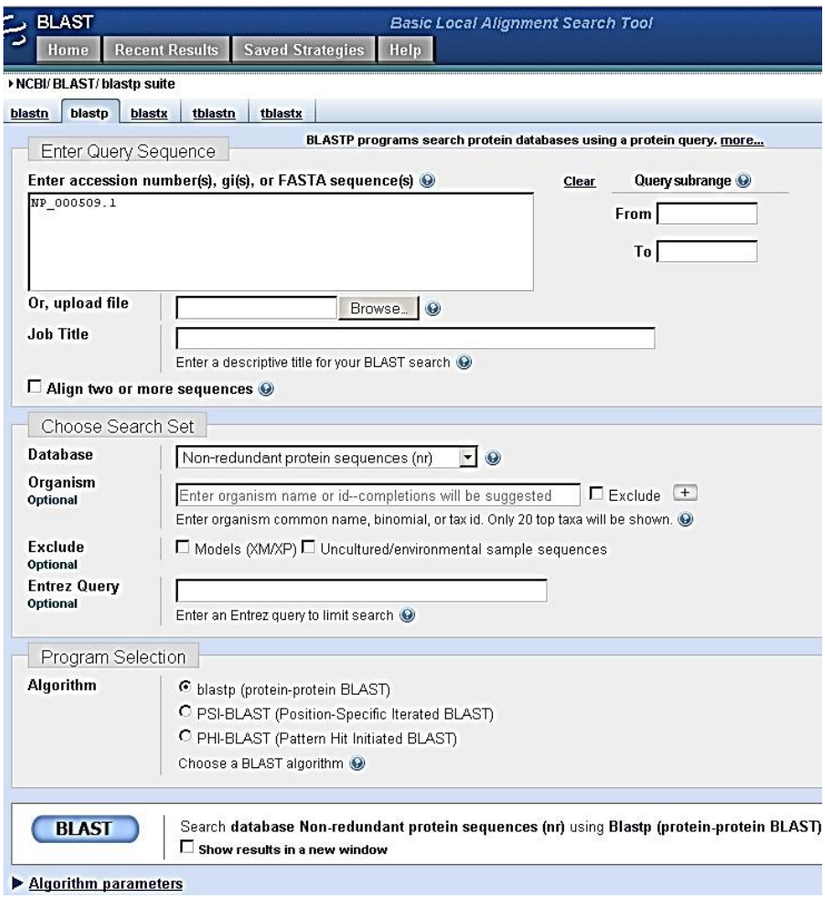


Step2: choose a type of BLAST

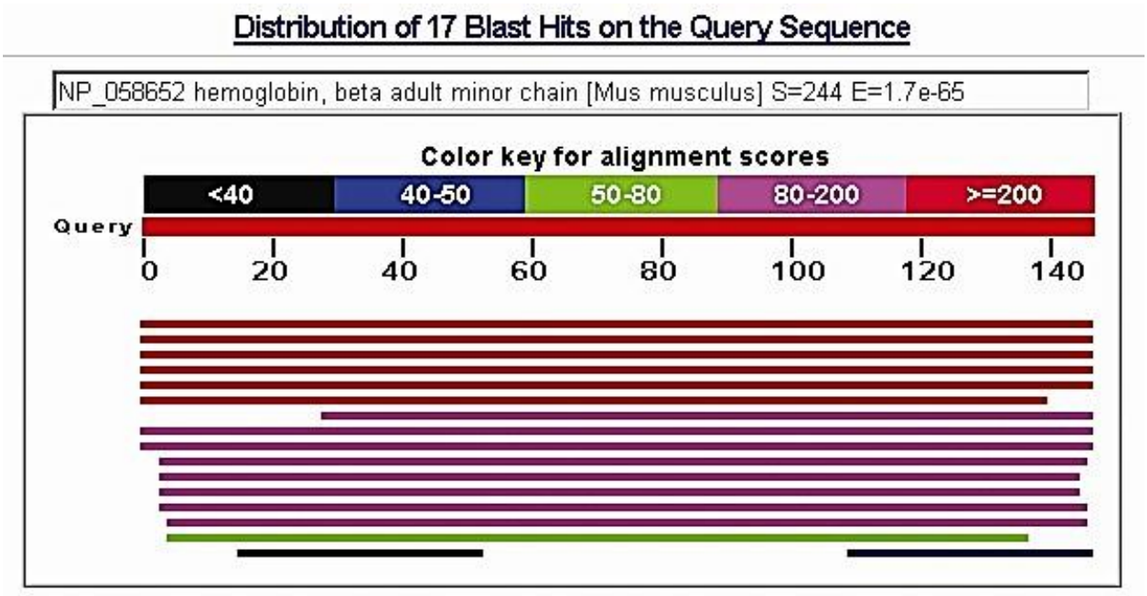


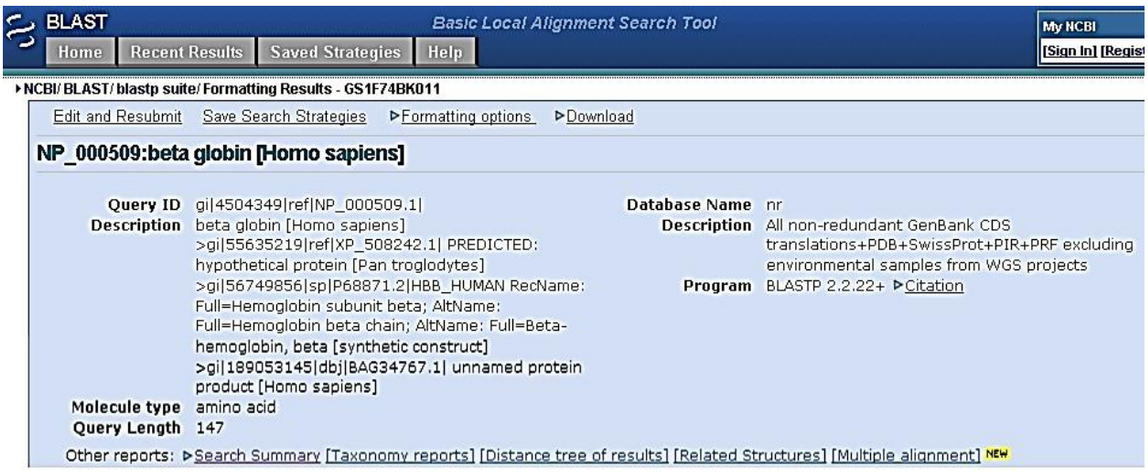


Step3: search parameter



Step4: tabulated search results





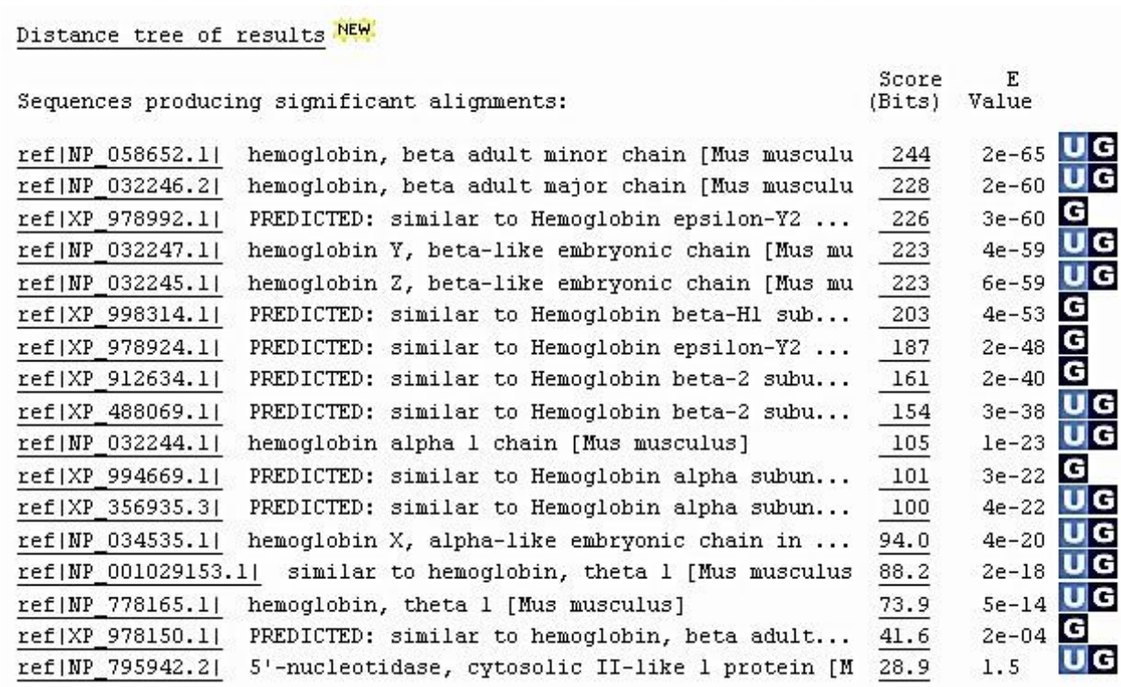


Figure 0.28 tabulated search results

65 Introduction to FastA-I

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.

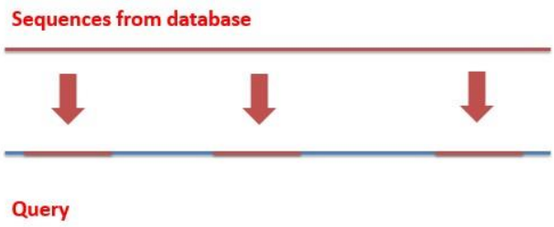
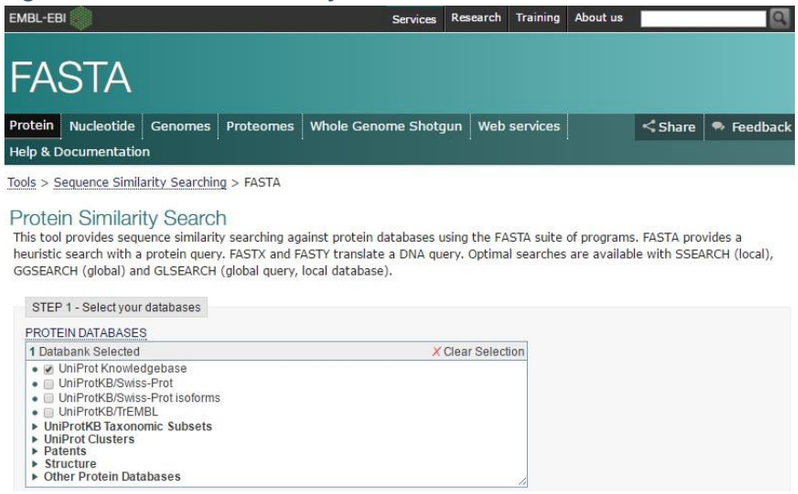
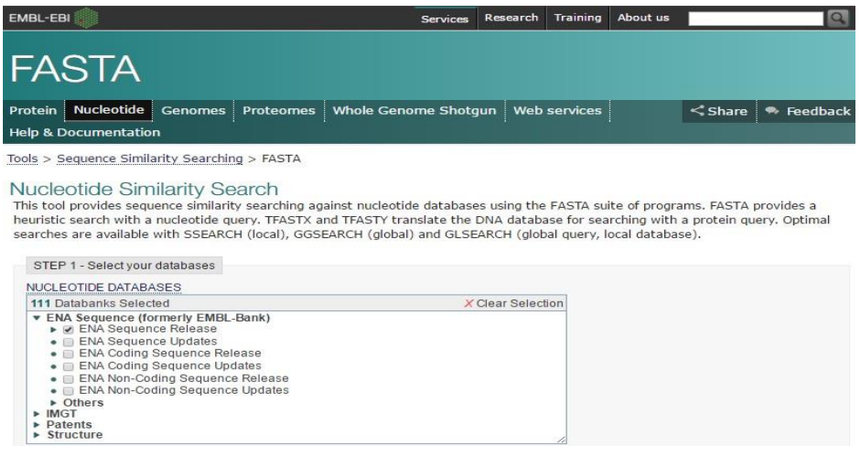
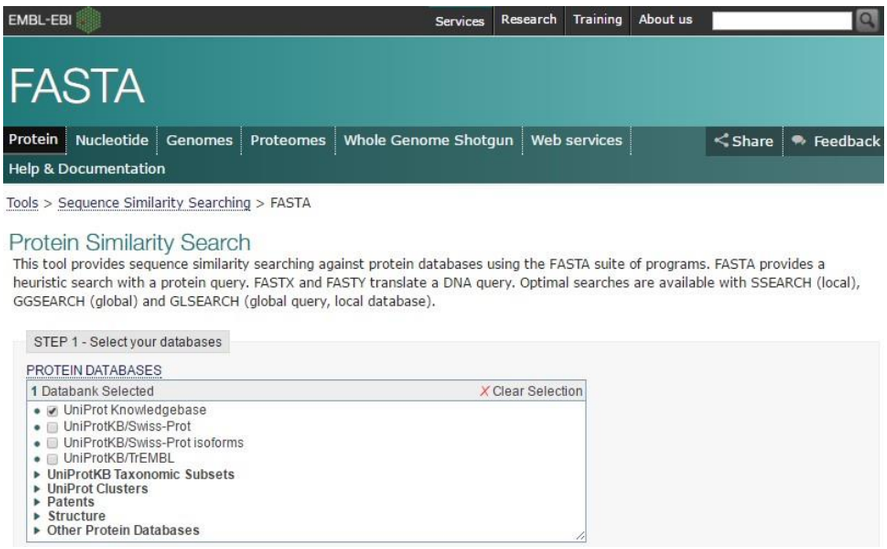


Figure 0.29 Regions of absolute identity



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





66 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.

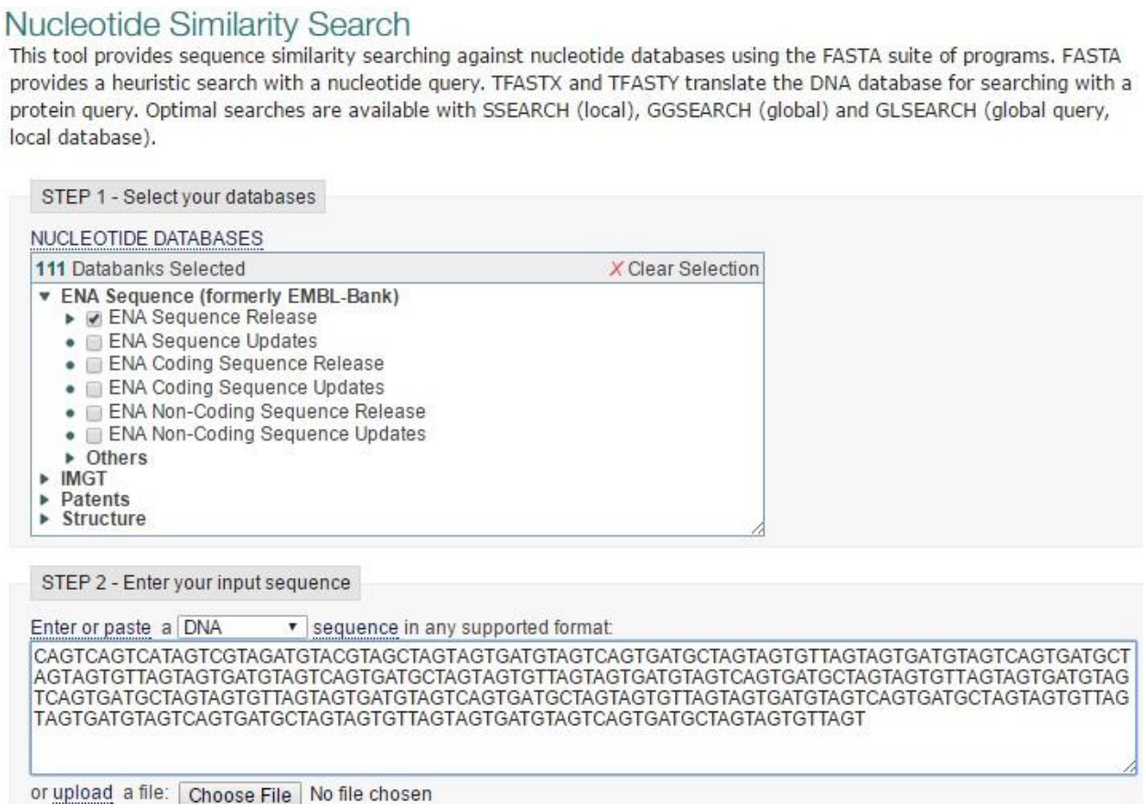


Figure 0.32 Input to FASTA: Gene IDs

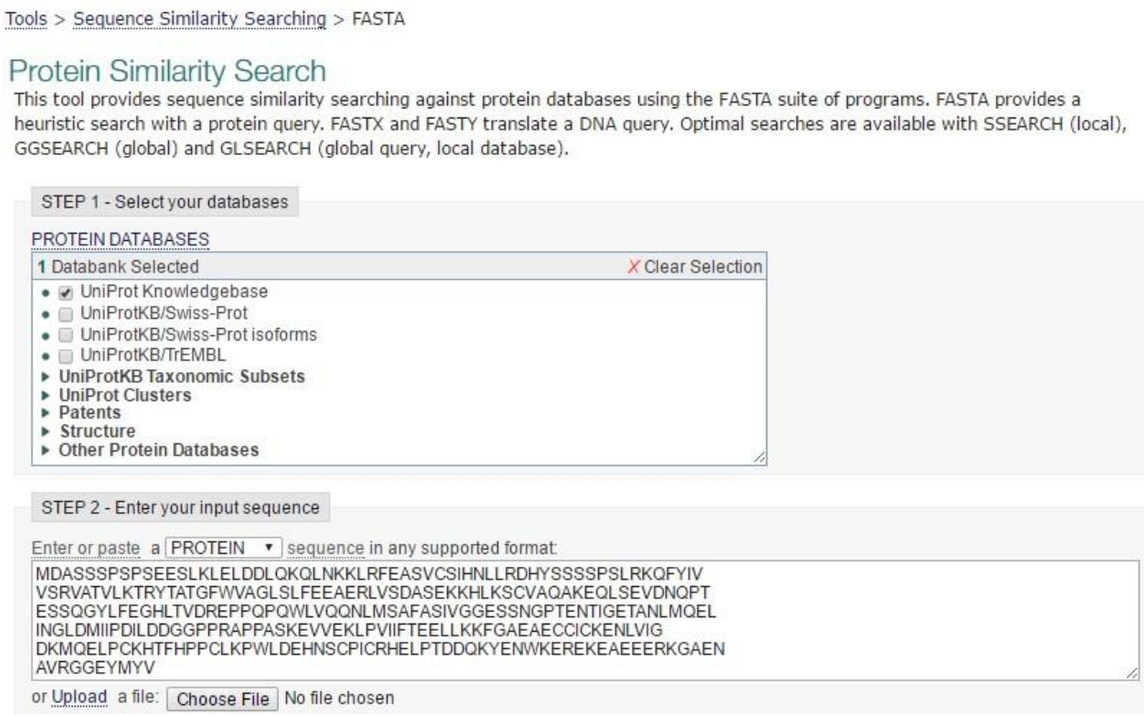


Figure 0.33 Input to FASTA: Protein Sequence

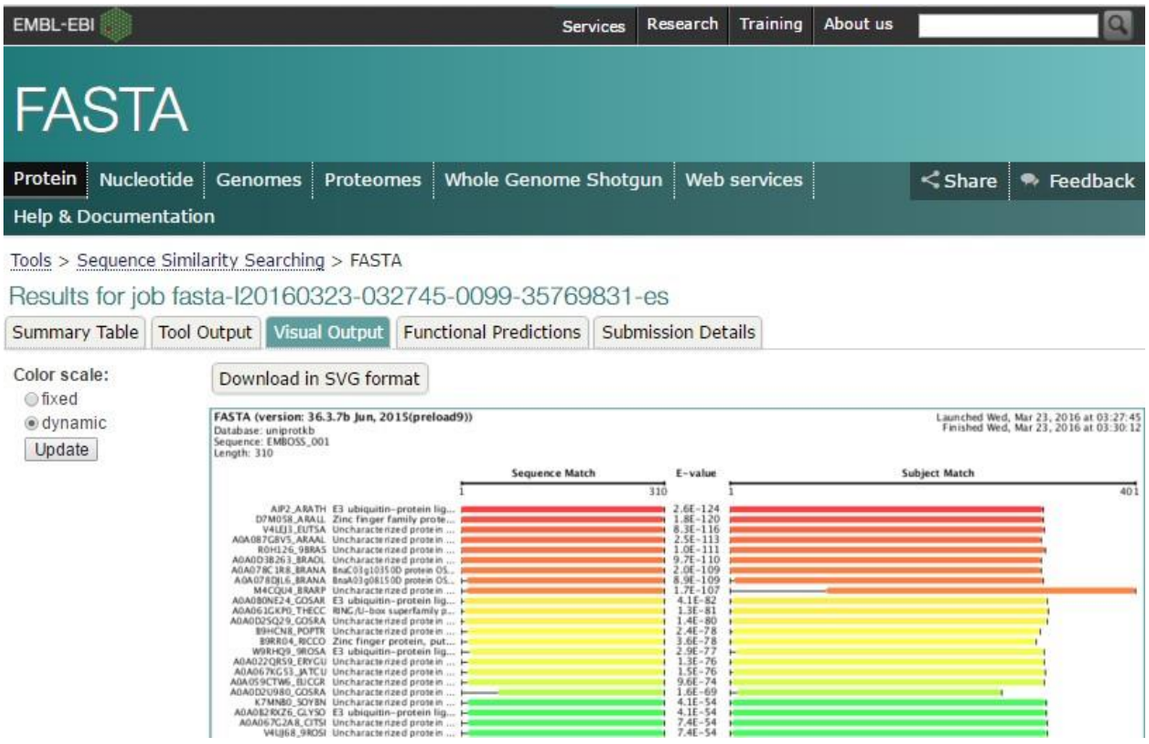
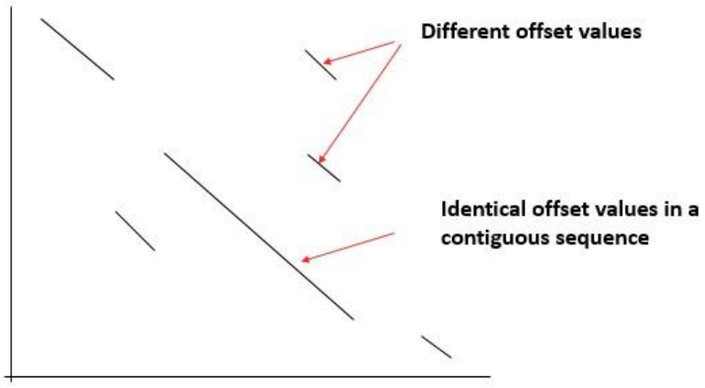


Figure 0.34 Results from FASTA

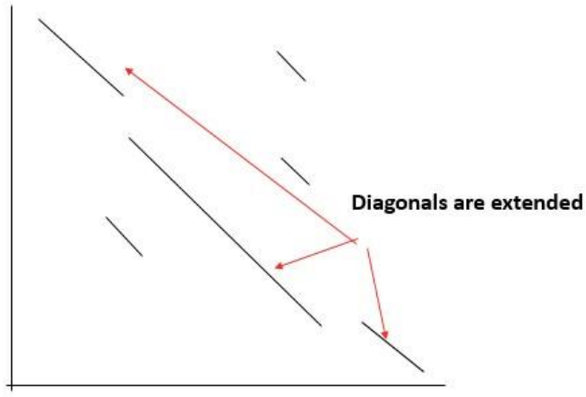
67 FASTA Algorithm

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.

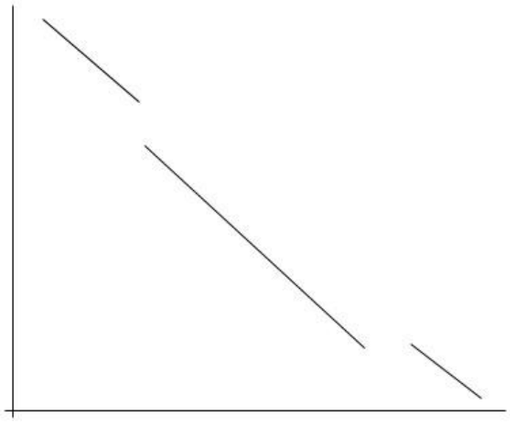
STEP1: Local regions of identity are found



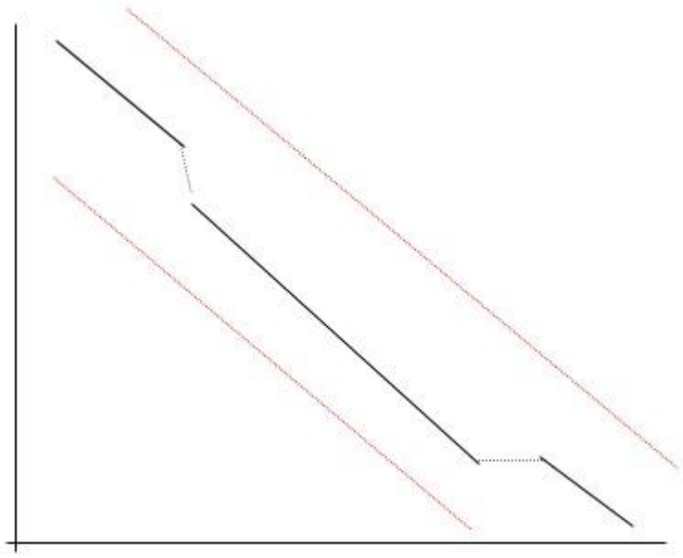
STEP2: Rescore the local regions using PAM or BLOSUM matrix



STEP3: Eliminate short diagonals below a cutoff score



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



68 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)

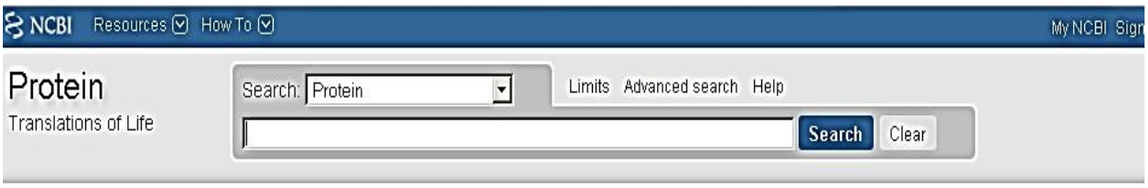
- **fasty35**

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

69 Summary 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



Display Settings: FASTA Send to: Change region shown

hemoglobin subunit beta [Homo sapiens]

NCBI Reference Sequence: NP_000509.1

GenPept Graphics

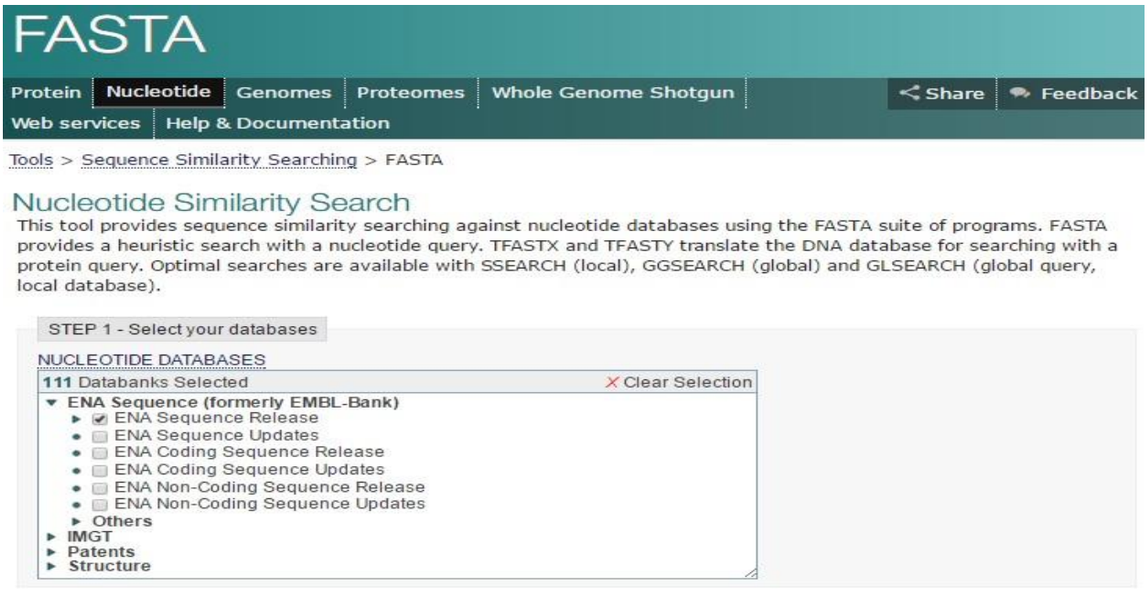
```
>gi|4504349|ref|NP_000509.1| hemoglobin subunit beta [Homo sapiens]
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKVKAHGKKVLG
AFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVAN
ALAHKYH
```

Analyze this sequence

Run BLAST

Identify Conserved Domains

Find in this Sequence



`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 0.36 Step 2: Choose a type of FASTA

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

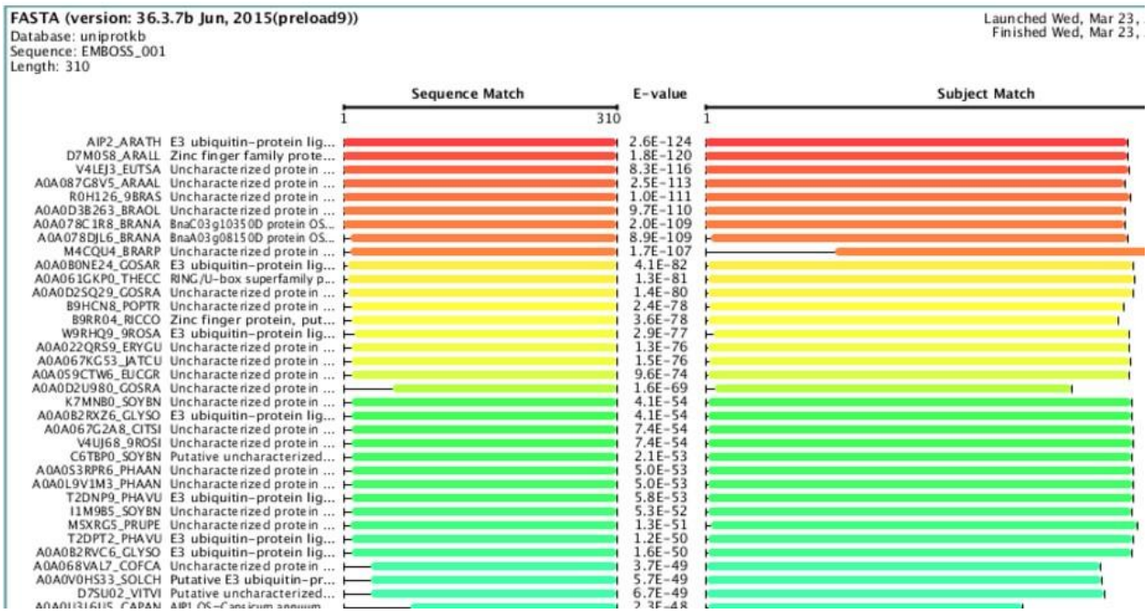


Figure 0.39 Step 4: Tabulated Search Results

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

70 Biological Databases 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.

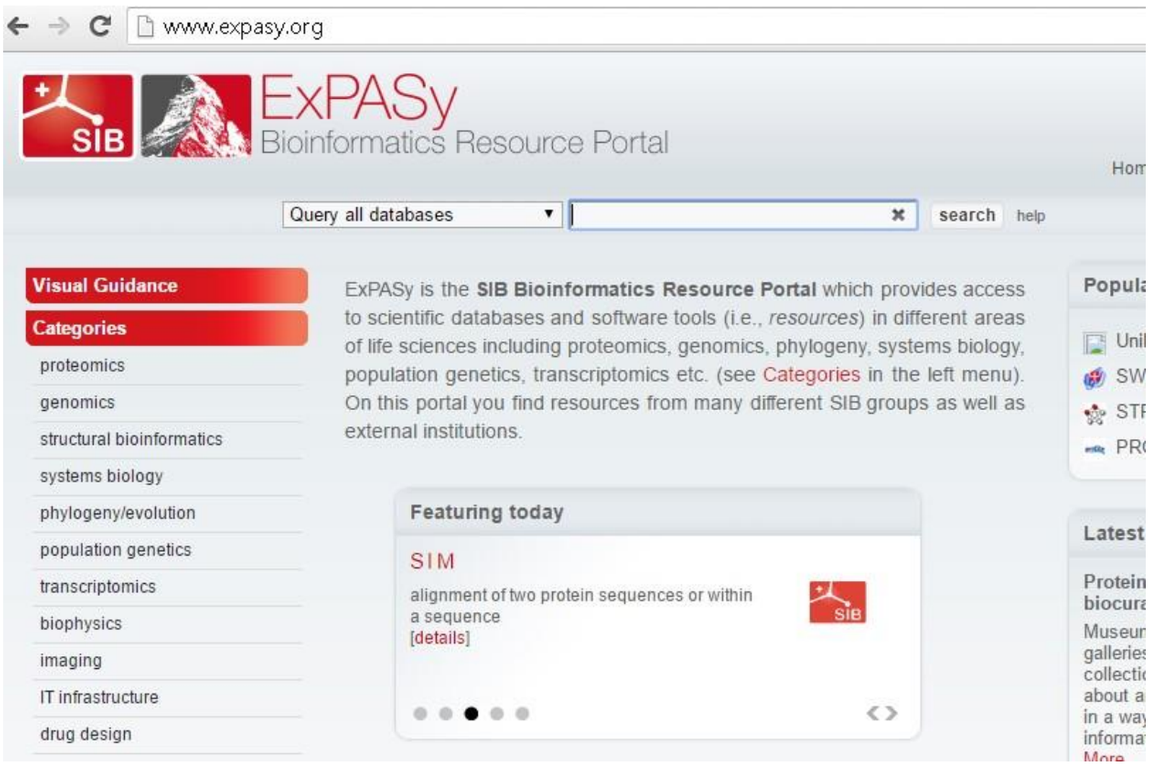
OBJECTIVE

- 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

71 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/**



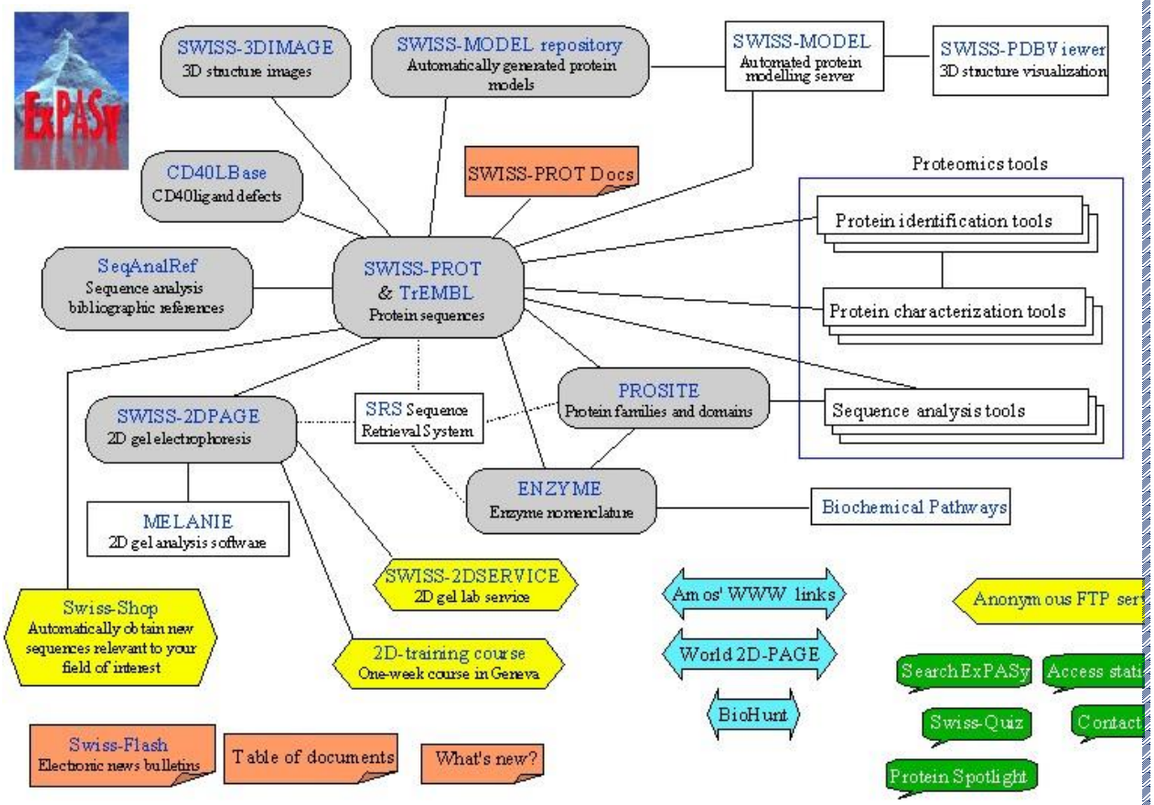


Figure 0.41 flowchart

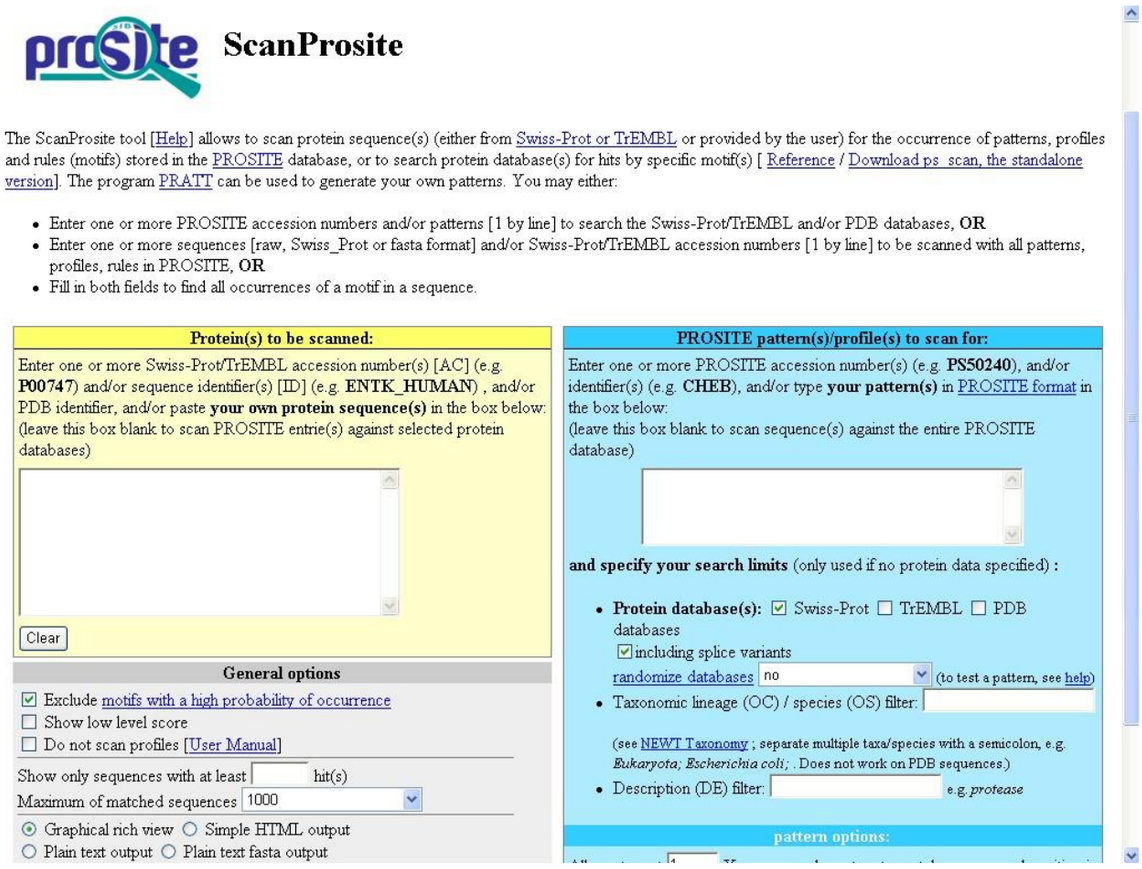


Figure 0.42 prosite scanning section

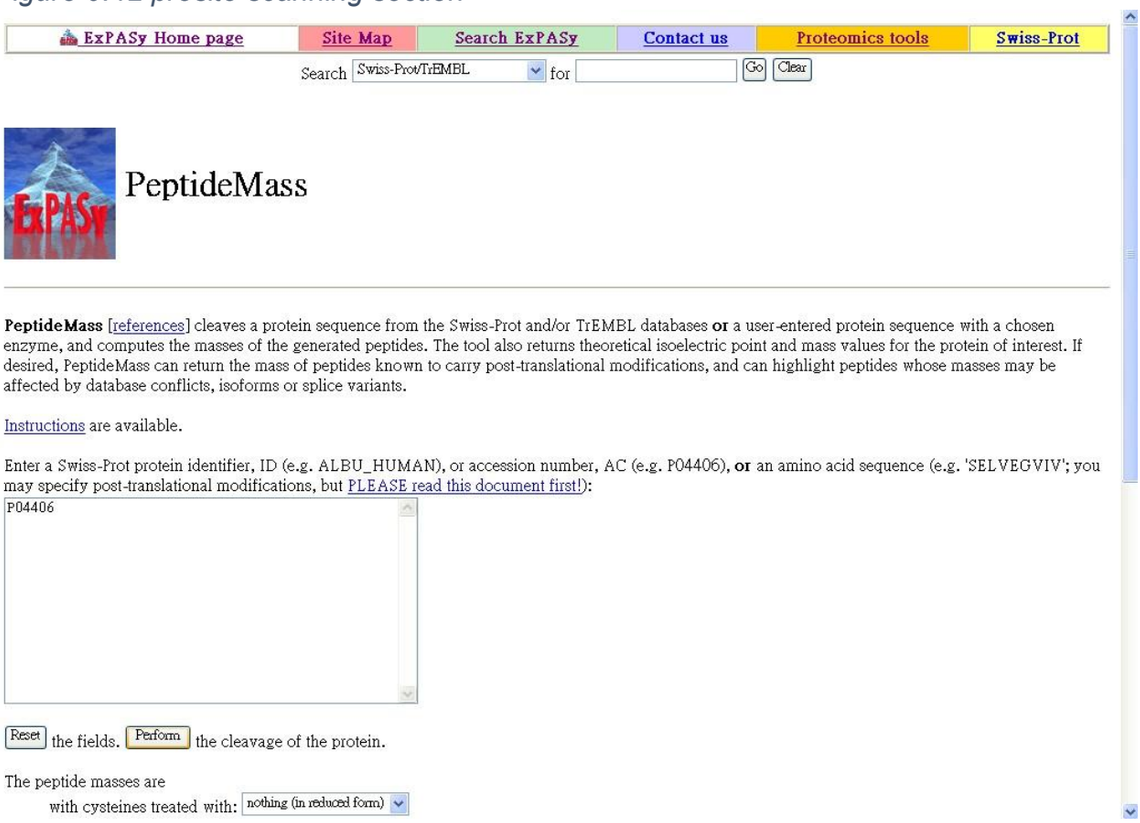


Figure 0.43 peptide mass finding

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:

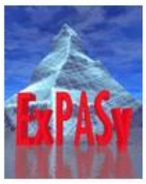
- **ProFound** - a tool for searching a protein sequence database using information from mass spectra of peptide maps. A Bayesian algorithm is used to rank the protein sequences in the database according to their probability of producing the peptide map.
- **PepFrag** - a tool for searching protein or 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

72 Uniprot, SwissProt

Both UniProt and SwissProt are the online database for proteins.

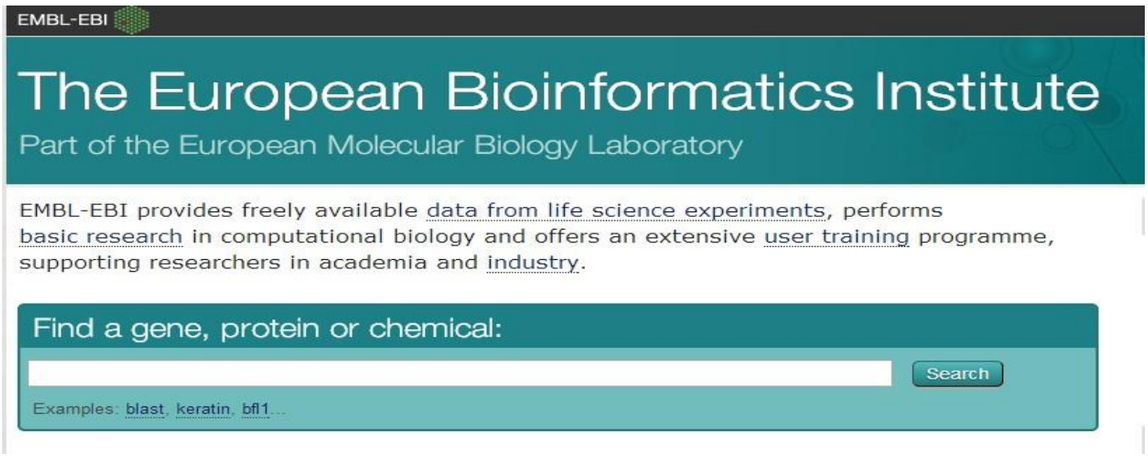


Figure 0.46 gene, protein or chemical can be find

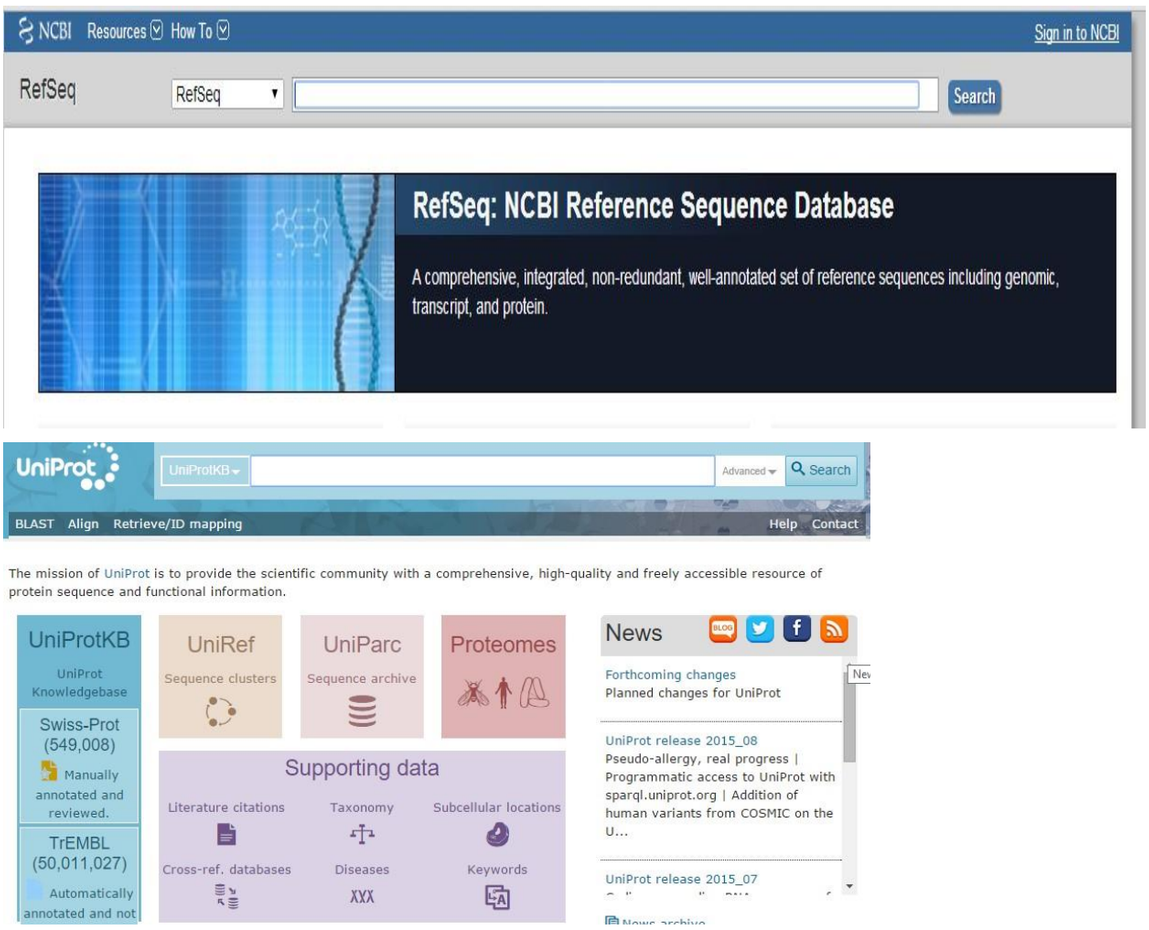
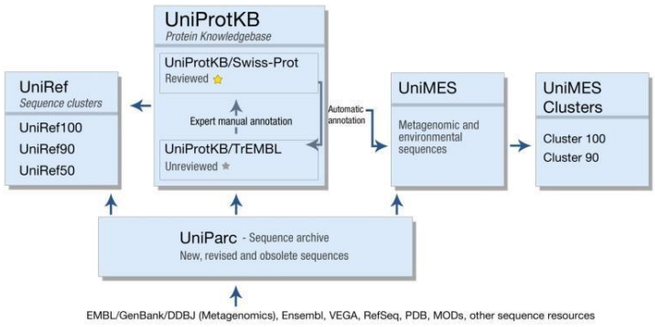


Figure 0.47 online database for proteins

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.

73 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



Figure 0.48 protein data bank



Figure 0.49 P0CG47 - UBB_HUMAN

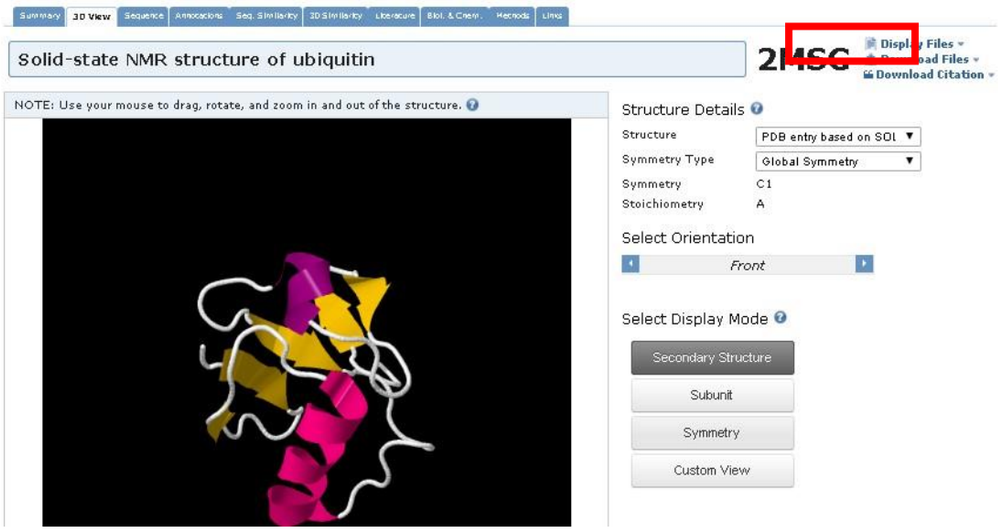


Figure 0.50 P0CG47 - UBB_HUMAN

```
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

74 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

75 GenBank

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

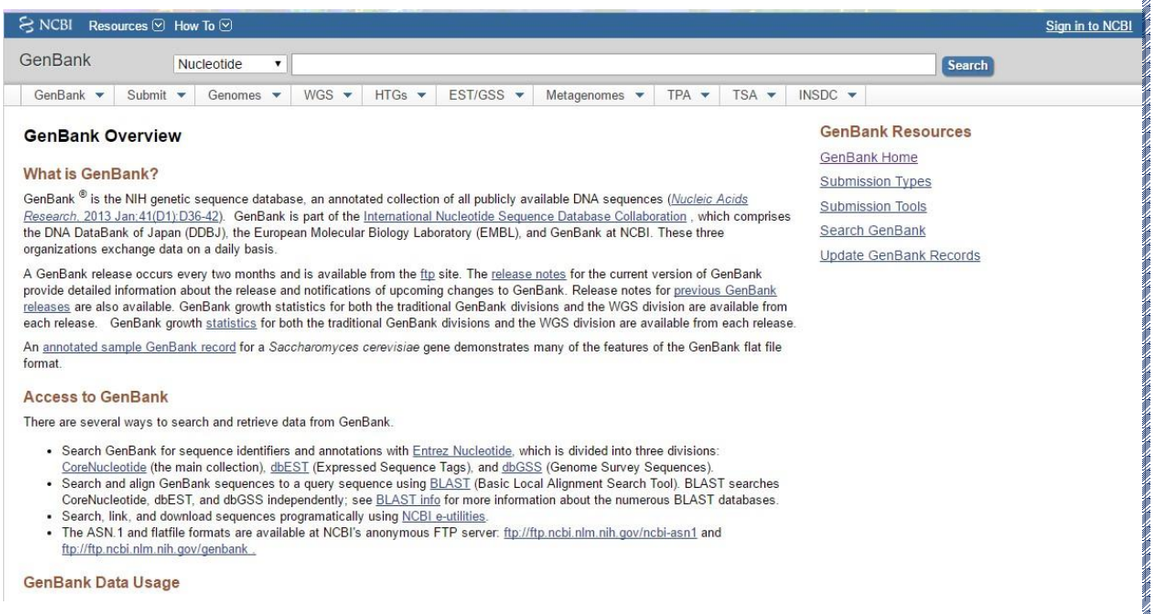
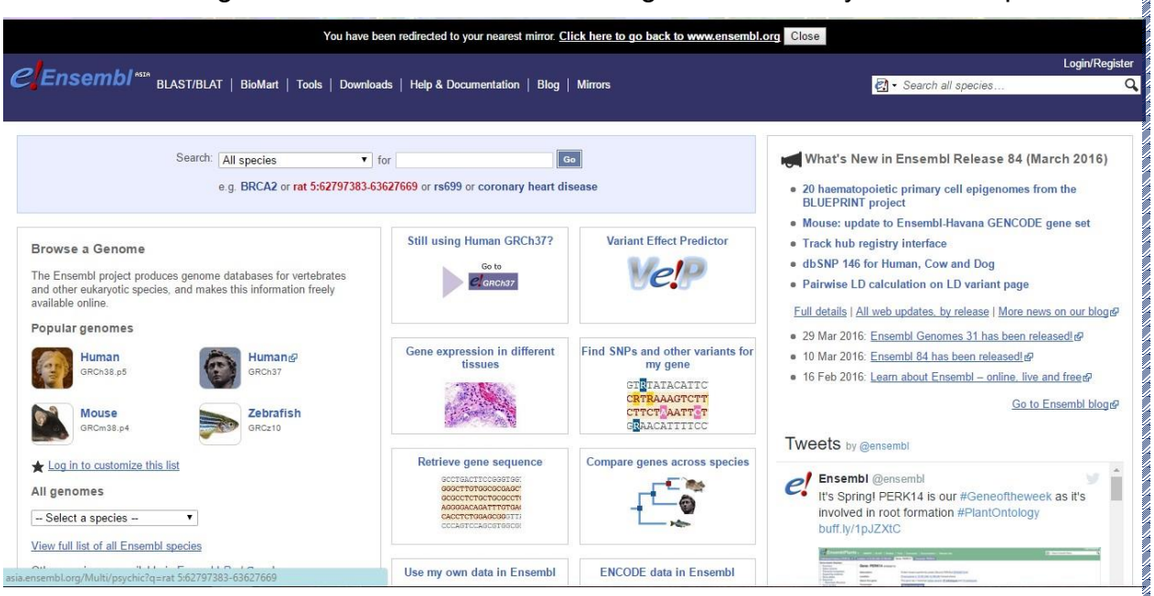


Figure 0.52 http://www.ncbi.nlm.nih.gov/genbank/

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

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.



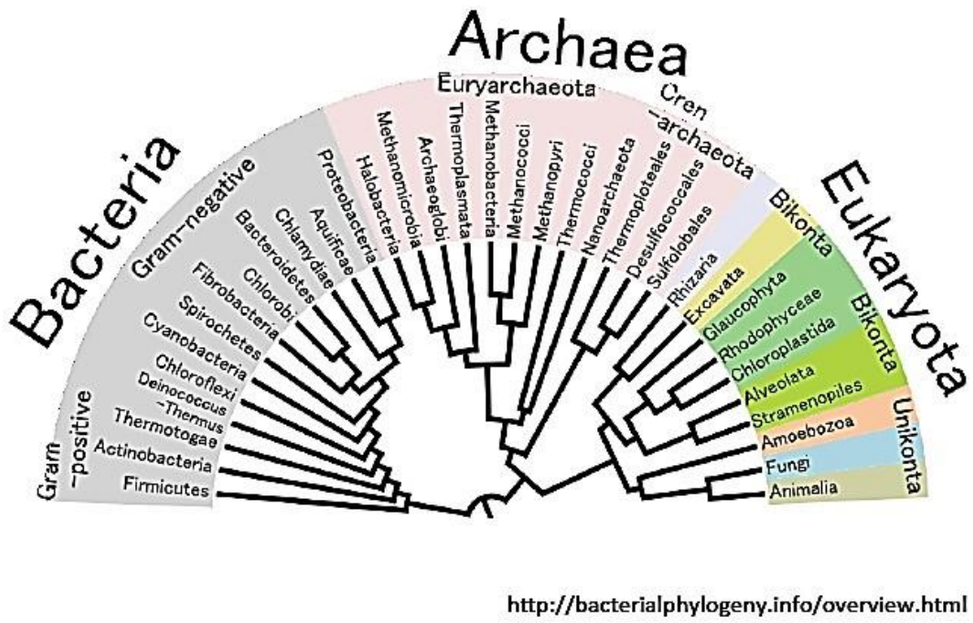
http://asia.ensembl.org/index.html

76 Molecular evolution and 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)



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

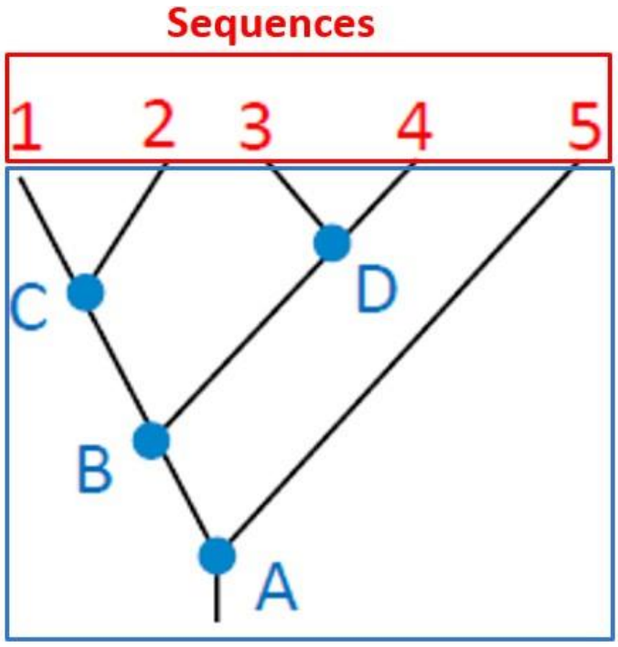


Figure 0.2 phylogenetic tree interference

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.

77 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.

78 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

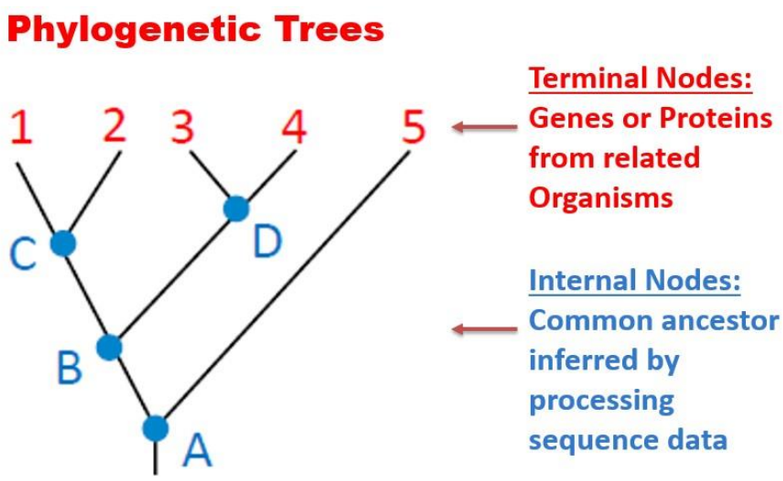


Figure 3 phylogenetic tree from ancestor to evolution

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.

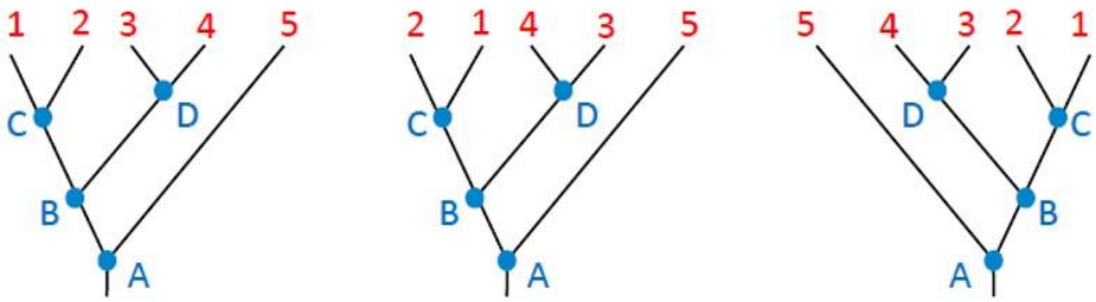


Figure 4 layout of trees

All trees have same meanings.

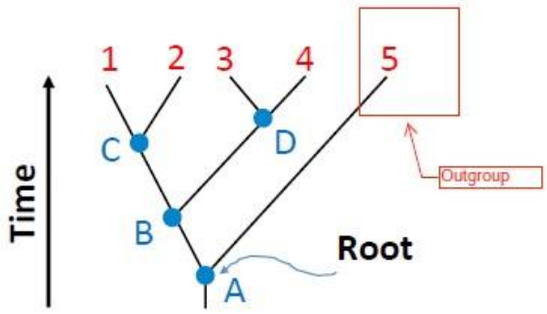


Figure 5 rooted tree

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.

79 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.

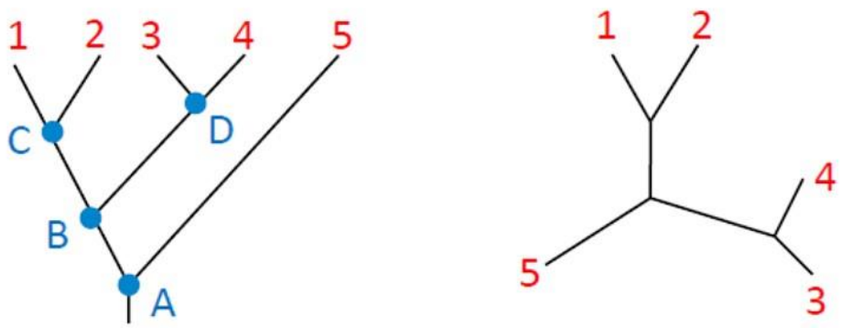


Figure 6 rooted tree vs unrooted

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.

80 Algorithm 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)*.

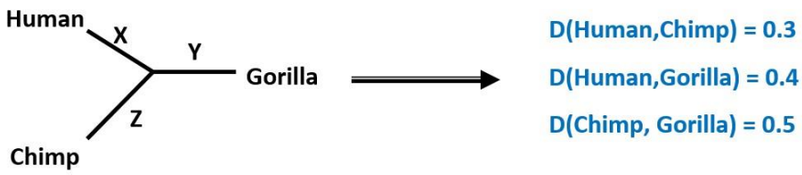


D(Human,Chimp) = 0.3

D(Human,Gorilla) = 0.4

D(Chimp, Gorilla) = 0.5

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



Conclusion

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

81 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:**

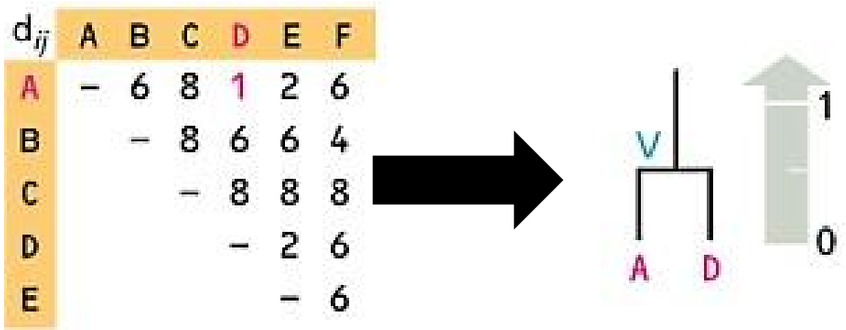
Assume we have N sequences

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**



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

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

Introduction to UPGMA

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

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

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

- **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

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.

82 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

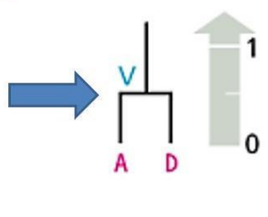


Figure 10 the distance matrix is obtained using pairwise sequence alignment

- **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:

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

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

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.

UPGMA - I

Methods for constructing trees

What are the distances between: **V and C (Calculate)!**

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

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

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

UPGMA - I

Methods for constructing trees

What are the distances between: **V and E (Calculate)!**

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

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

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

UPGMA - I

Methods for constructing trees

What are the distances between: **V and F (Calculate)!**

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

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

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

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.

83 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_{ij}	B	C	E	F	V
B	-	8	6	4	6
C		-	8	8	8
E			-	6	2
F				-	6

UPGMA - II

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

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

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

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

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

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.

84 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

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

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.

New matrix

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

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

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

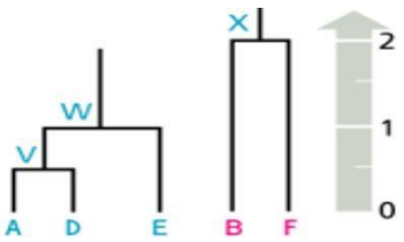
d_{ij}	B	C	E	F	V
B	–	8	6	4	6
C		–	8	8	8
E			–	6	2
F				–	6



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



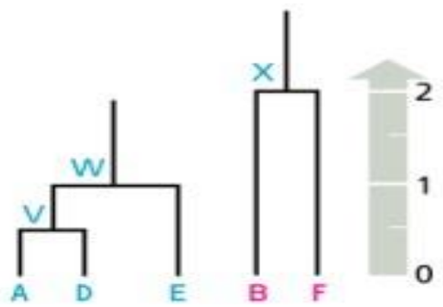
Conclusion

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.

85 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



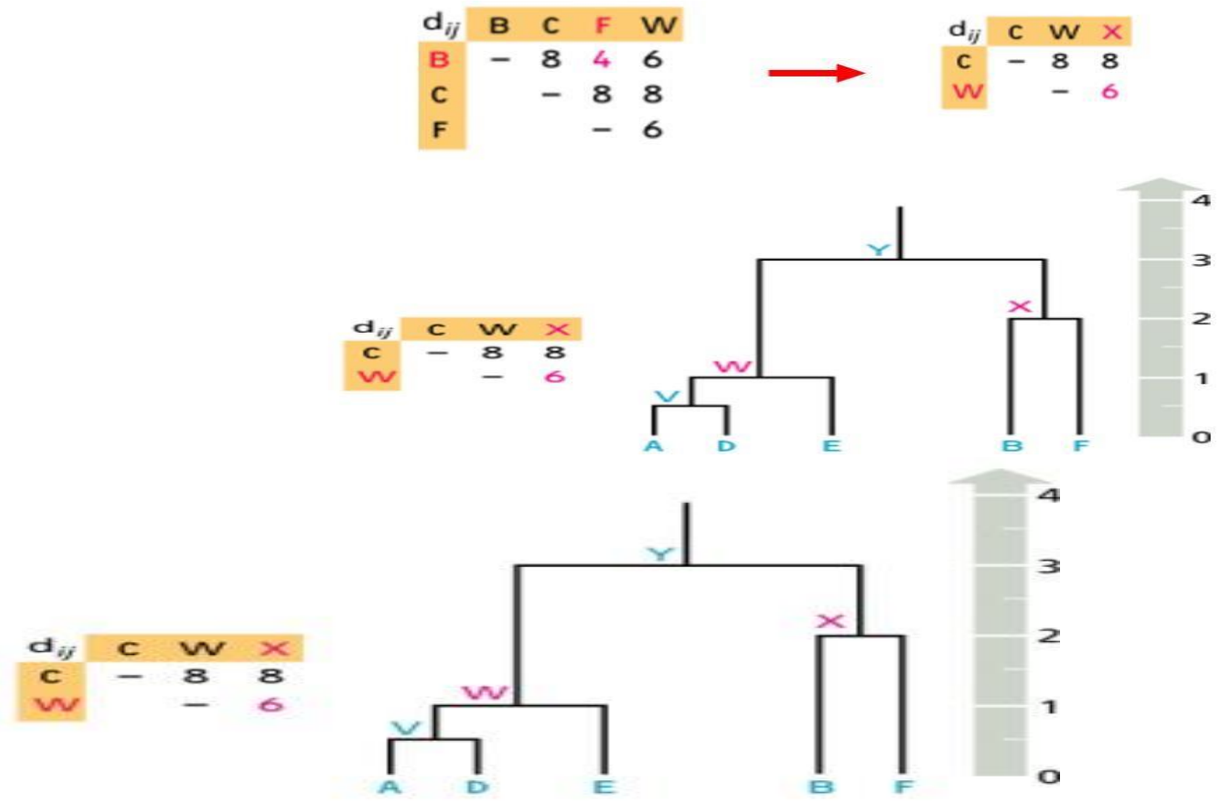
F – B becomes a new cluster lets say X. We have to modify the distance matrix yet again. What is the distance between trees: W and X.

d_{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

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

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

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



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

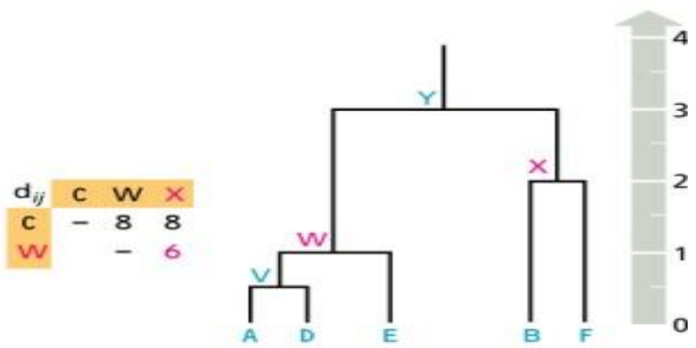
What is the distance between: Y and C.

Conclusion

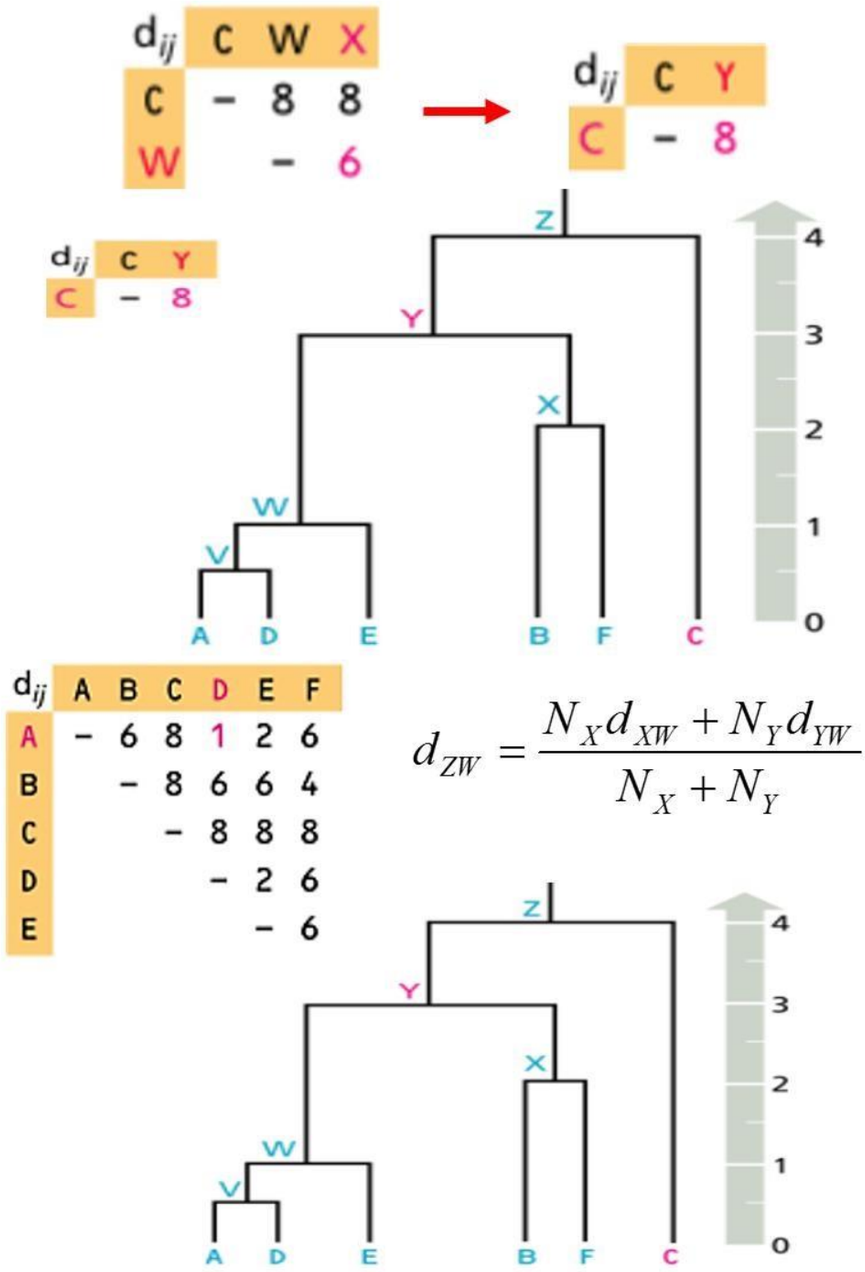
We have now seen how trees are generated and connected. Next, we need to finalize the tree by adding the last two clusters.

86 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.



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

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.

87 DNA to RNA Sequences, Base Complementarity

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.

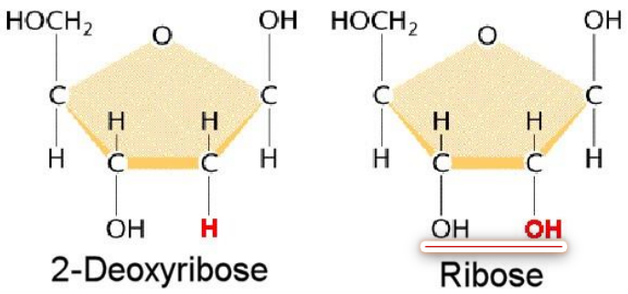


Figure 0.1 Ribose sugar has (OH) and Deoxyribose (H)

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

88 Types of RNA and their Function

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)

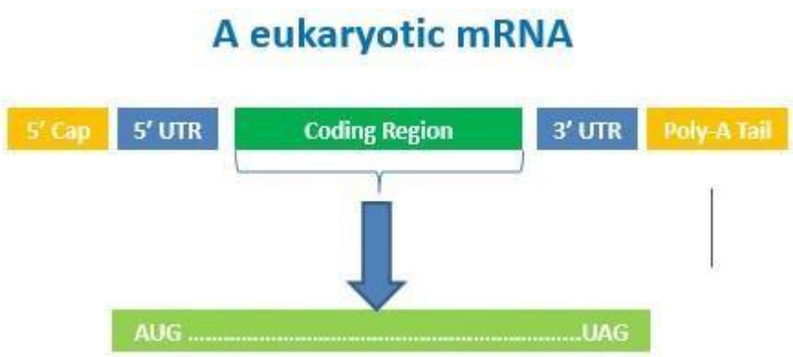


Figure 2 RNA sequence is complementary to the DNA sequence and is translated as codons of three nucleotides

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

89 Significance of RNA Structures

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)

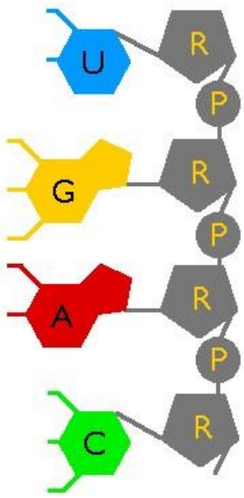


Figure 3 In RNA ribose is used in place of deoxyribose 3 In RNA uracil is used in place of thymine

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

90 RNA Folding, Energies of 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

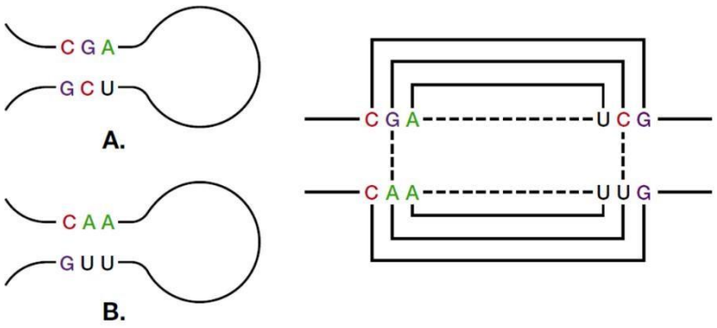


Figure 4 Energy is continuously given out as the RNA molecule folds by pairing complementary bases

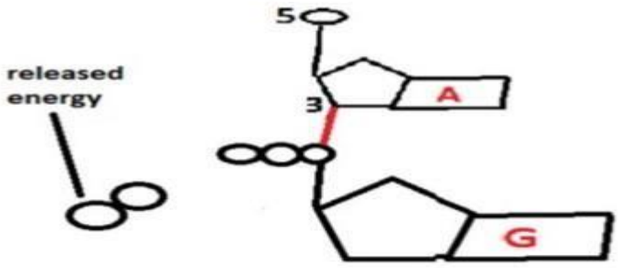
X \ Y	A	C	G	U
A	.	.	.	-2.1
C	.	.	-3.3	.
G	.	-2.4	.	-1.4
U	-2.1	.	-2.1	.

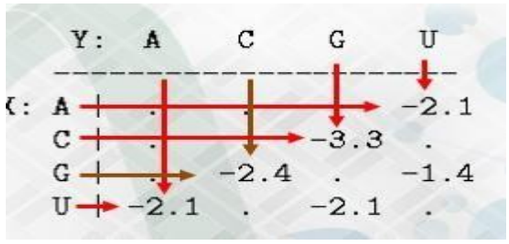
We can calculate the overall energy of RNA structures by summing up energies given out during the process of folding. For knowing the positive and negative values of calculations of stabilizing and destabilizing energies we may factor in ways in which RNA can be destabilized.

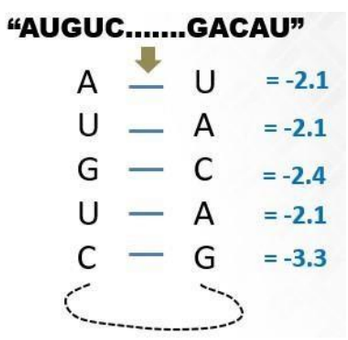
91 Calculating Energies of Folding - An Example

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.







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.

92 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**

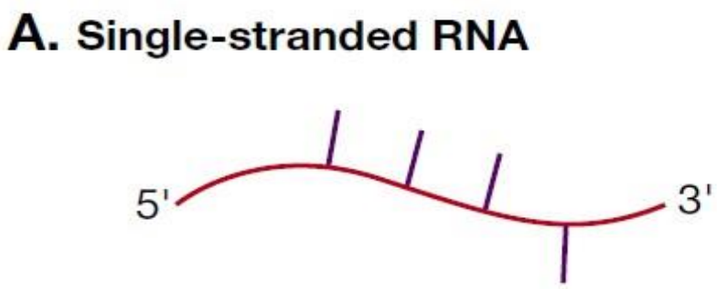
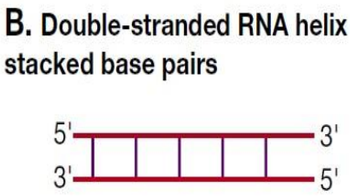
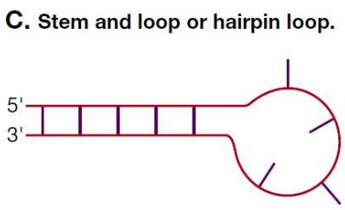


Figure 8 RNA sequence extends from its 5’ end to 3’ end. Upon folding, 3’ end may fold on to the 5’ end

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



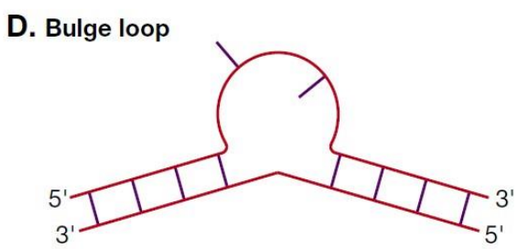
The loop of the hairpin must at least four bases long to avoid steric hindrance with base-pairing in the stem part of the structure.

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

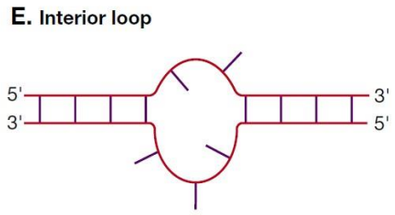
93 Types of RNA Secondary Structures – II

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.



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.



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

94 Types of RNA Secondary Structures – III

Another 2' RNA structure is the Junction or Intersection.

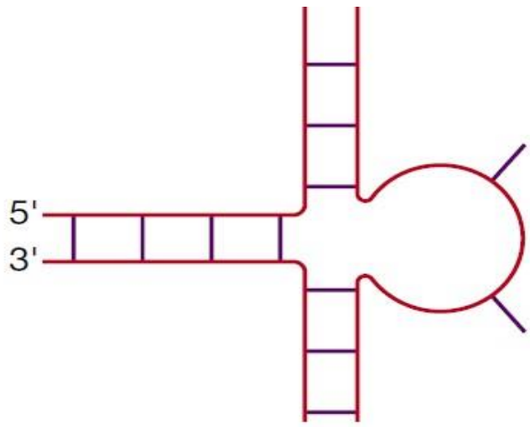


Figure 9 2' RNA structure called junction

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)

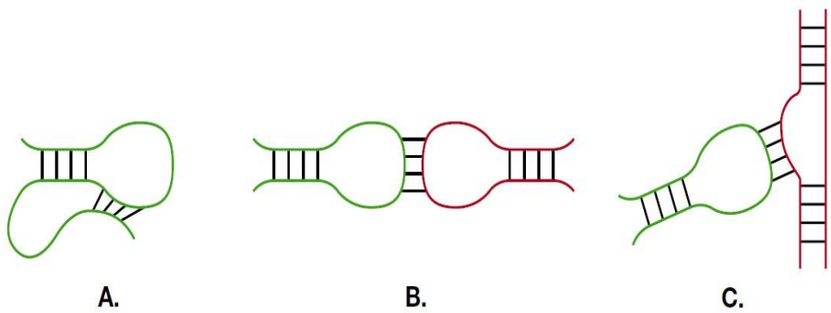


Figure 10 Unpaired bases in two 2' structures form hydrogen bonds with each other

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

95 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.

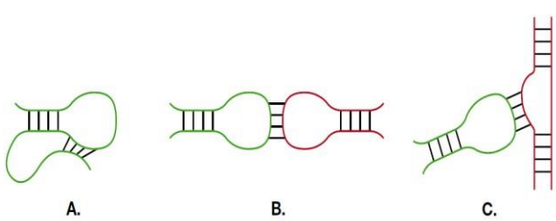
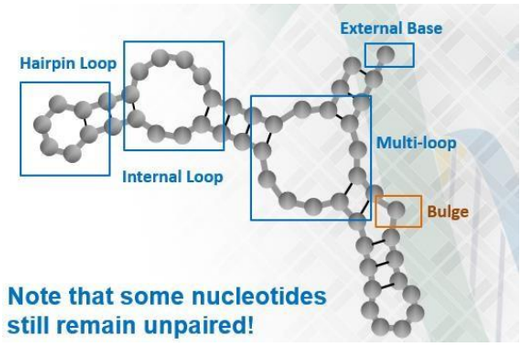


Figure 11 Hydrogen bonding formation in open nucleotides.

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.



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.

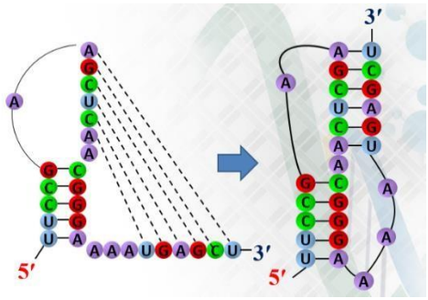
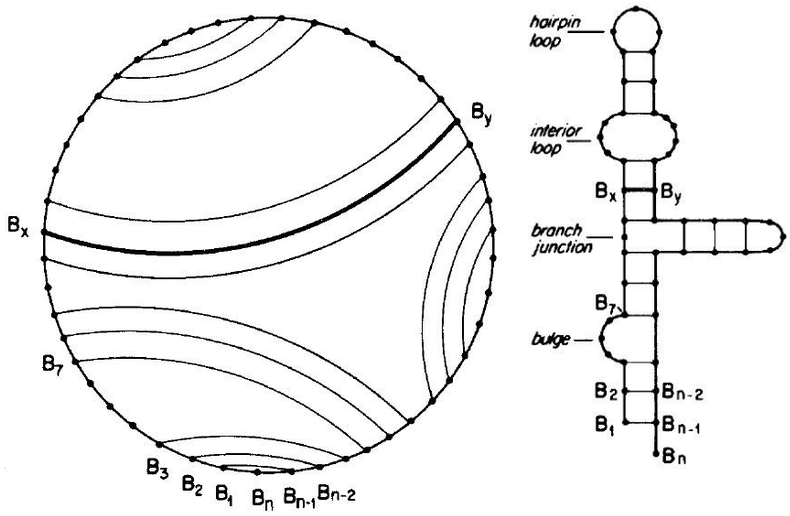


Figure 12 pseudoknots

96 Circular Representation of RNA 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.

97 Experimental Methods for Determining RNA Strucutres

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

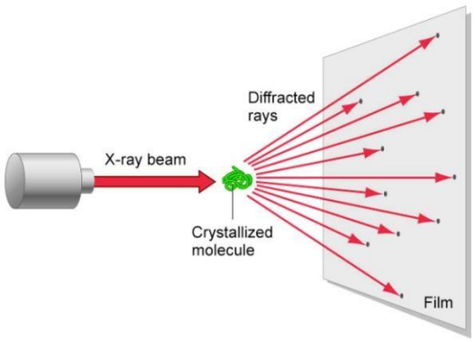


Figure 13 X-ray Crystallography
https://260h.pbworks.com/w/page/30814223/X%20Ray%20Crystallography

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.

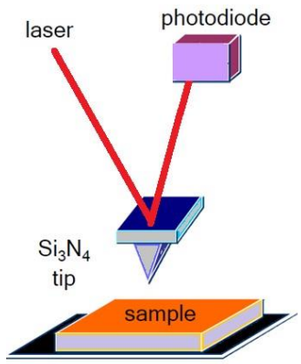
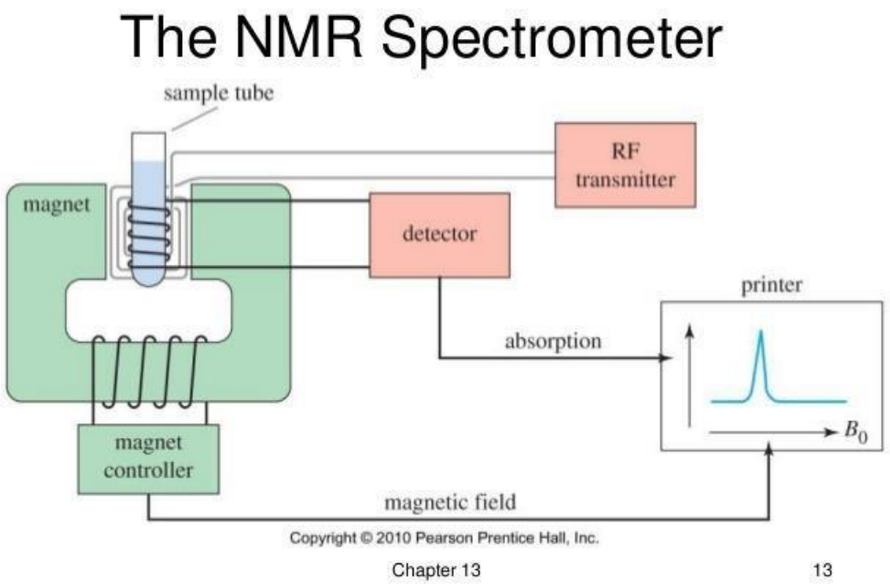


Figure 14 Atomic microscopy

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



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.

98 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.

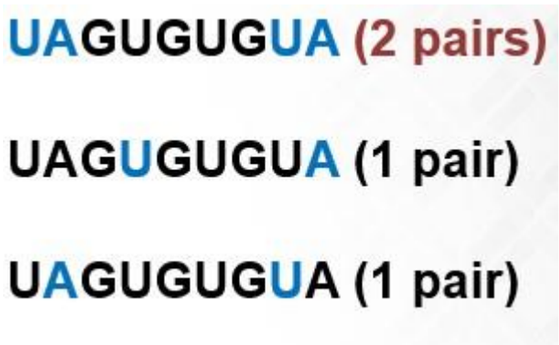


Figure 15 pairs represent the stability of the RNA molecule

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

99 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:

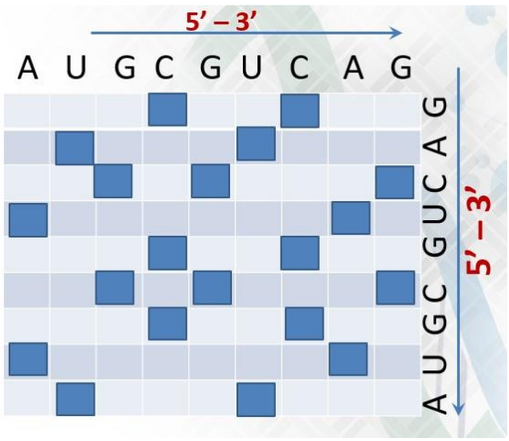


Figure 16 dot are placed at complimentary base pair place.

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

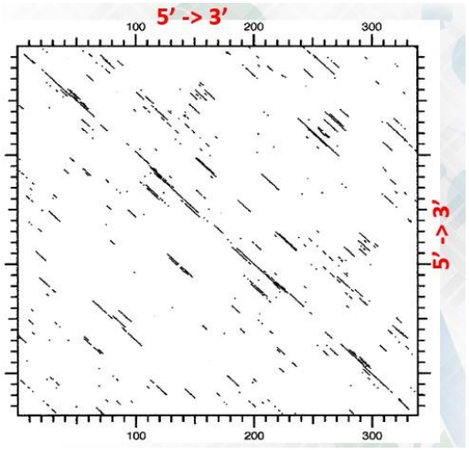


Figure 17 Potato Tuber Spindle Viroid

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

100 Energy Based 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	.

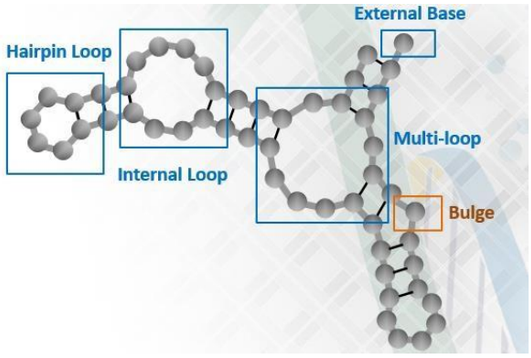
Figure 18 energy table

- **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.



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

101 Zuker 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.

Zuker's Algorithm

Zuker's Algorithm (2003)

Stacking Energies

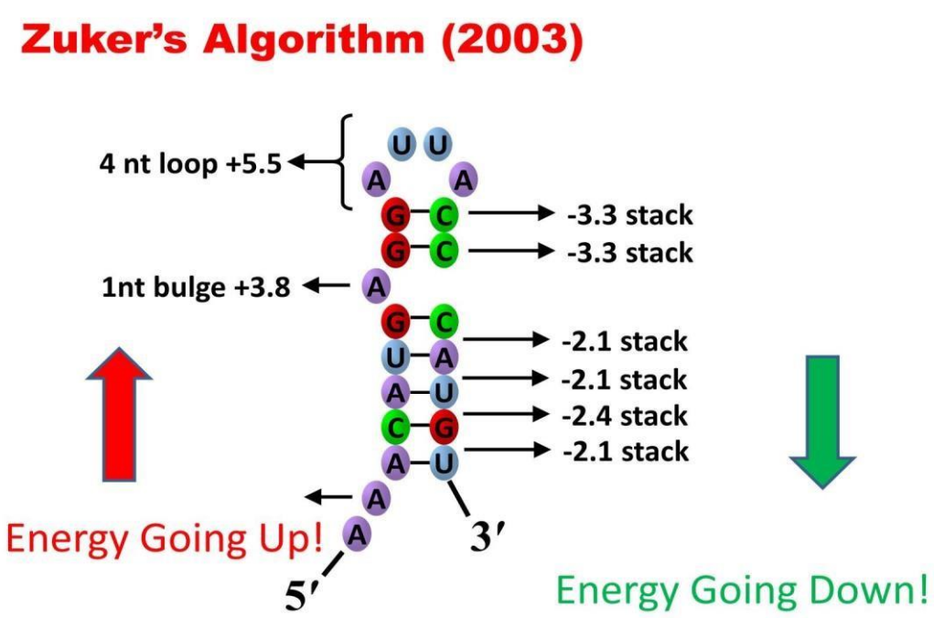
X \ Y:	A	C	G	U
A	.	.	.	-2.1
C	.	.	-3.3	.
G	.	-2.4	.	-1.4
U	-2.1	.	-2.1	.

kcal/mole

Destabilization 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

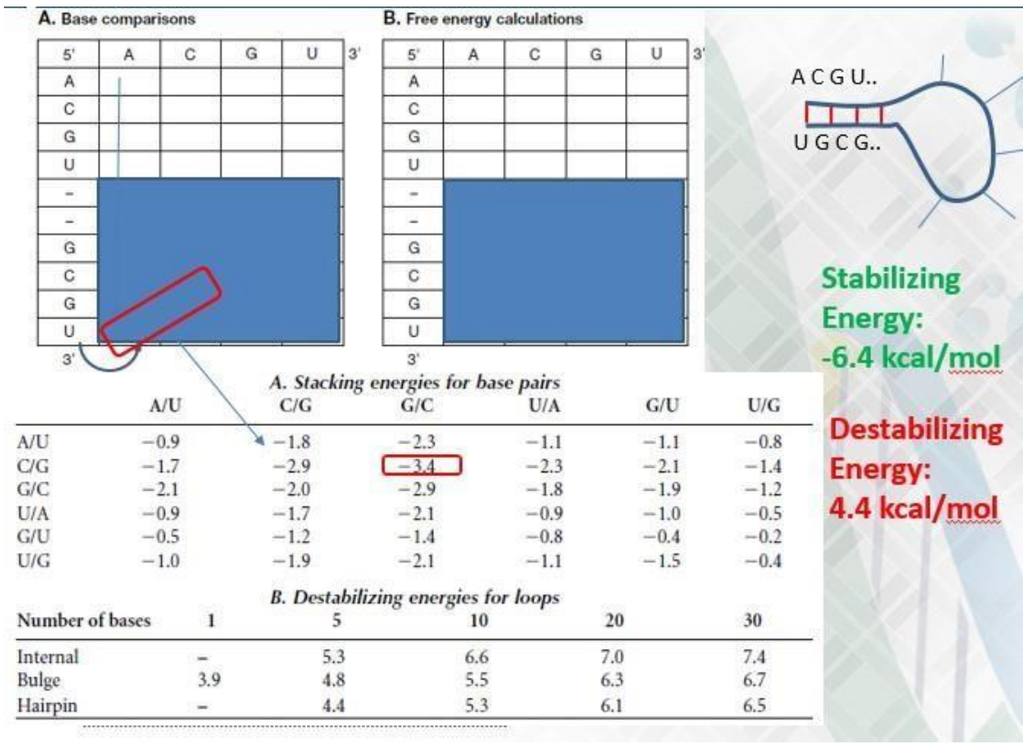
kcal/mole



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

102 Example - Zuker Algorithm

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

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.

103 Zuker Algorithm – Flowchart

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

A. Base comparisons

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

B. Free energy calculations

5'	A	C	G	U	3'
A					
C					
G					
U					
–					
–					
G				−6.4	
C			−5.2		
G		−1.8			
U					
3'					

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

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

The flow chart for energies is:

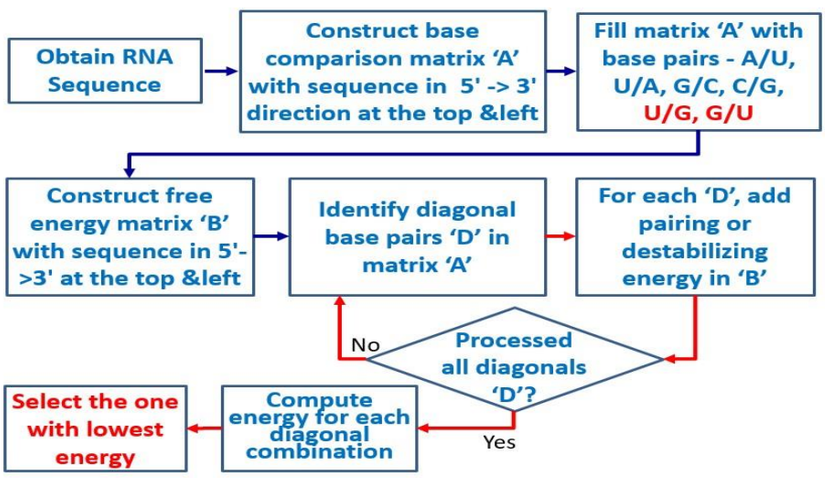


Figure 24 flow chart

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

104 Martinez Algorithm

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

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

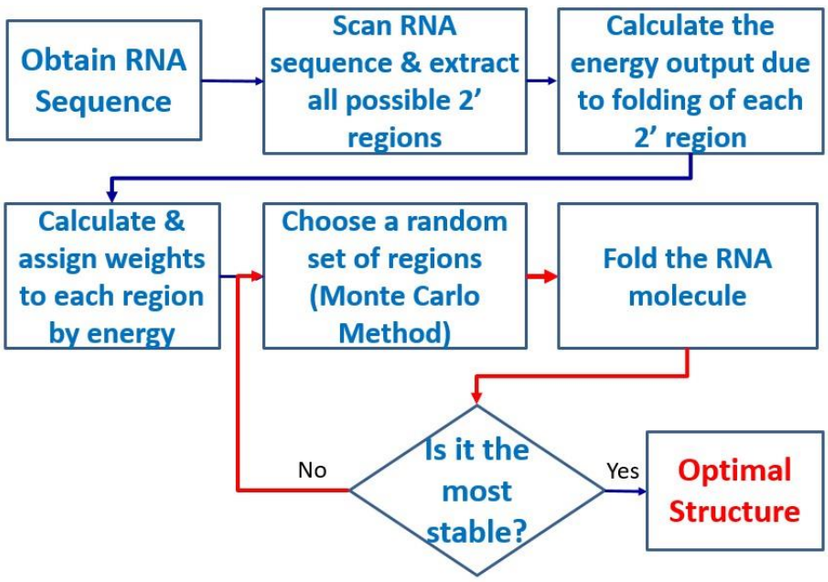


Figure 25 Martinez Algorithm flow chart

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

And Monto Carlo method do not provide a definitive solution.

105 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:

AUGCUGUCAUCG

AUGCUGUCAUCG

AUGCUGUCAUCG

AUGCUGUCAUCG

Figure 26 all possible complementary bases combinations.

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

106 Nussinov -Jacobson Algorithm 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

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

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

Left Row Bottom Column

107 NJ Algorithm 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.

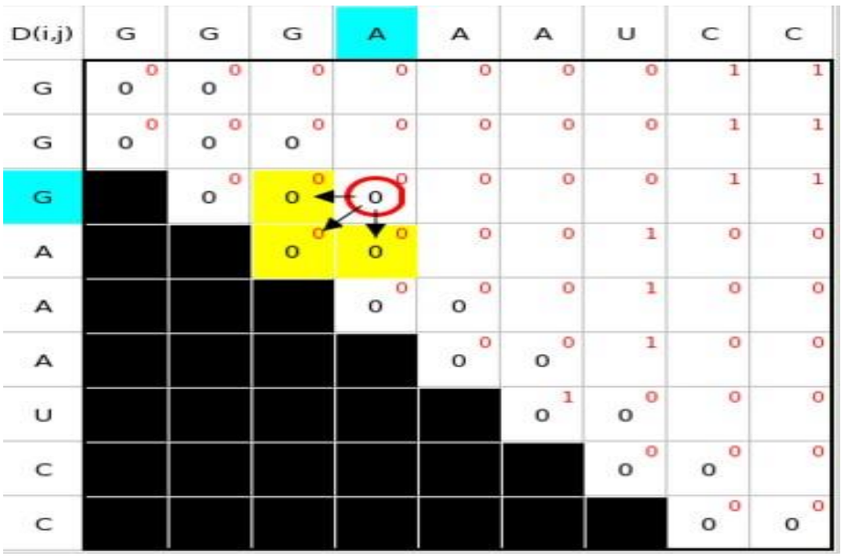


Figure 0.3 for maximum score 4 positions are used in scoring

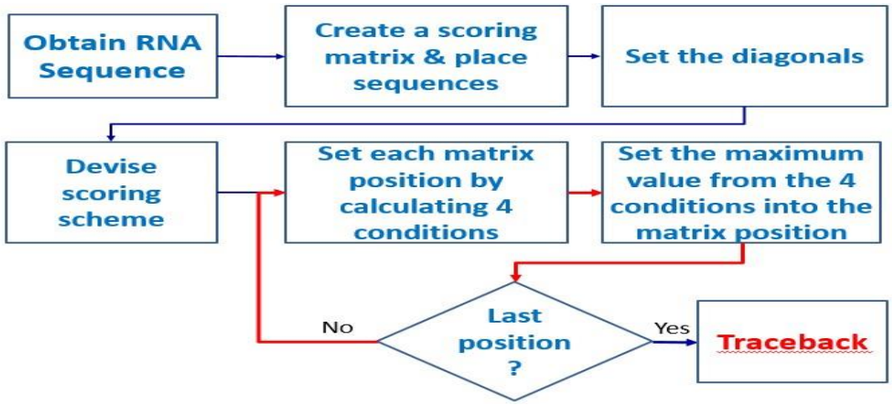


Figure 28 flow chart of NJ Algorithm.

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

108 Example - Jacobson Algorithm

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

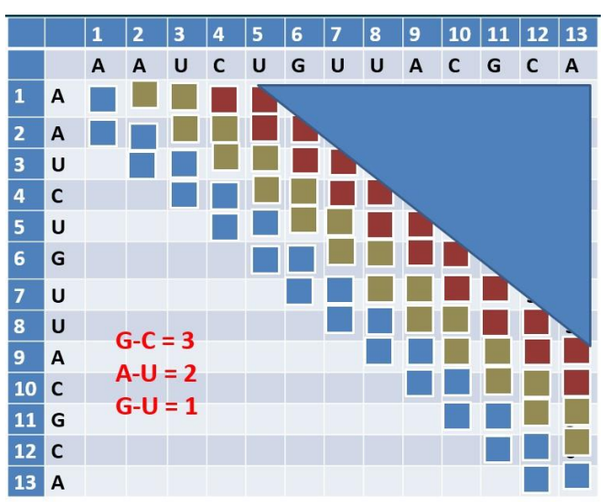


Figure 29 N-J Algorithm scoring

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

109 Score Calculations and 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	G-C = 3							0	0	0	3	3	3
10	C	A-U = 2								0	0	3	3	3
11	G	G-U = 1									0	0	3	3
12	C											0	0	0
13	A	4th condition of max{S(i,k)+S(k+1,j)}.											0	0

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.

110 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

111 Web Resources: RNA Bricks

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

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

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

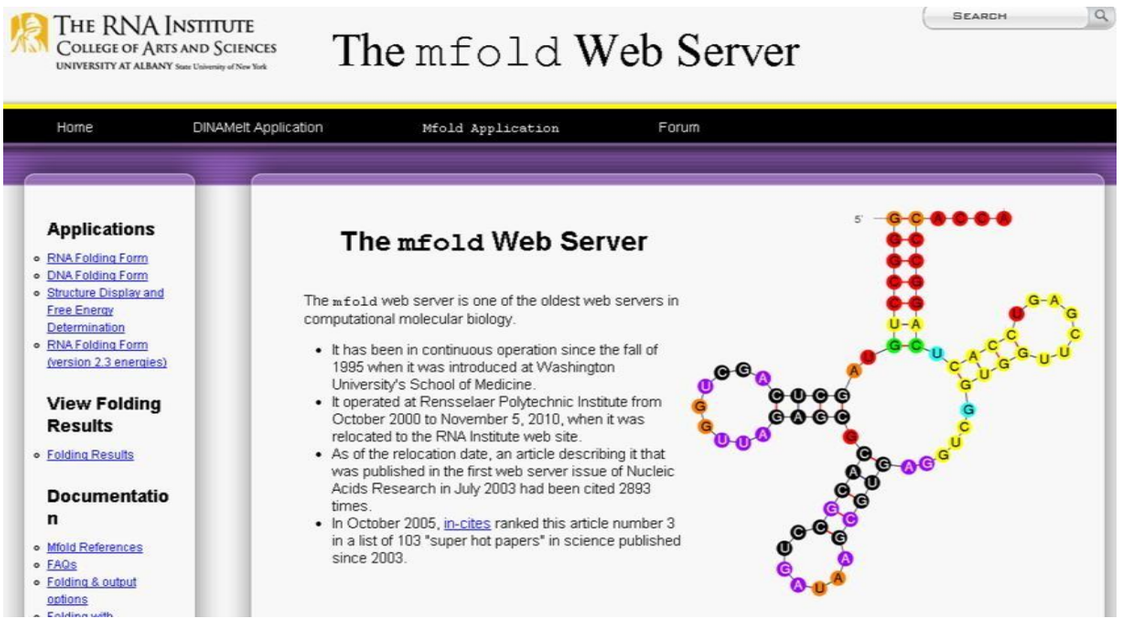


Figure 31 http://unafold.rna.albany.edu/?q=mfold

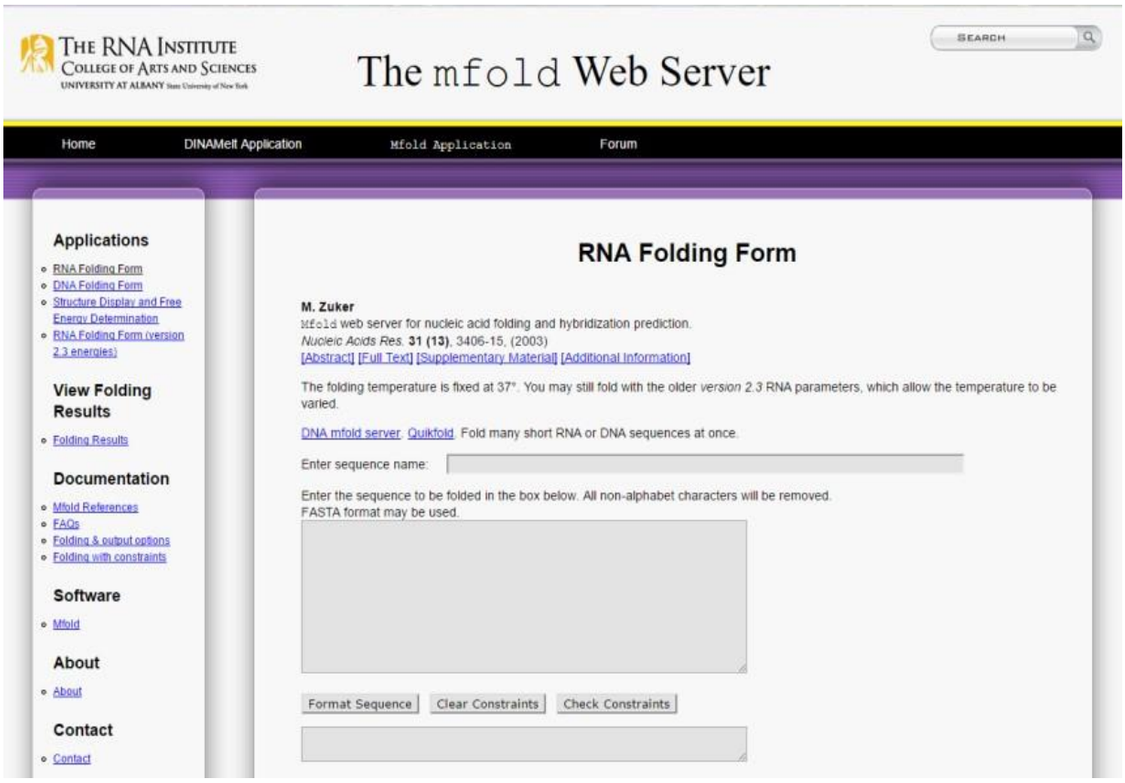


Figure 32 http://unafold.rna.albany.edu/?q=mfold/RNA-Folding-Form

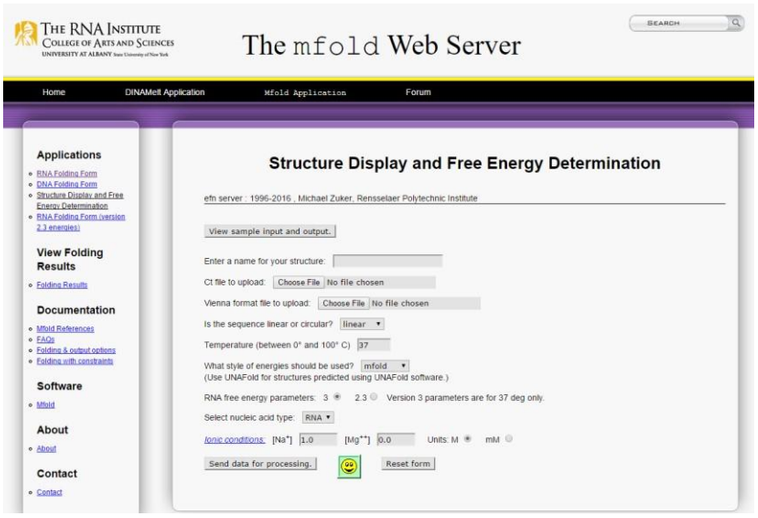


Figure 33 http://unafold.rna.albany.edu/?q=mfold/Structure-display-and-free-energy-determination

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

112 Web resources: MFOLD

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

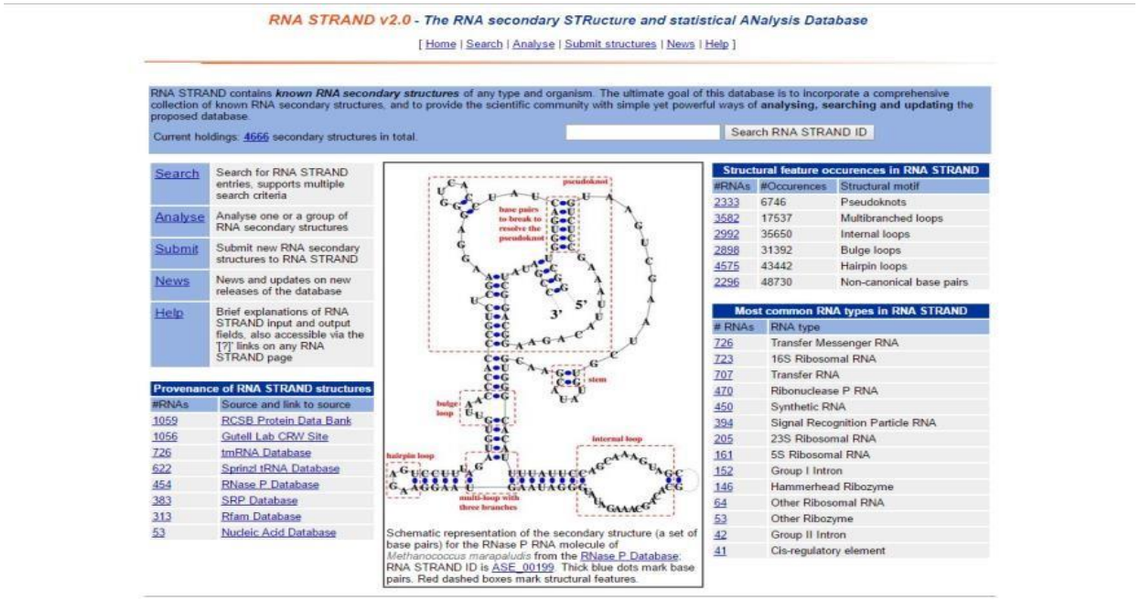


Figure 34 http://www.rnasoft.ca/strand/

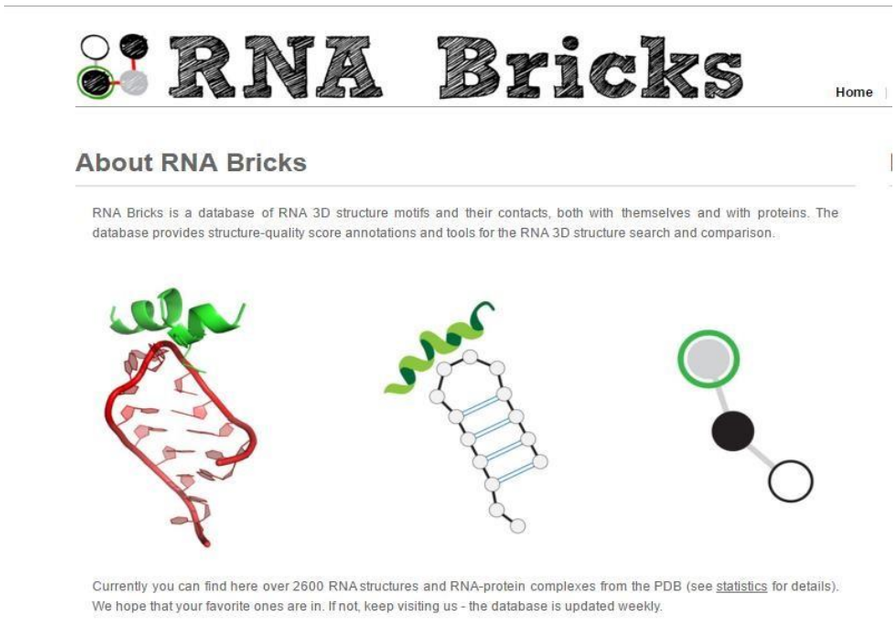


Figure 35 http://iimcb.genesilico.pl/rnabricks

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

113 From DNA/RNA Sequences to Proteins

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

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

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

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

114 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.

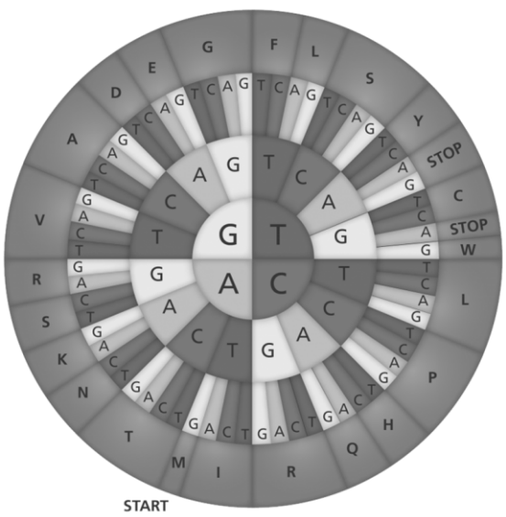


Figure 3 coding of amino acids

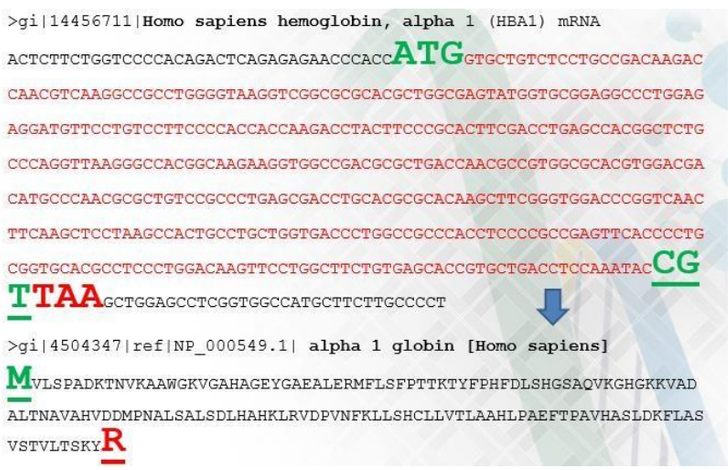


Figure 4 Start Codon ATG and Stop Codon TAG, TGA or TAA

115 Open Reading Frames

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

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

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

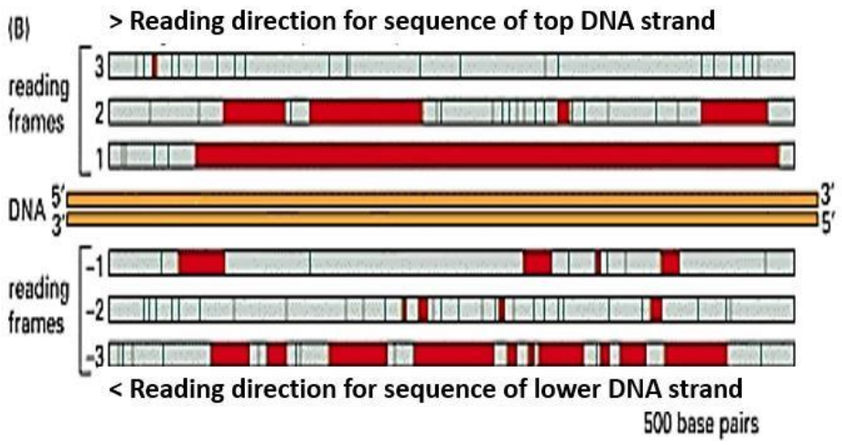


Figure 5 ORF 1 is valid, as it is the longest

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

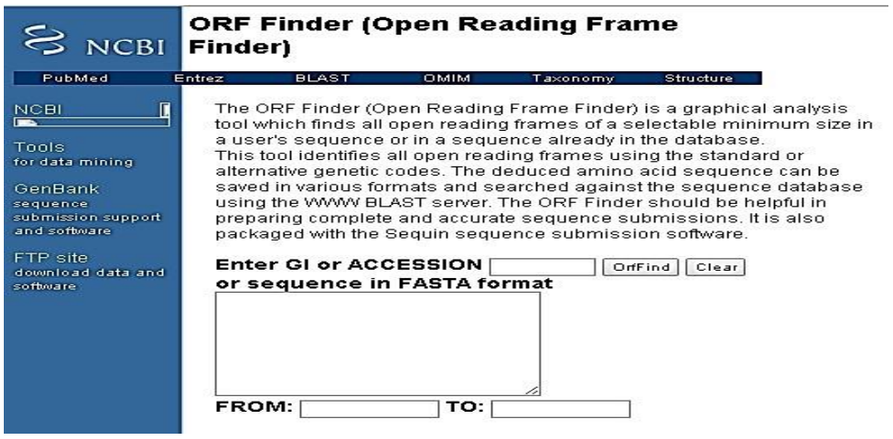


Figure 6 NCBI, ORF Finder

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

116 ORF Extraction – A Flowchart

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

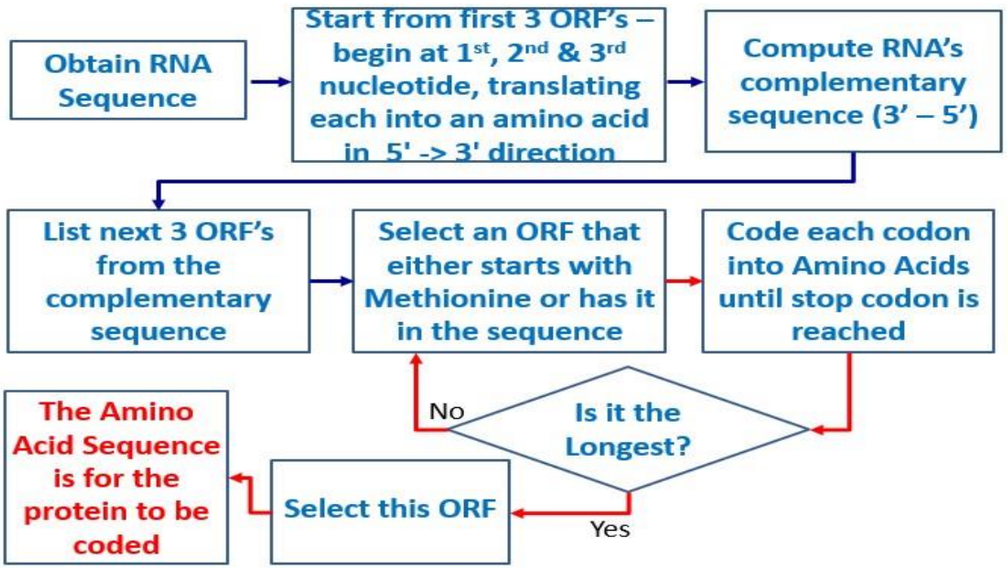


Figure 7 ORF extraction flowchart

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

117 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

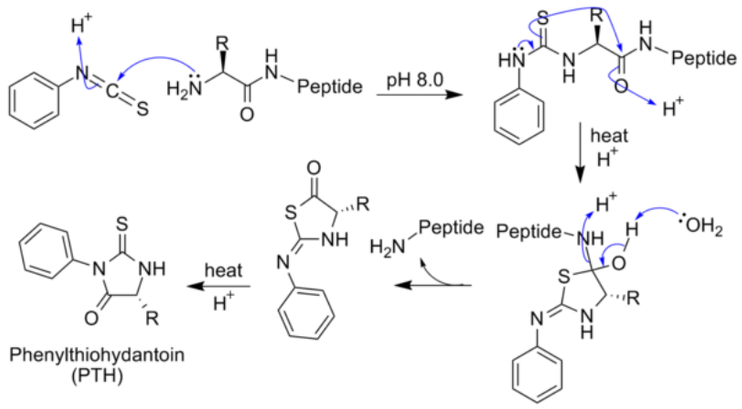
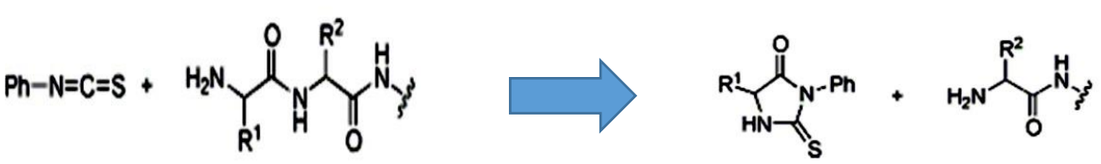


Figure 8 Mechanism

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.



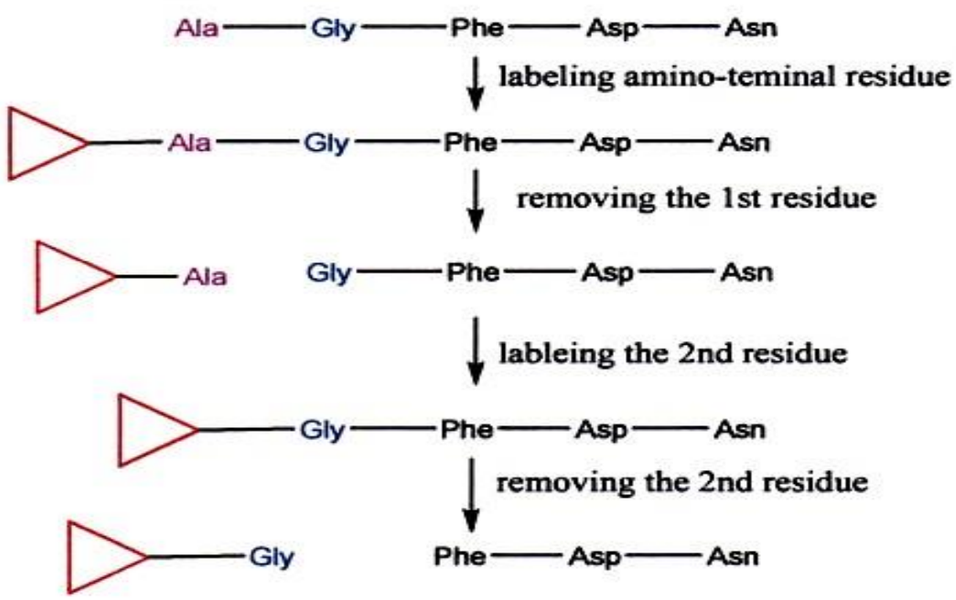


Figure 9 working of Edam degradation

 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.

118 Application of MS in 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.

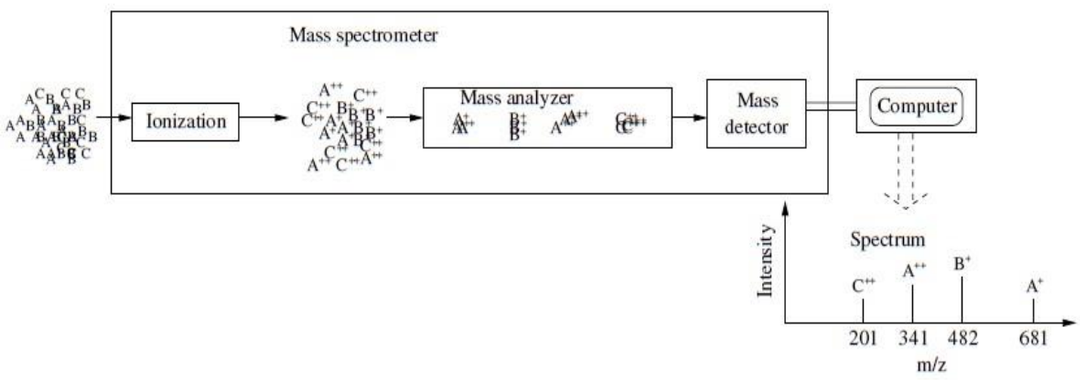
$$\mathbf{F} = Q(\mathbf{E} + \mathbf{v} \times \mathbf{B}), \quad \text{(Lorentz force law)}$$

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

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

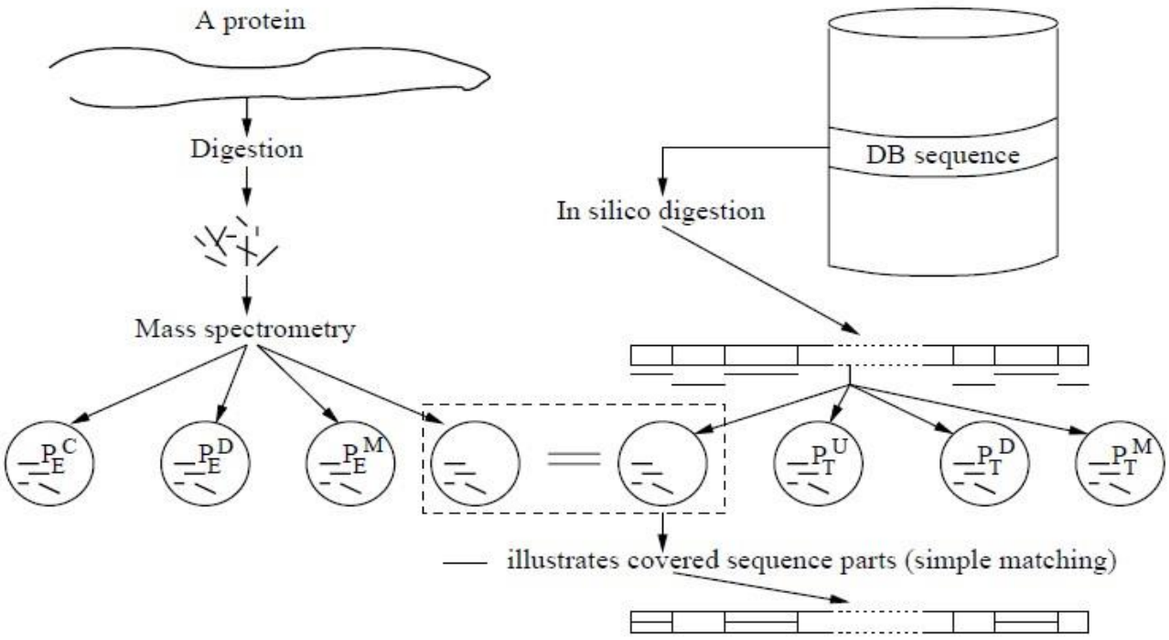
- COMPUNENTS
  - Sample Injection
  - Ionization Source
  - Mass Analyzer
  - Ion Detector
  - Spectra search using computational tools
- CONCLUSION

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)

119 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.

120 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.

121 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.

122 Two Approaches for Bottom Up Proteomics

There are two approaches for bottom up proteomics.

1. Peptide Mass Fingerprinting.
2. Shotgun Proteomics

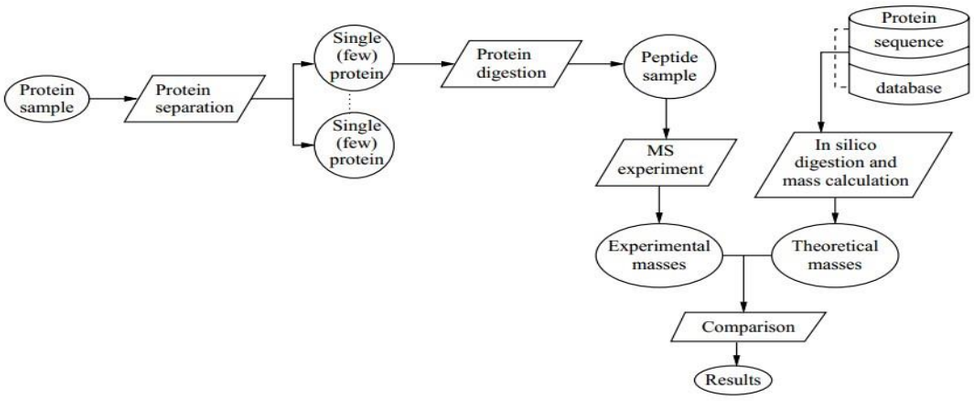


Figure 0.33 Peptide mass fingerprinting

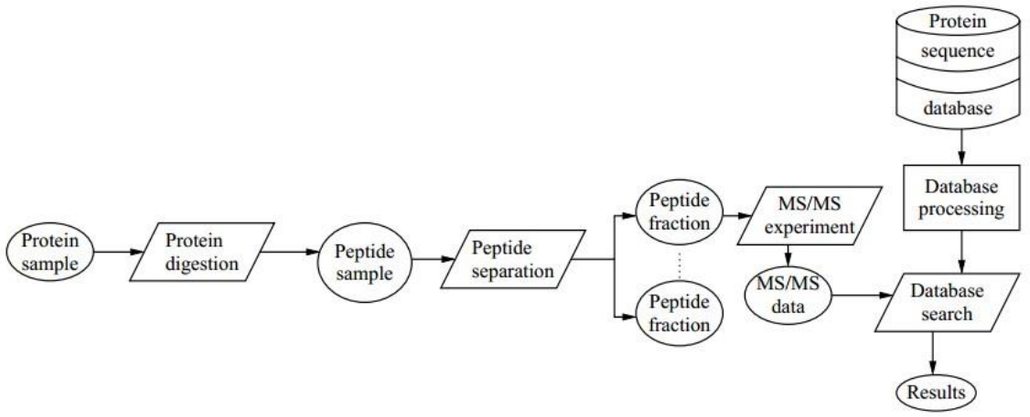


Figure 14 Shotgun Proteomics

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.

123 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.

124 Protein Identification

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

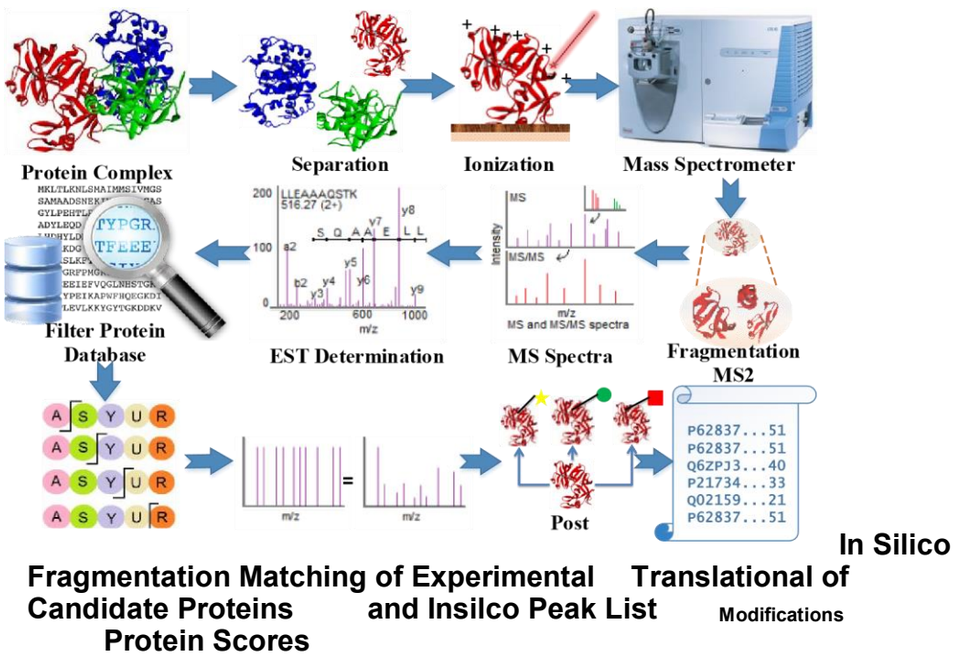
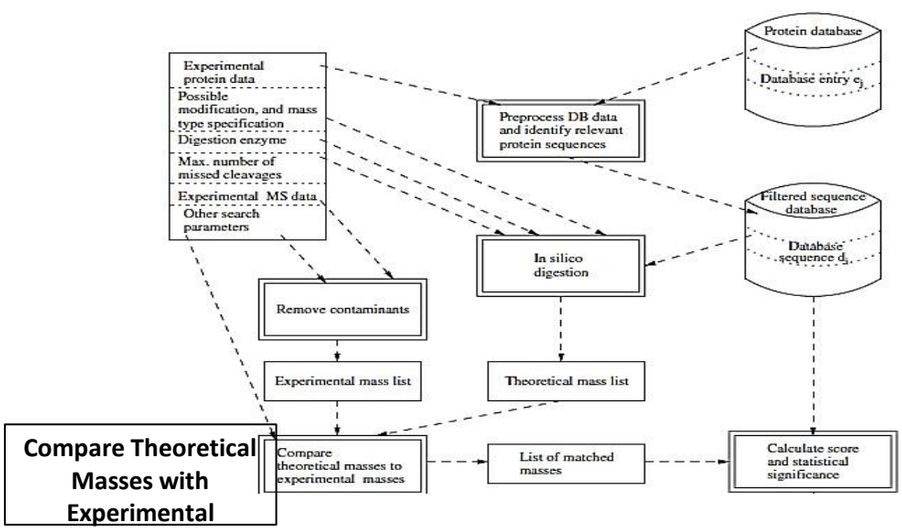


Figure 0.45 protein sequence identification flowchart



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

125 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

ions with charge +1: $m/z = (2000 + 1)/1 = 2001$

ions with charge +2: $m/z = (2000 + 2)/2 = 1001$

ions with charge +3: $m/z = (2000 + 3)/3 = 666.7$

Figure 0.56 resolving multiple charges

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

126 MS1 and 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.

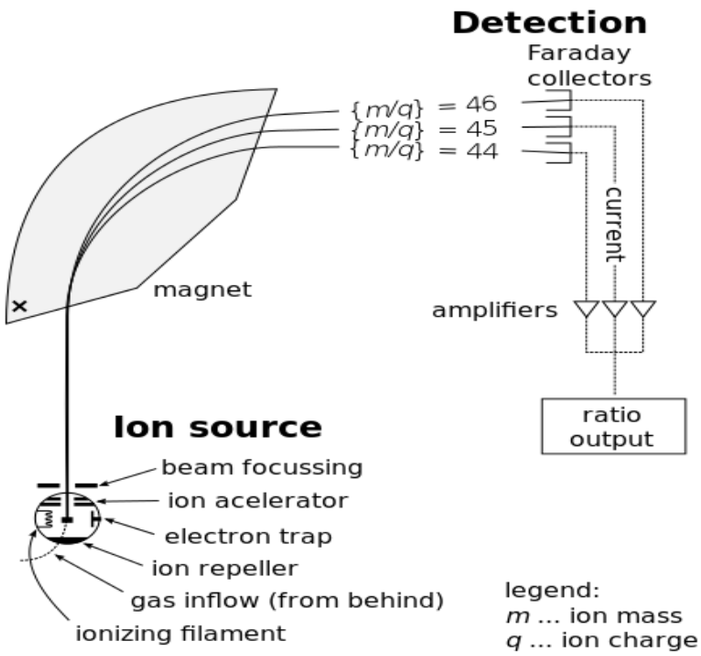


Figure 0.67 MS1 Schematic (Image courtesy Wikipedia)

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.

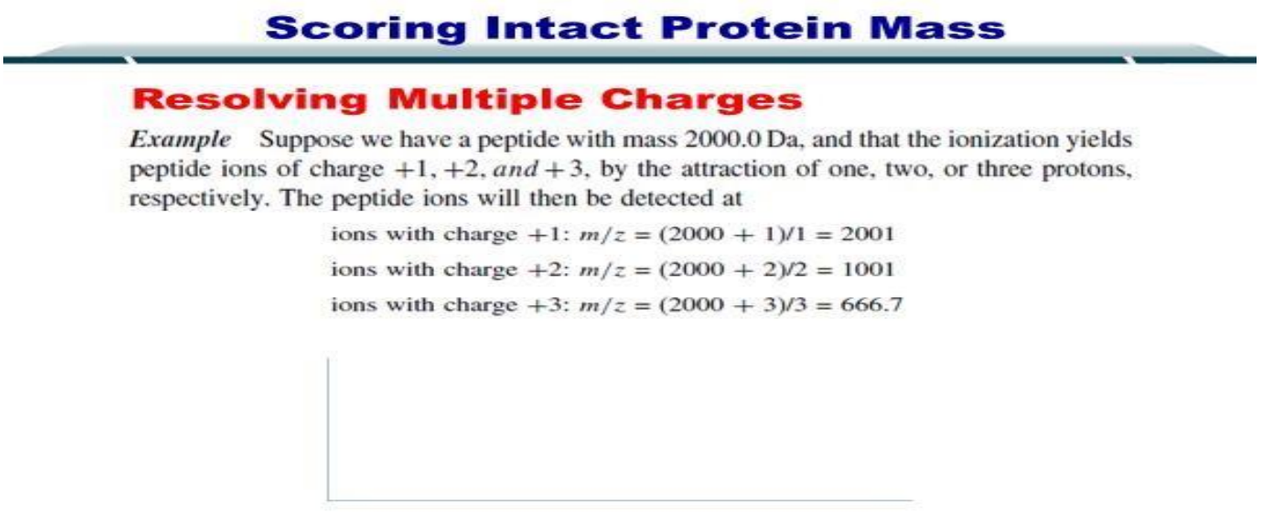
127 Scoring Intact Protein Mass

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**

Scoring Intact Protein Mass

Scoring Scheme:

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

Example: A Protein Sequence from Database "MQLF"

MW: 537.67

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

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

128 Protein Fragmentation Techniques

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

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

There are different techniques for protein fragmentation.

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

Each fragmentation technique gives result of specific type of fragments.

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

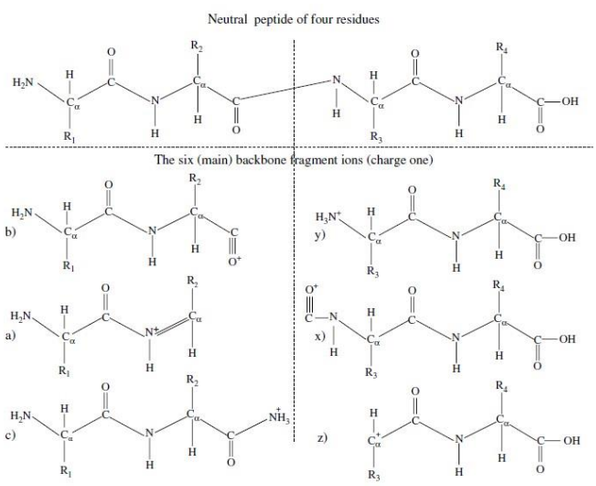
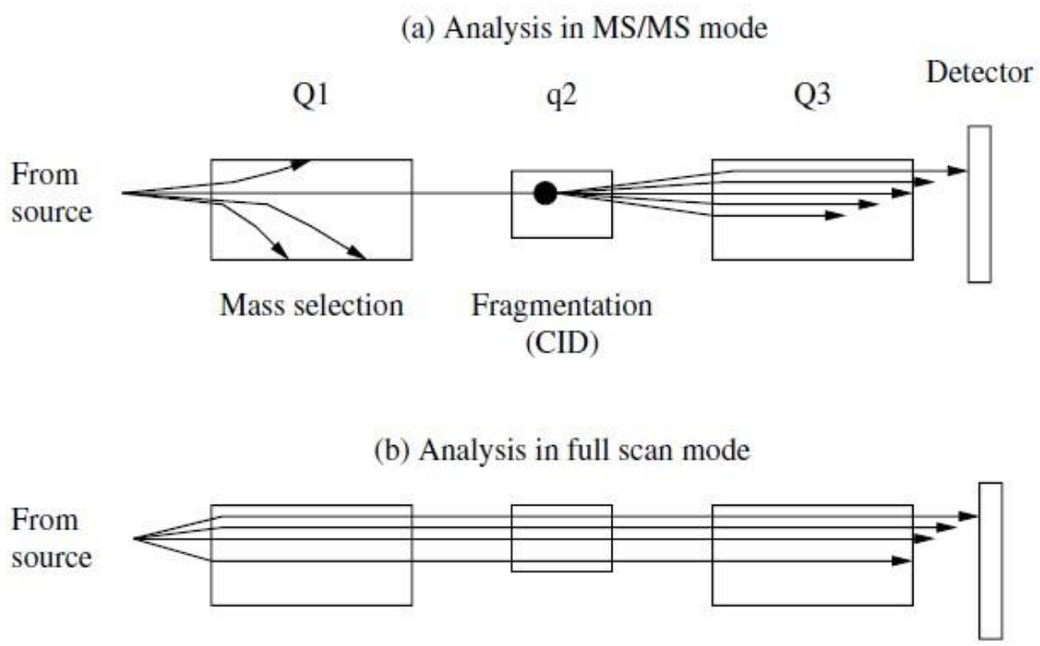


Figure 0.78 natural peptide of four residue

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

129 Tandem MS

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



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

(i) select fragment's mass range.

(ii) Fragment the precursor fragment.

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

130 Measuring Experimental Fragment's Mass

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

FRAGMENTATION TECHNIQUES AND MOLECULAR WEIGHT

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

FRAGMENT MASS

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

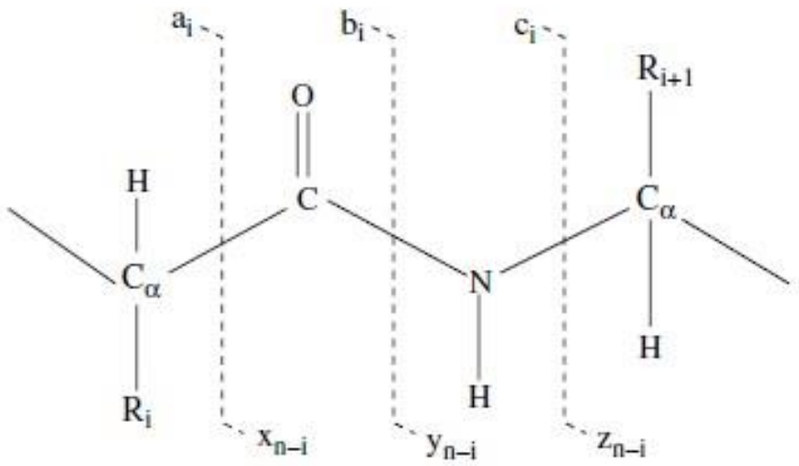


Figure 0.89 Masses after Fragmentation by ECD, CID & ETD

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

131 Calculating Theoretical Fragment's Mass

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

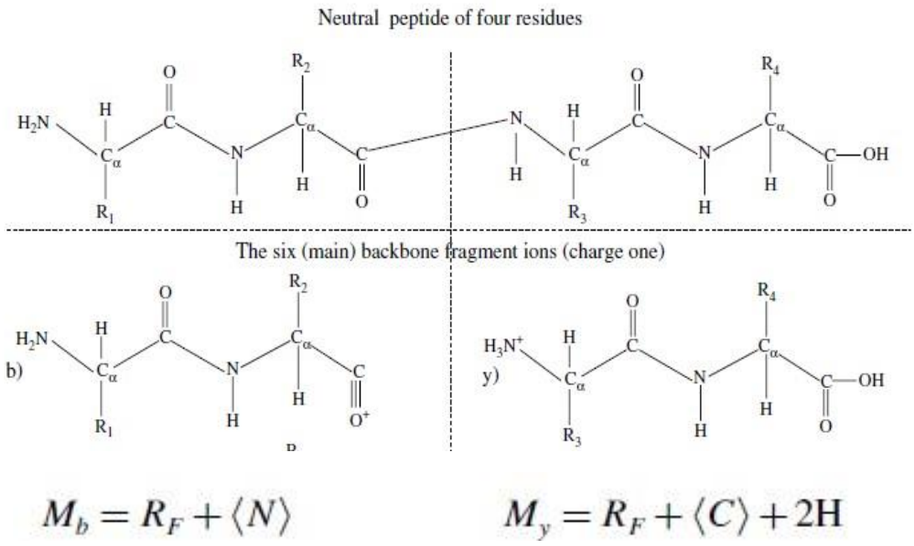


Figure 20 Masses after Fragmentation by CID

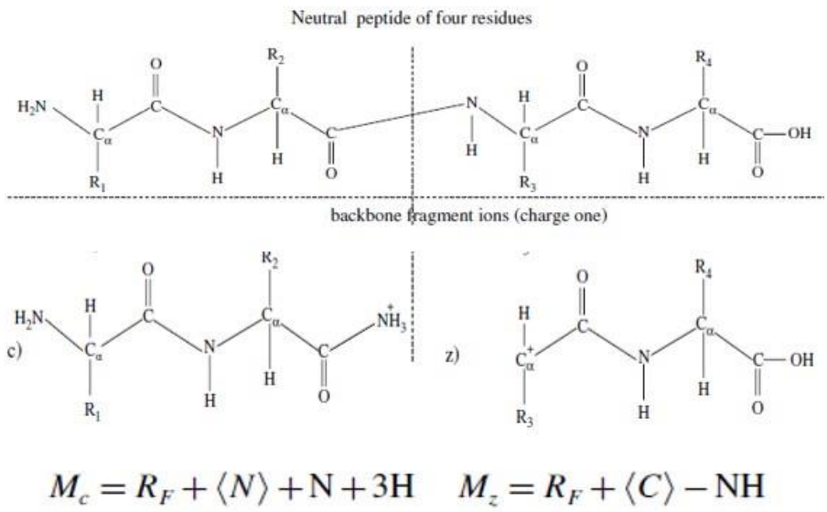


Figure 0.91 Masses after Fragmentation by ECD

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

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

Figure 22 Masses after Fragmentation by ECD

132 Peptide Sequence Tags

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

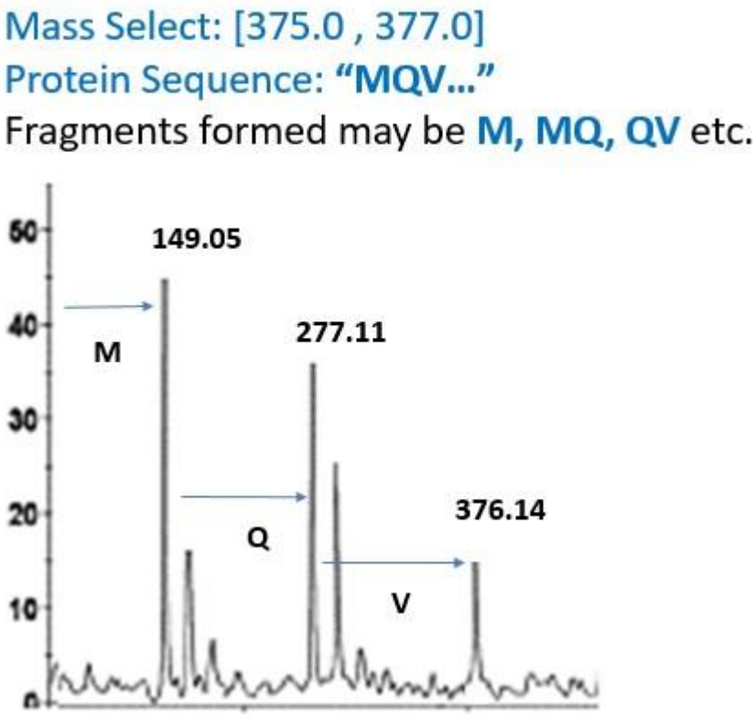


Figure 23 variation sites

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

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

133 Extracting Peptide Sequence Tags

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

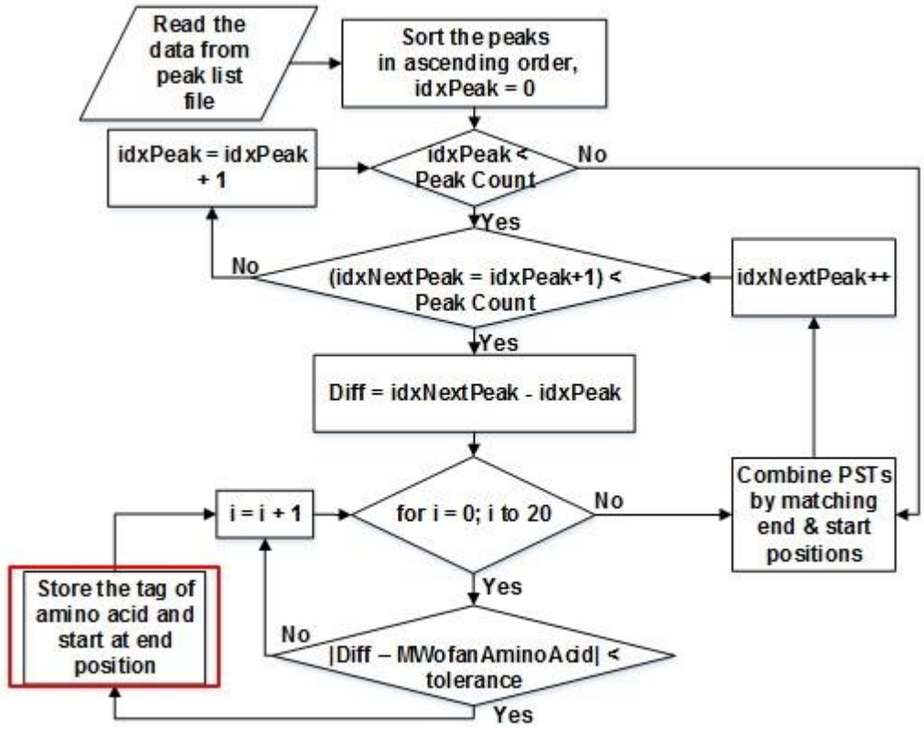


Figure 24 peptide sequencing tagging

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

134 Using Peptide Sequence Tags in Protein Search

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

Sample sequence in protein DB

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

MAFSAEDVLKEYDRRRRMEALLLSLYYPNDRKLLDYKEWSPPRVQVECPKAPVEWNNPPSEKGLIVG
HFSGI
KYKGEKAQASEVDVNKMCCWVSKFKDAMRRYQGIQTCKIPGQVLSDLDAKIKAYNLTVEGVEGFVRYS
RVT
KQHVAAFLKELRHSKQYENVNLIHYILTDKRVDIQHLEKDLVKDFKALVESAHRMRQGHMQNVKYILYQ
LLK KHGHGPDGPDILTVKTGSMQLYDDSFRKIYTDLGWKFTPL

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

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

135 Scoring Peptide Sequence Tags

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

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

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

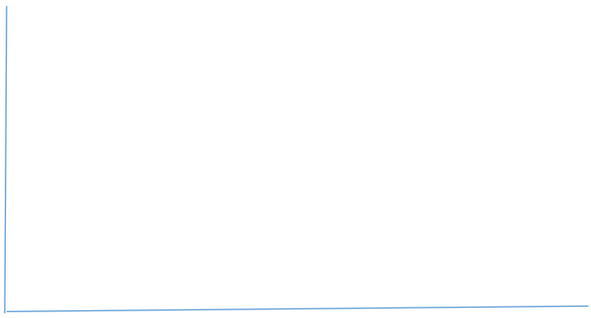


Figure 25 root mean square error

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

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

So, the updated relationship is:

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

136 In silico Protein Fragmentation

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

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

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

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

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

Consider a random protein sequence from DB:

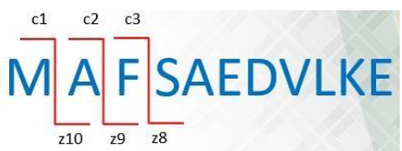


Figure 26 random protein sequence

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

137 In silico Fragment Comparison and Scoring

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

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

- Accumulate the score

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

Scoring scheme should also consider the errors in peak matching

138 Protein Sequence Database Search Algorithm

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

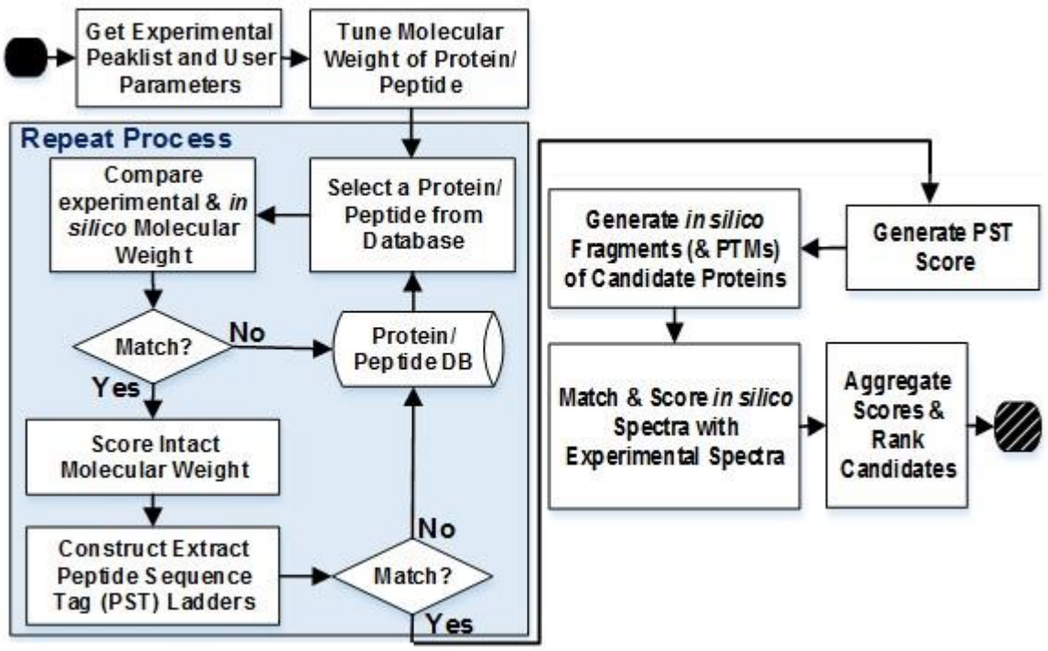


Figure 27 protein sequencing flowchart

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

139 Integrative Scoring Schemes

Three individual scores can be obtained:

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

For overall cumulative score computation:

- We simply sum the scores up (a linear function).

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

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

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

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

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

140 Large Scale Proteomics

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

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

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

TECHNIQUE

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

STEP BY STEP TECHNIQUE

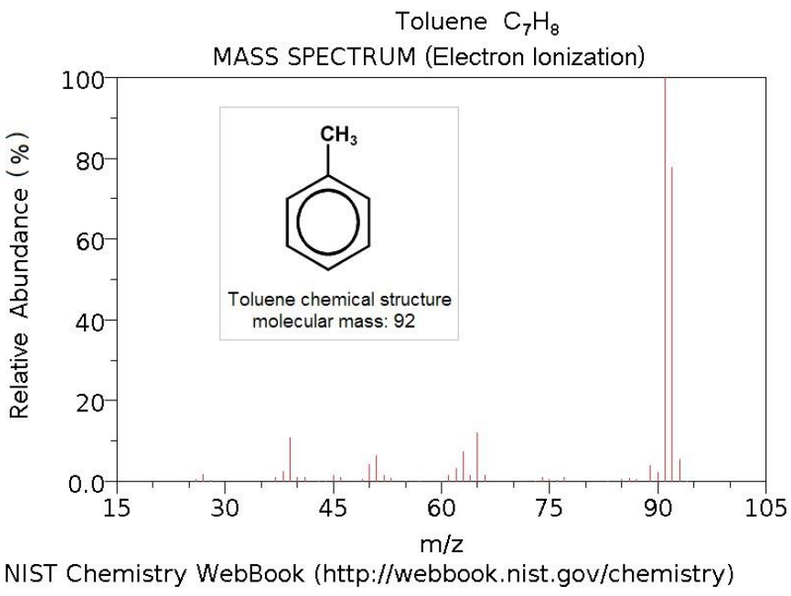
1. The instrument alternates between MS and MS/MS modes.
2. Three most intense peaks are chosen for MS/MS analysis.
3. After the initial MS scan, an MS/MS spectrum from peptide A is obtained by selectively fragmenting this mass only.
4. Next, a spectrum for peptide B is produced, followed by a recording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

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

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

141 Proteomics Data File Formats

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



Followings are the formats for proteomics data:

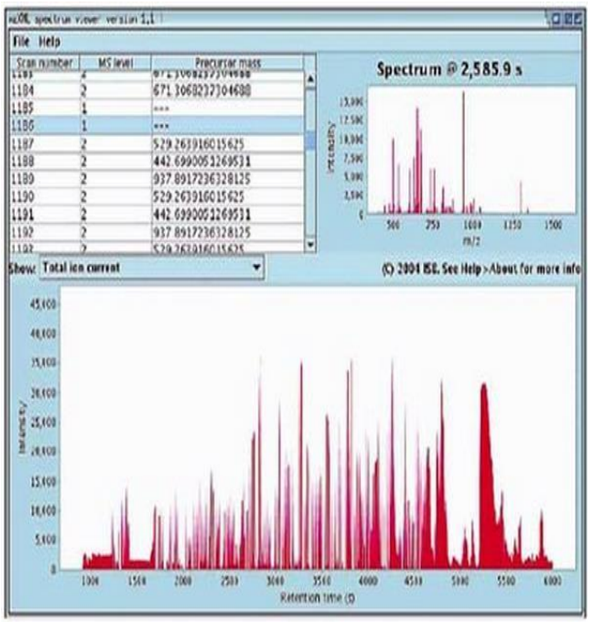
Agilent	.D/.YEP
Bruker	.BAF
ABI/Sciex	.WIFF
	.t2d
Thermo Xcalibur	.RAW
Micromass	
PerkinElmer	
Waters	
Shimadzu	.QGD

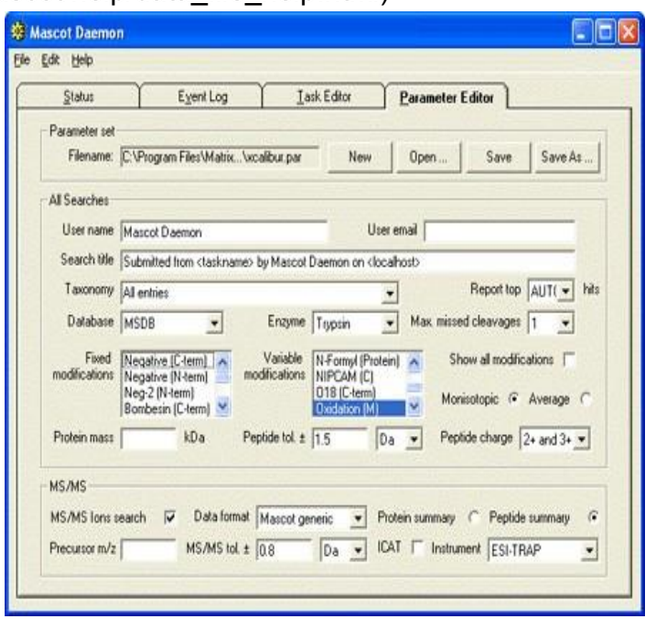
Figure 28 Formats used for proteomics data

OPEN FORMTAS:

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

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





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

142 RAW File Format

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

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

143 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

144 Open MS Data Formats

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

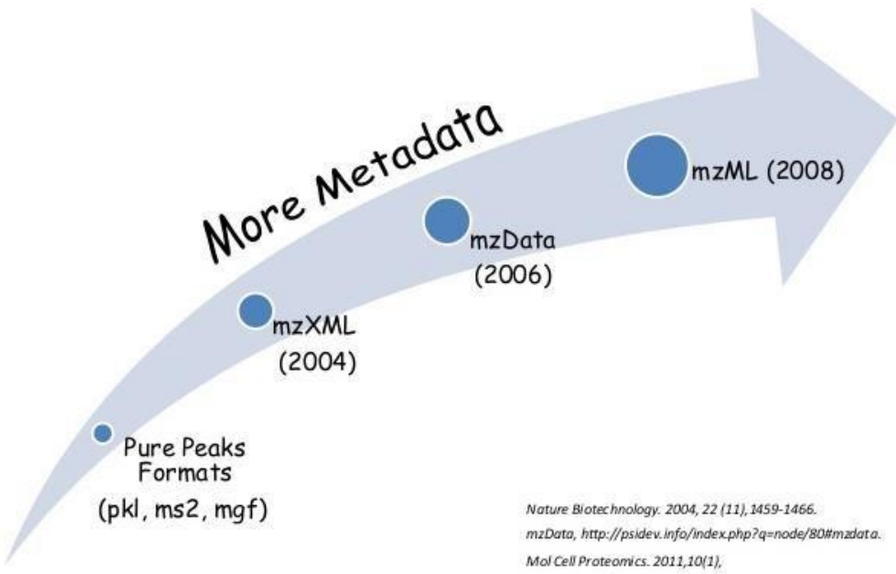
NEED

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

SOLUTION

- mzData was developed by HUPO-PSI

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



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

Figure 31 Formats used for open use

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

145 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

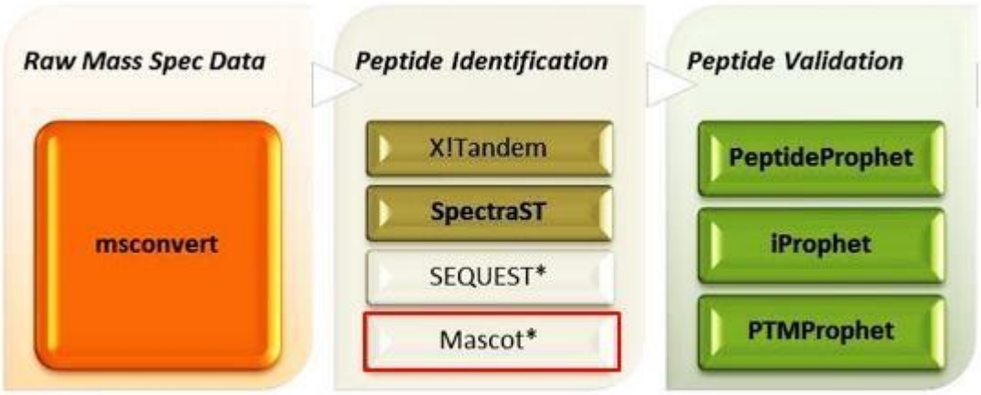


Figure 32 mzML http://tools.proteomecenter.org/software.php

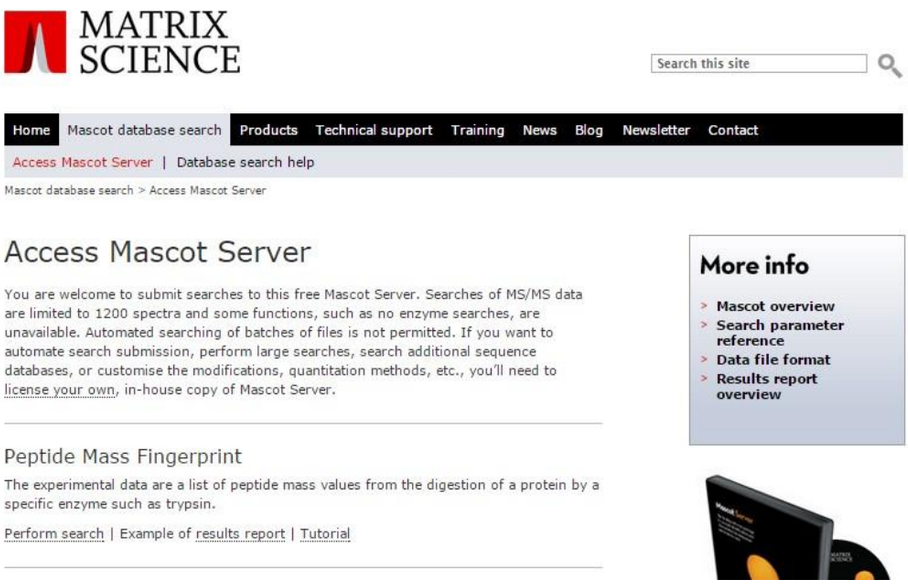
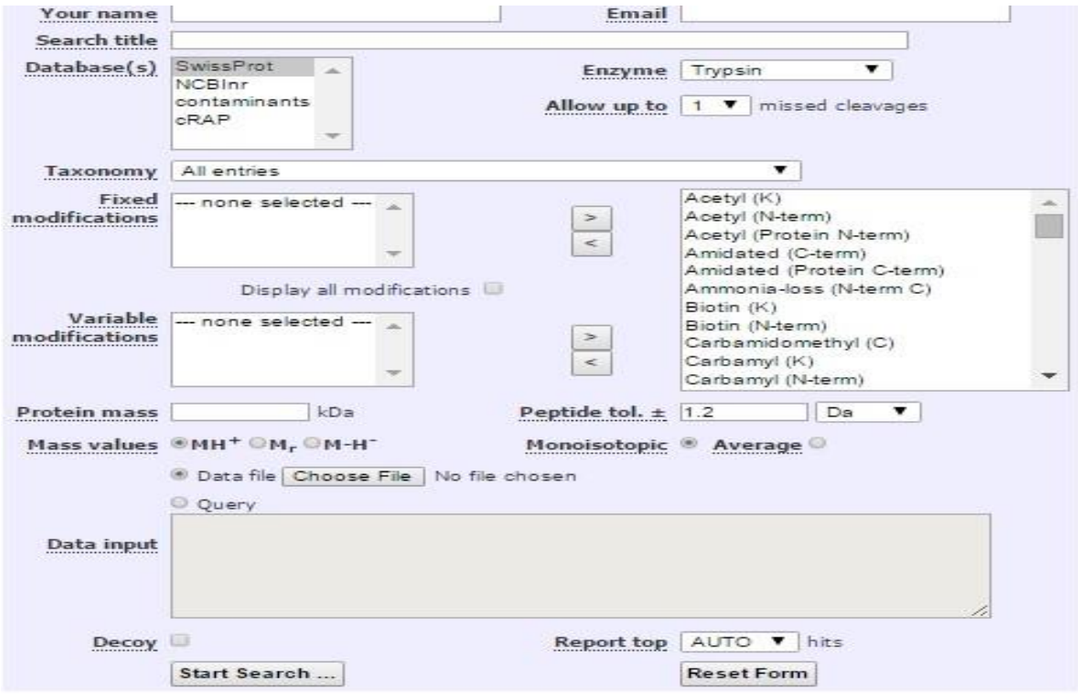
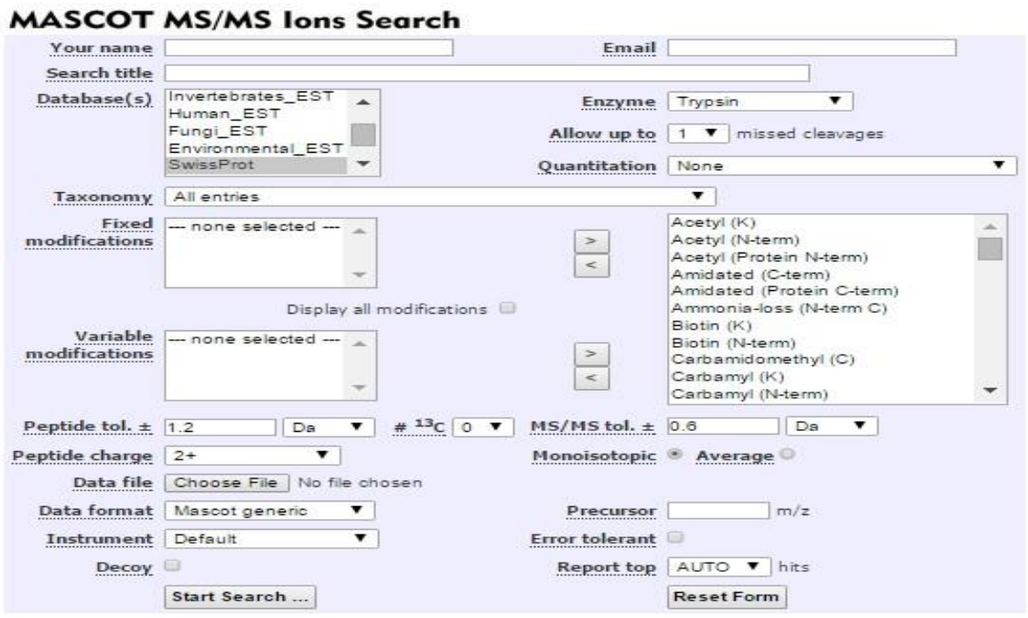


Figure 33
http://www.matrixscience.com/search_form_select.html

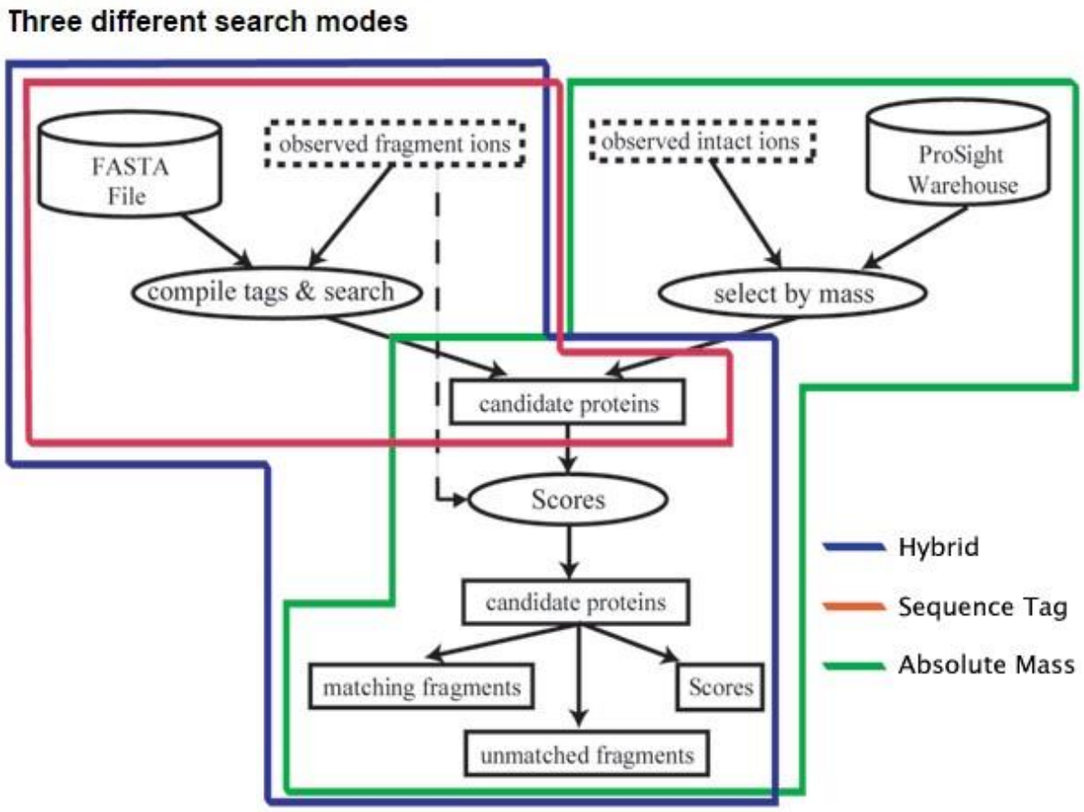




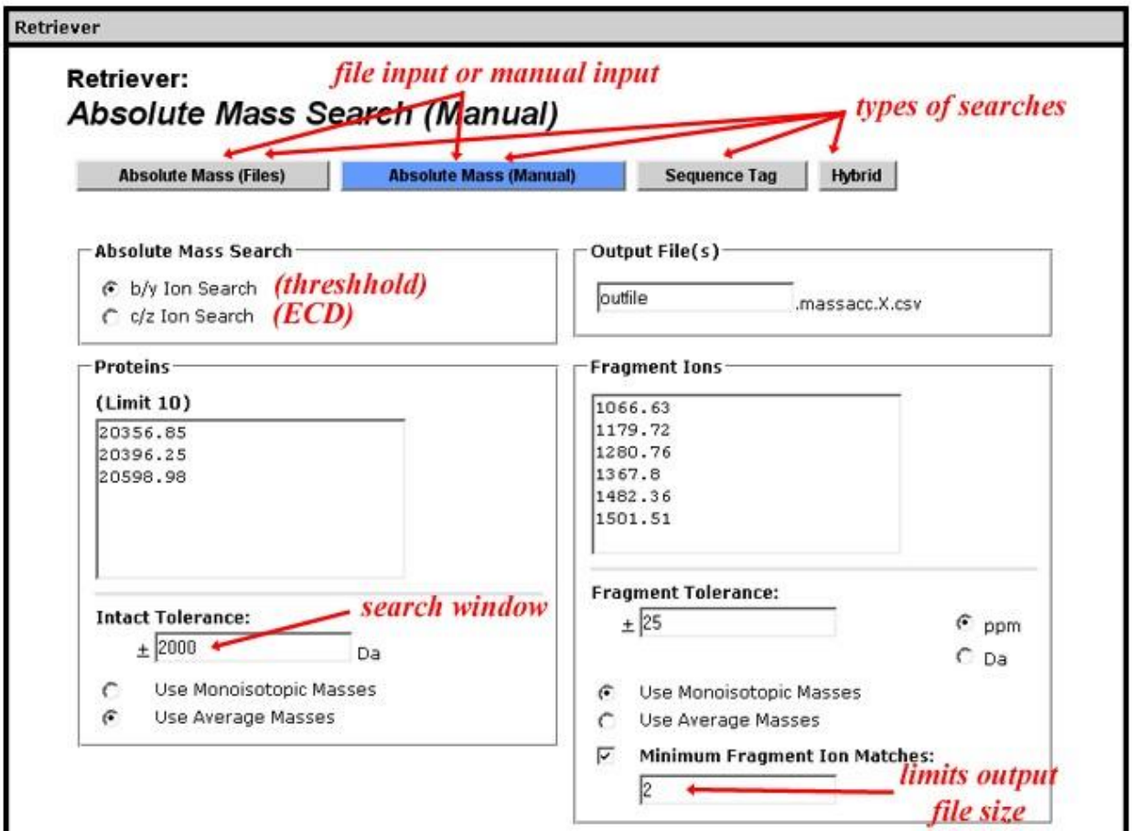
Mascot is the most widely used online search tool for proteomics data. However, it lacks a batch processing mode. Also, it does not cater for top-down proteomics data.

146 Online Proteomics Tools - ProSight PTM

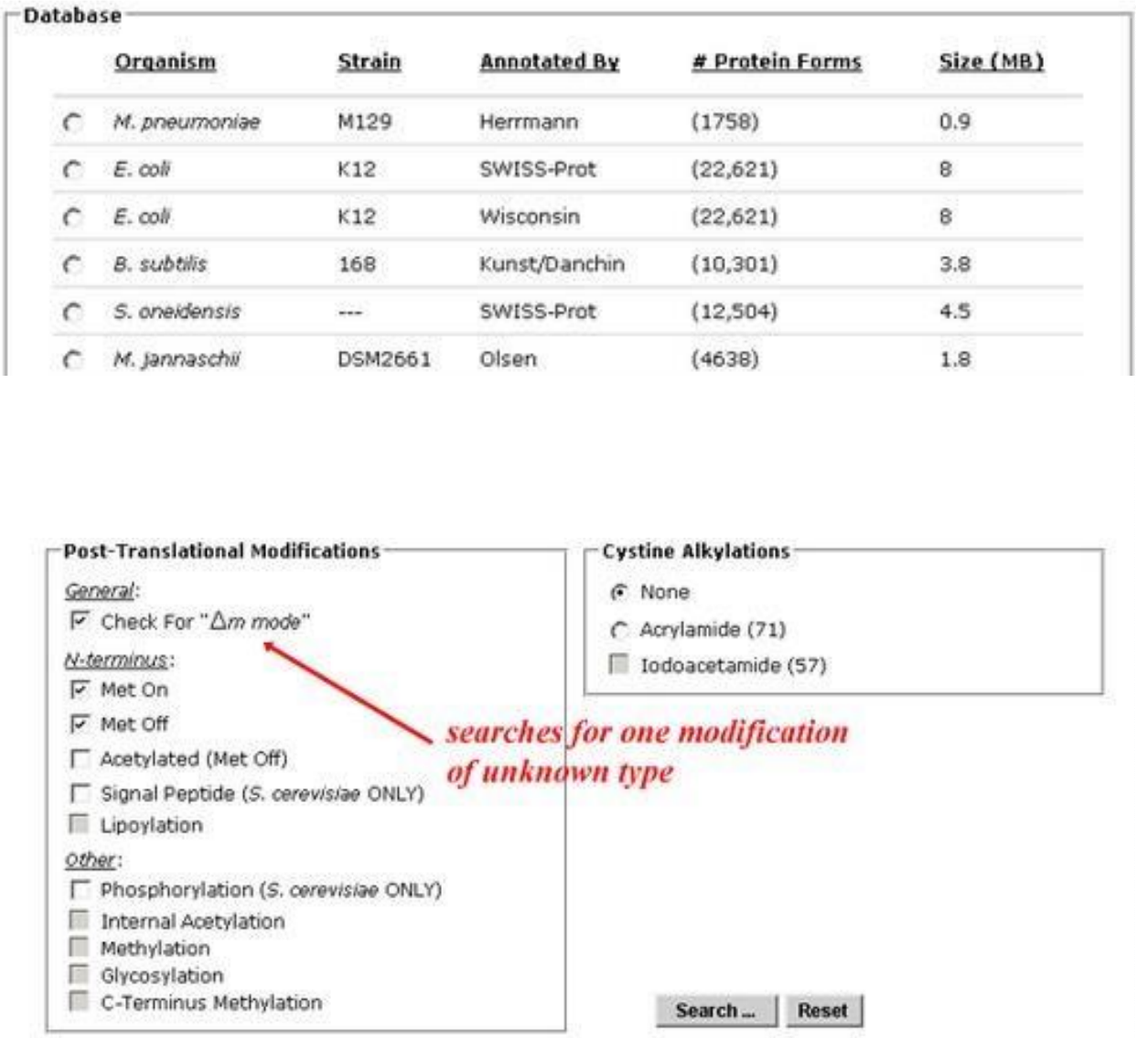
Kelleher et al have developed an online Top down Proteomics Search Engine. “Prosight PTM”. ProsightPTM searches top down proteomics data and reports the precursor protein



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



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



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

147 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

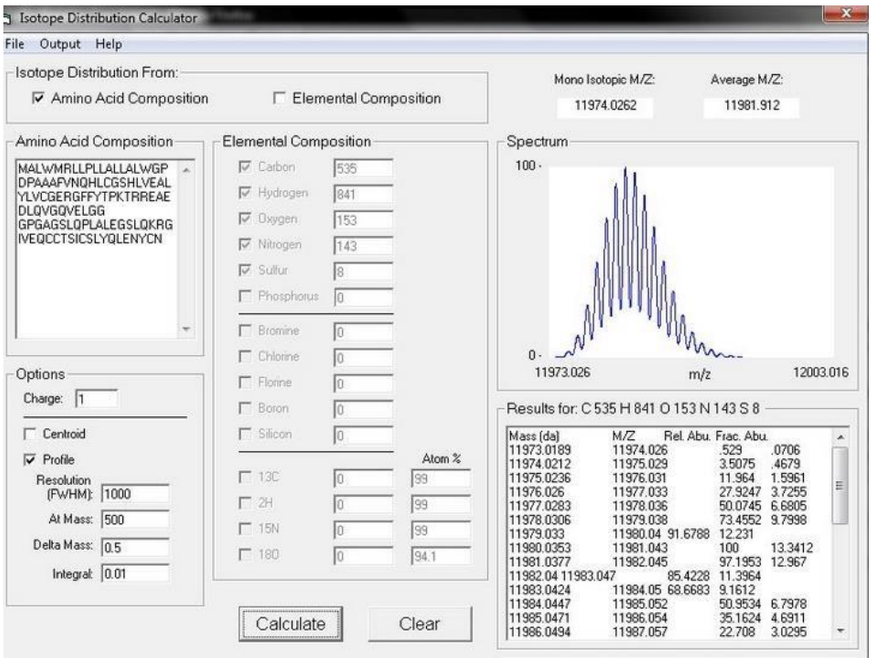


Figure 35 Detecting Monoisotopic Peaks

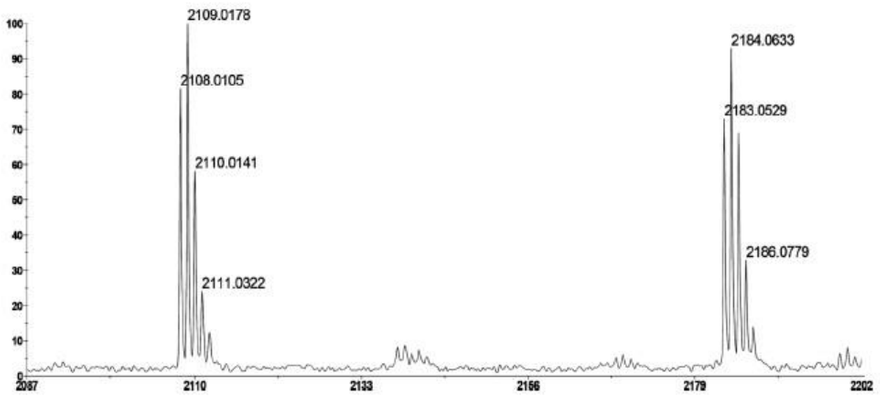


Figure 36 Detecting Monoisotopic Peaks

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

148 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

149 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



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.

150 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

151 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…

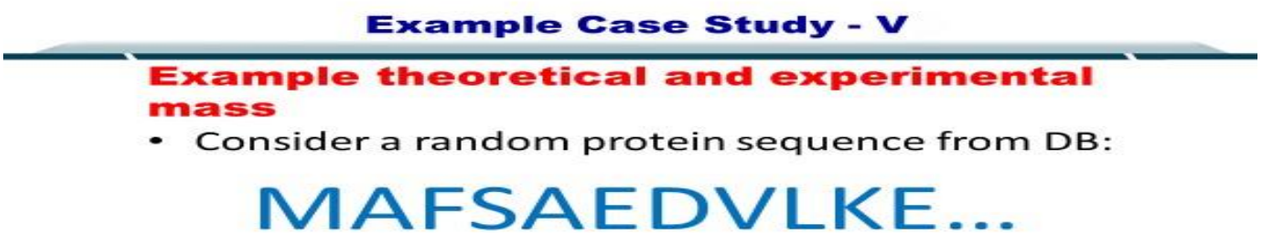
$$M_{Score} = \frac{1}{\sqrt{(M_{Exp} - MT)^2}}$$

Figure 38 Example intact protein mass score

MAFSAEDVLKE…

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

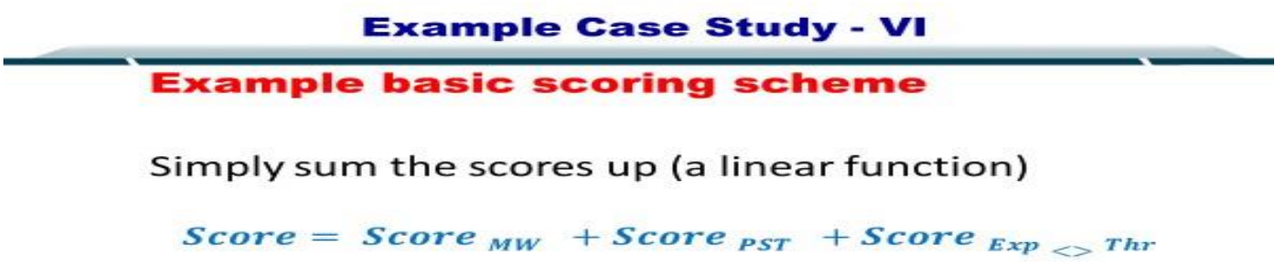
Figure 39 Example peptide sequence tags



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.

152 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.



For comprehensive scoring

Example Case Study - VI

Example comprehensive scoring scheme

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

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

153 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.

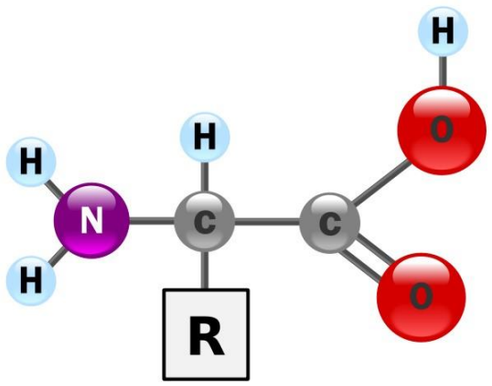


Figure 0.1 chemical structure of amino acid

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

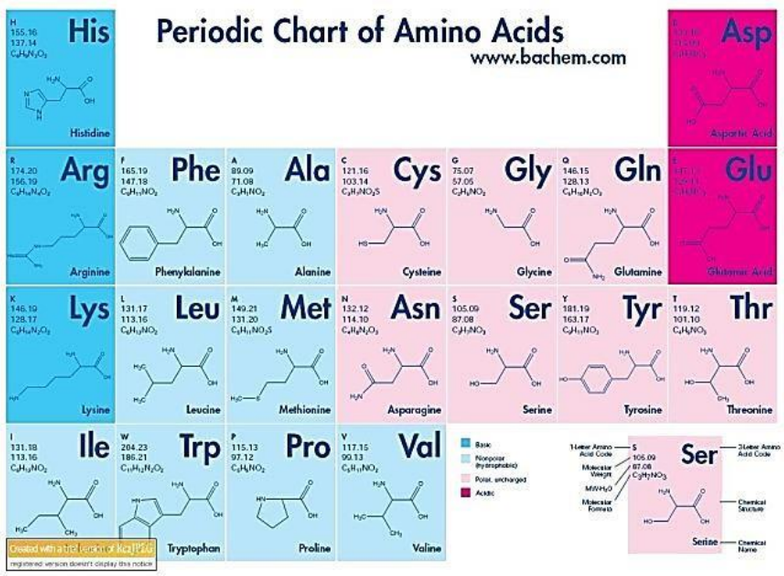


Figure 0.2 periodic chart of amino acid

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

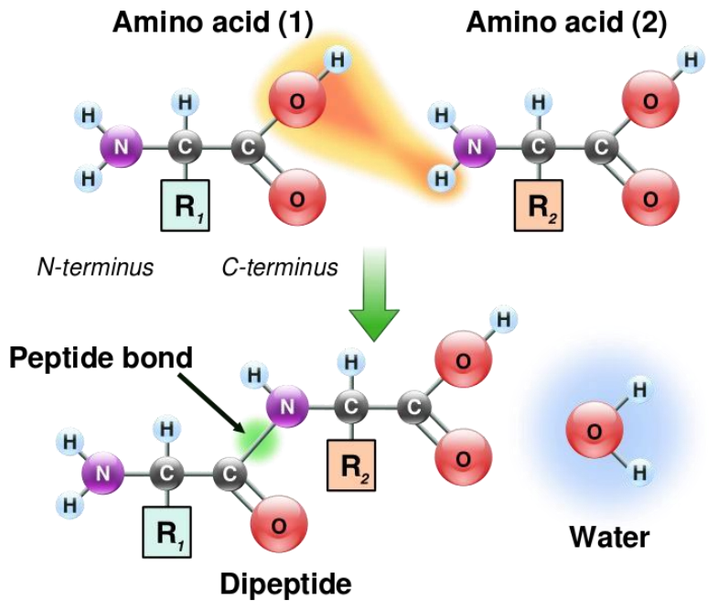


Figure 0.3 polymerization of amino acids

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.

154 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).

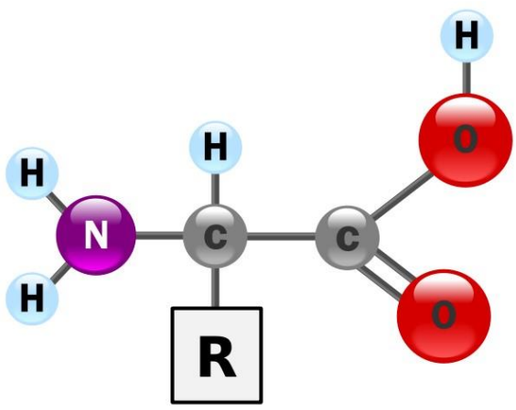


Figure 4 R group in amino acid

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.

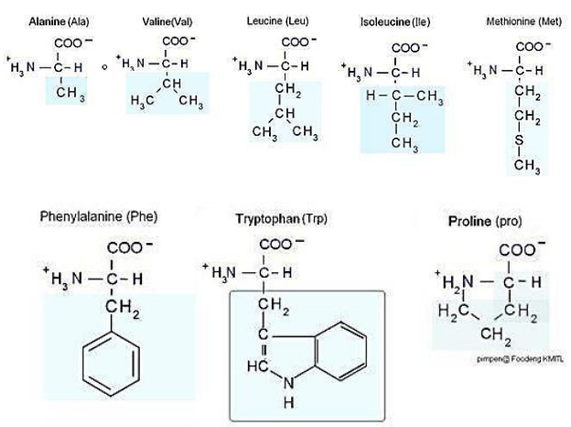


Figure 5 hydrophilic group

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.

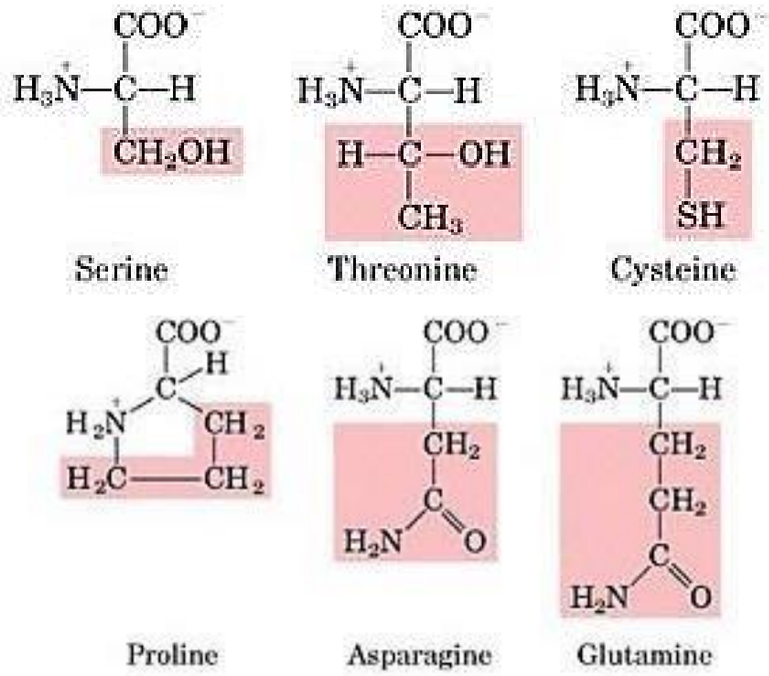


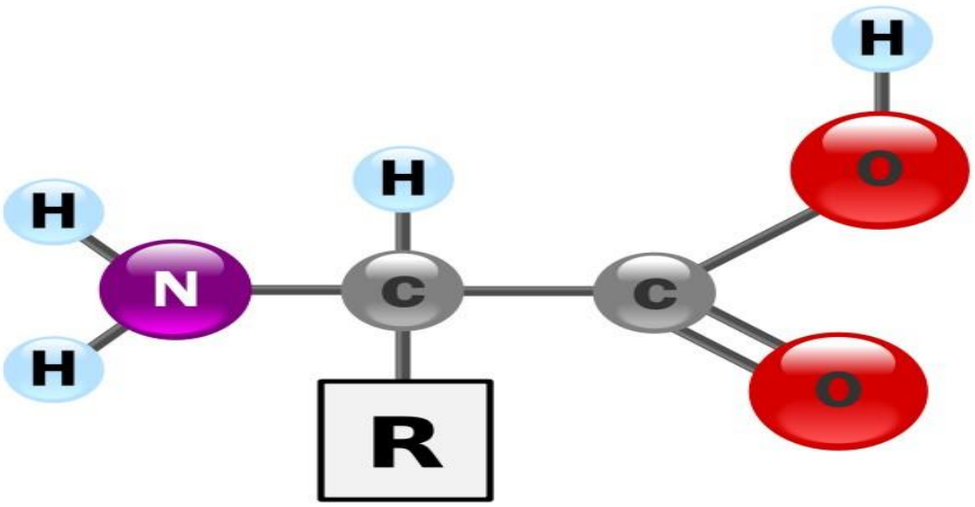
Figure 6 Polar amino acids

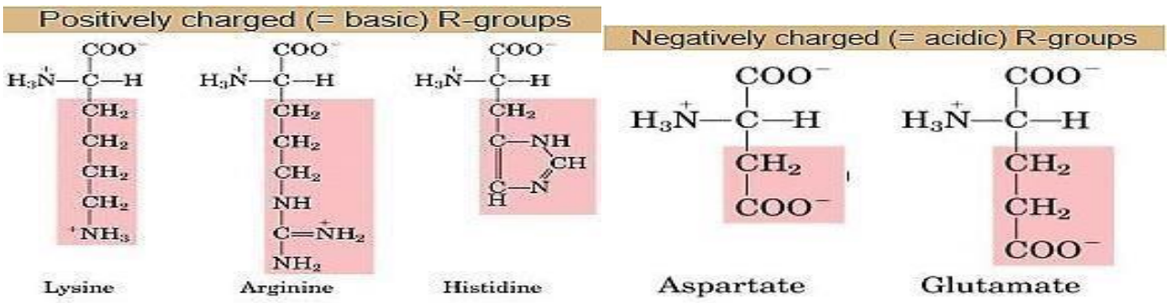
Amino acids have unique properties such as polarity, charge states and interactions with water. Each of these properties describes the overall characteristic of an amino acid.

155 Properties of Amino Acids – III

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

- **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)**





Change in Charge

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

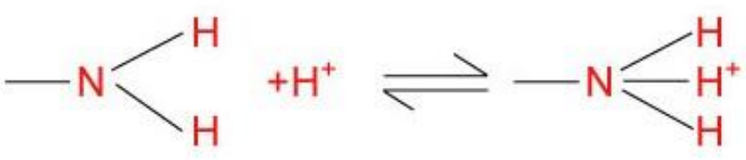
Upon polymerization of amino acids into polypeptide chains, charged amino acids get neutralized. At pH=7, five amino acids are charged, 2 negatively and 3 positively.

156 Properties of Amino Acids – IV

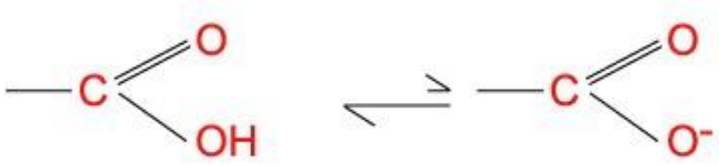
Some amino acids are positively charged and some have negative charge. pK is the values for an amino acid is the pH at which exactly half of the chargeable group is charged.

	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.



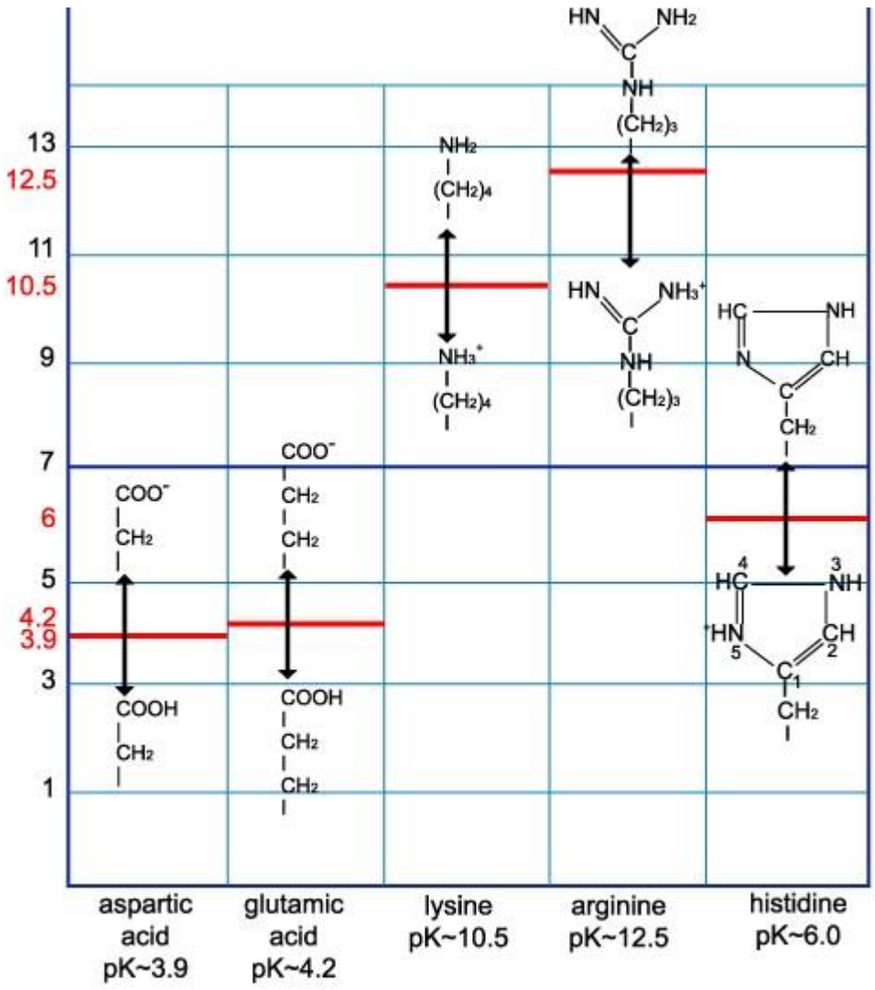


Figure 9 properties of amino acids according to pK and PH.

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

157 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.

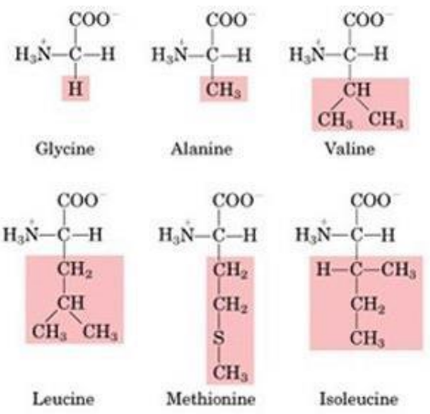


Figure 10 Aliphatic Amino Acids (Non polar C and H chains)

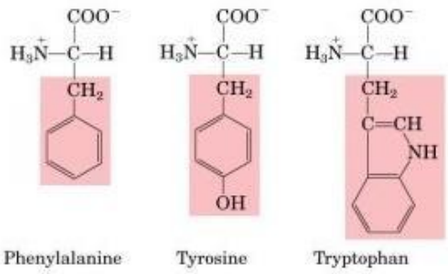


Figure 11 Aromatic R groups

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

- Chemically inert,
- Poorly soluble in water

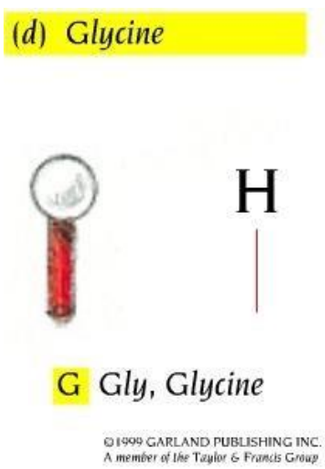
However, side chains containing organic acids are very different. They are chemically reactive and Soluble in water. Elemental composition plays a very important role in determining properties of amino acids. Solubility and reactivity are key factors participating in protein folding.

158 Structural Traits of Amino Acids – 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.



EXAMPLE-2: Proline

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

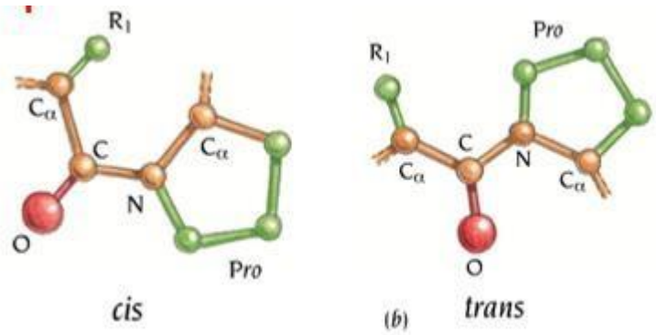


Figure 12 cis and Tran's form of proline

EXAMPLE-3: Cystine

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

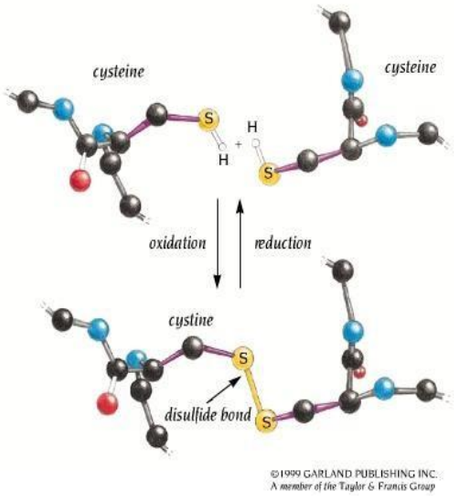


Figure 13 cystine

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

159 Structural Traits of Amino Acids – 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!

160 Structural Traits of Amino Acids – 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.

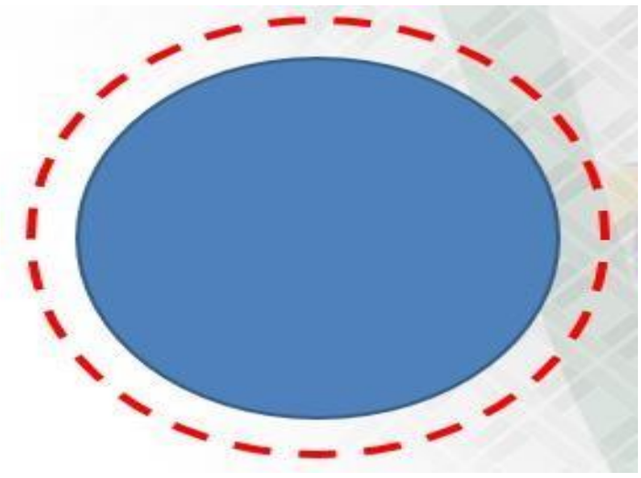


Figure 0.44 Organization of core and surface in a protein

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.

161 Structural Traits of Amino Acids – 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.

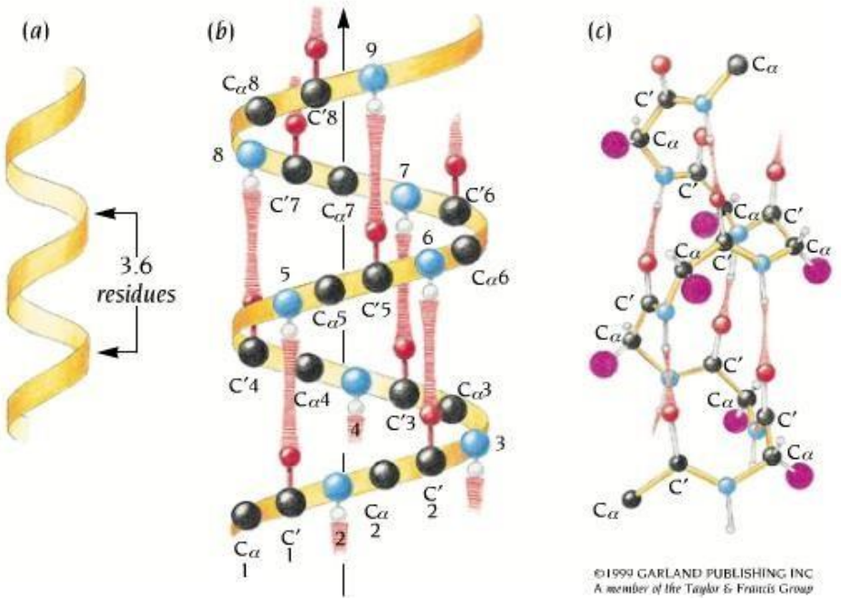


*Figure 0.55 Alpha Helix C = black **O = red N = blue***

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.

162 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.

Background:

- **Proteins comprise 20 different amino acids**
- **Amino acids polymerize and form protein molecules**
- **Proteins fold together to take 3D forms** Introduction:

- **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**

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. Folding of Proteins

- **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 How do we know the final folding state of a protein?

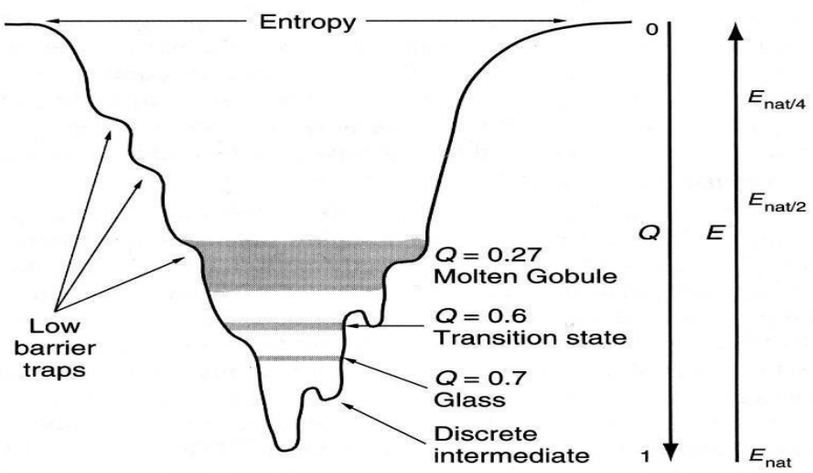


Figure 19.16 Cross section through a folding funnel. *E* corresponds to free energy. [Courtesy of P. G. Wolynes]

Proteins fold spontaneously in water. Proteins fold to achieve thermodynamic stability. Proteins fold to organize themselves for performing functions in cells.

163 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.

Background:

Proteins fold spontaneously
Proteins fold to achieve thermodynamic stability
Proteins fold to organize themselves for performing functions in cells

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)

Why study folding

Proteins are the functional machines in cells
Dysregulated protein expressions are a major cause of disease
Understanding protein folding helps design suitable drugs

Computational folding of proteins

- **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**
- **How do we study folding, computationally?** o **First, we collect clues & evidences from experimentally reported structures** o **We utilize these observations to analyze unknown structures**

Conclusions

- **Given algorithms and procedures to fold a protein, we can fold amino acid chains to form 3D proteins**
- **This can help us study misfoldings, interactions between drugs and proteins etc.**

Dobson, C. M. (2003). "Protein folding and misfolding." Nature **426**(6968): 884-890

164 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 $100^3 = 5 \times 10^{47}$ will be the combination. If it take $1/10^{th}$ of a Nano-second (10^{-10}), then to compute all the folding possibilities will take 1.6×10^{30} years.

In fact, it take a protein less than a second to fold. It's the Amazing speed of folding.

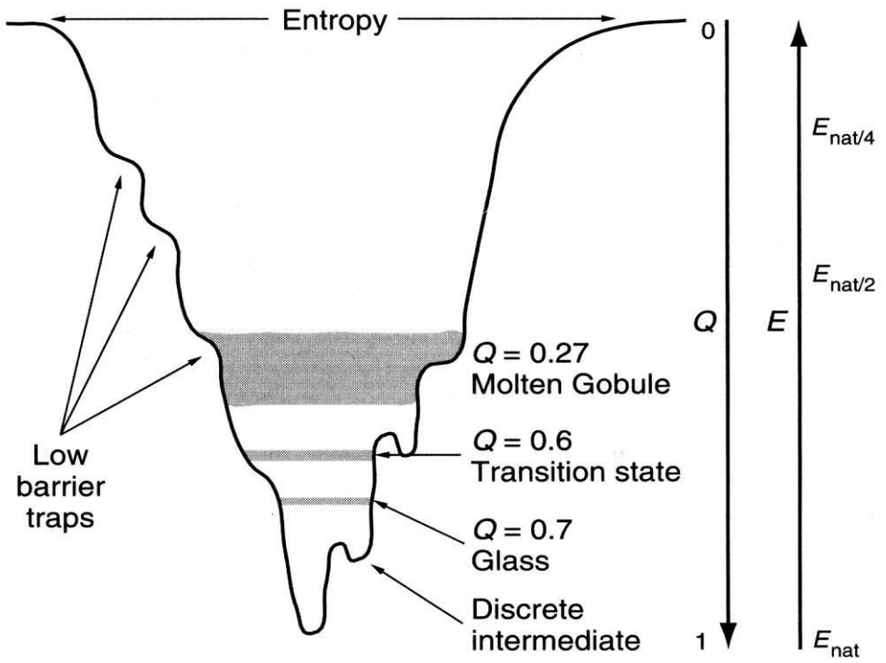


Figure 19.16 Cross section through a folding funnel. *E* corresponds to free energy. [Courtesy of P. G. Wolynes]

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.

165 Process 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

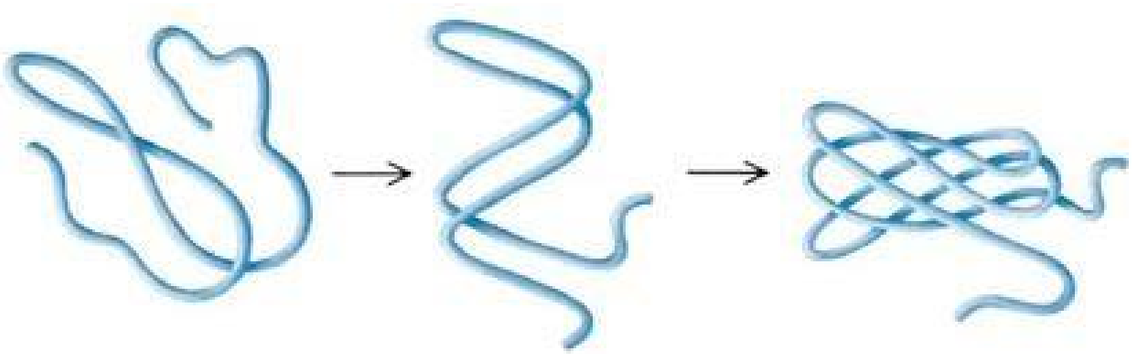


Figure 0.88 Protein folding

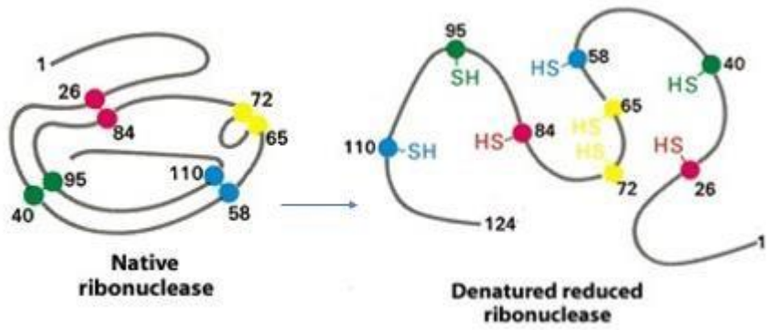


Figure 19 Anfinsen's Experiment

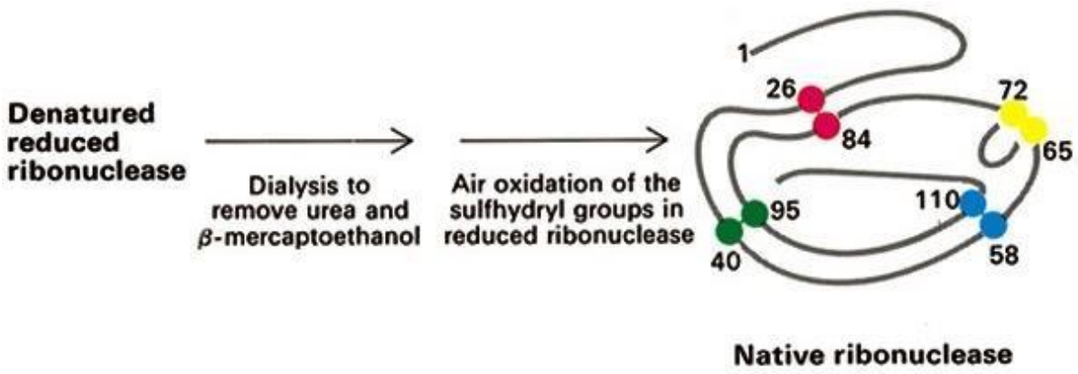


Figure 20 Anfinsen's Experiment

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

166 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

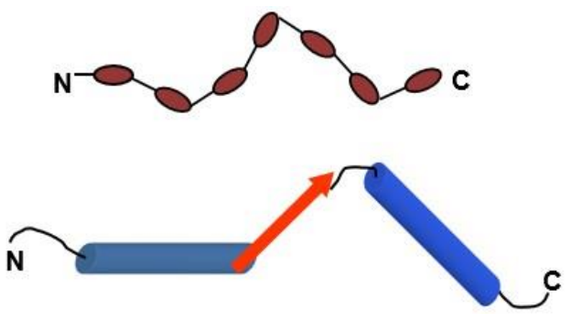


Figure 21 Step 1: Formation of secondary structures



Figure 0.92 Step 2: Arrangement of secondary structures

NUCLEAR CONDENSATION MODEL

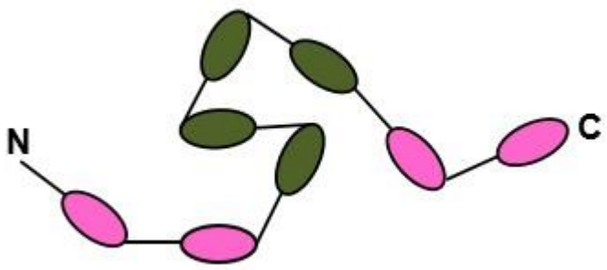


Figure 20.10 Step 1: Formation of a Hydrophobic Core

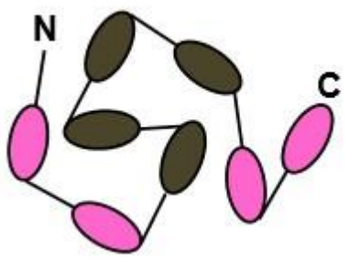


*Figure 20.11 Step 2: Including remaining amino acids and **expanding the nucleus***

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.

167 Protein Structures

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.

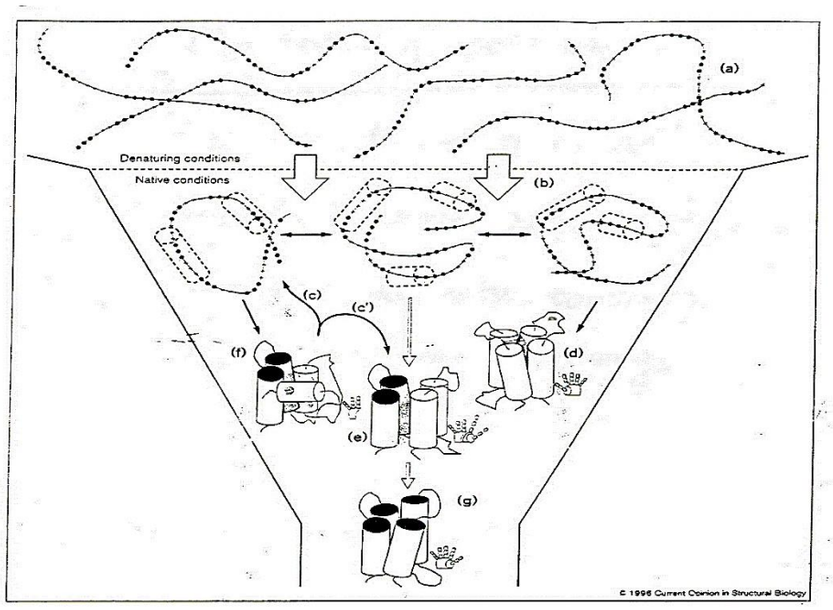


Figure 25 Folding funnel

Bond Type	kJ/mol
Covalent Bond	250
Electrostatic	5
van der Waals	5
Hydrogen bond	20

Figure 0.126 Energies of Various Bonds & Interactions

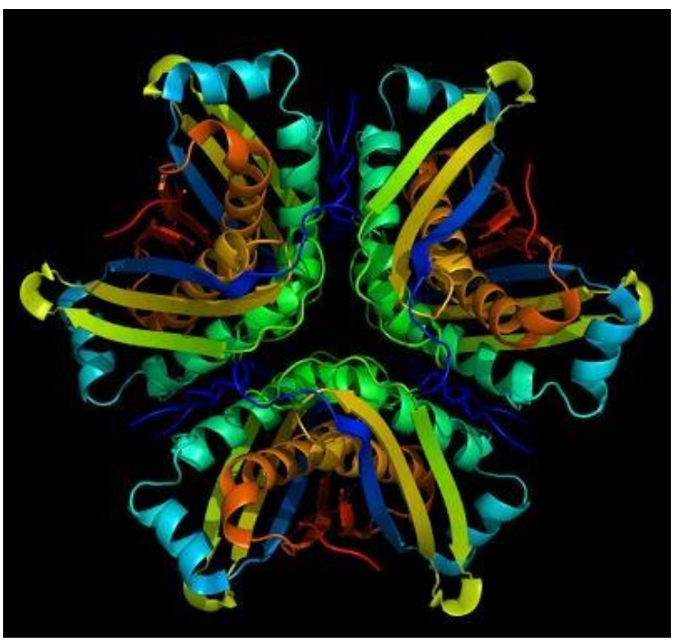


Figure 27 Hydro peroxide resistance protein OsmC (1vla)

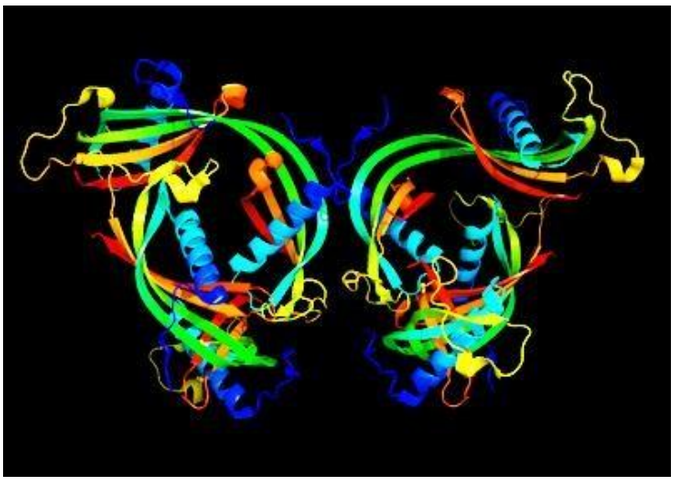


Figure 28 Cystatin – 3 (C) http://beautifulproteins.blogspot.com/

Protein structures are very complex yet they form spontaneously. We will investigate how to develop algorithms to predict such structures.

168 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

169 Primary Structure of Proteins

Protein structures are organized into 1', 2', 3' and 4' modular conformations. 1' structures are essentially the amino acid sequence of the proteins.

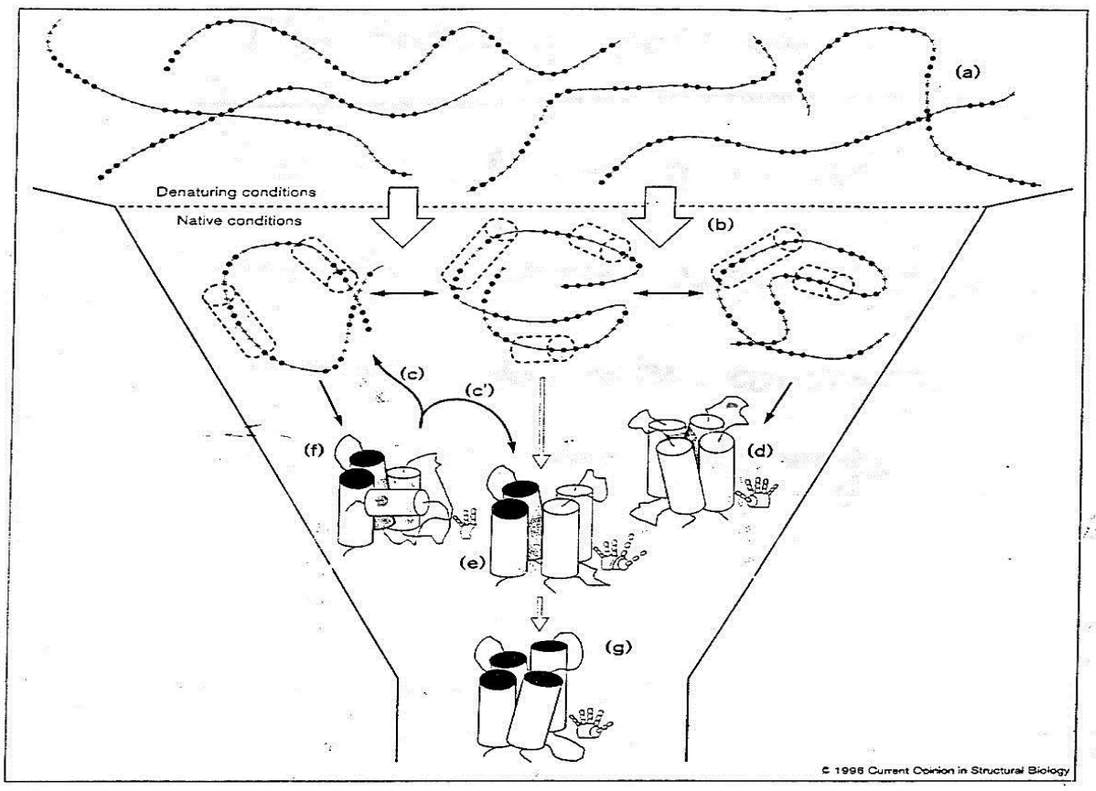


Figure 29 protein folding funnel

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

170 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.

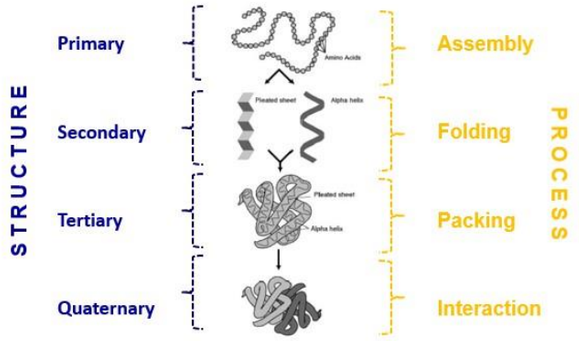


Figure 30.13 Organization of Secondary Structure

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.

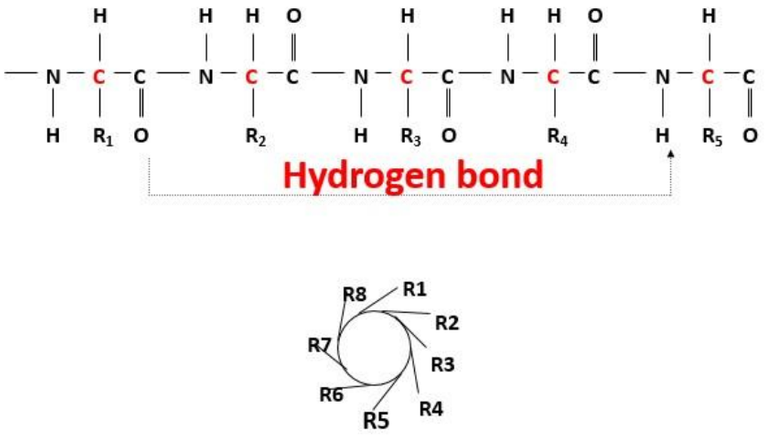


Figure 30.14 Forming Secondary Structure

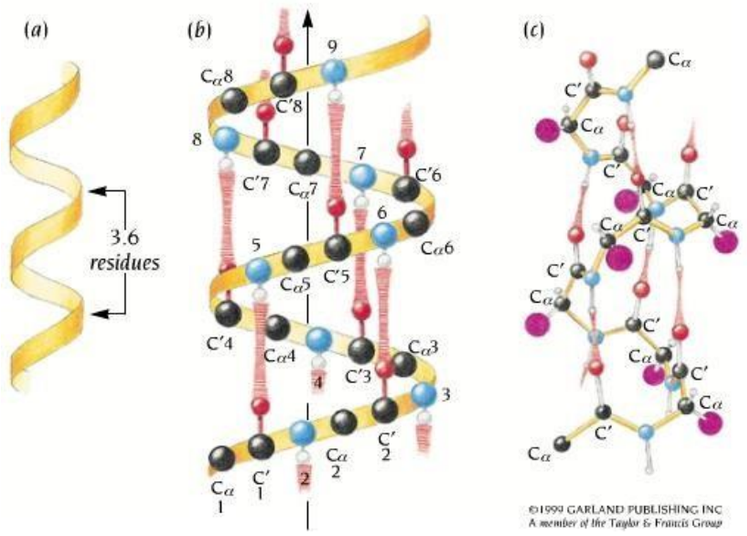


Figure 30.15 Types of Secondary Structures – Alpha Helix

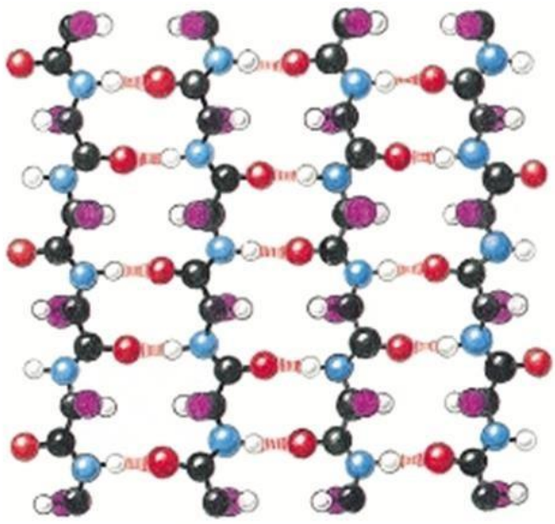


Figure 30.16 Types of Secondary Structures – Beta Sheets

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.

171 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.

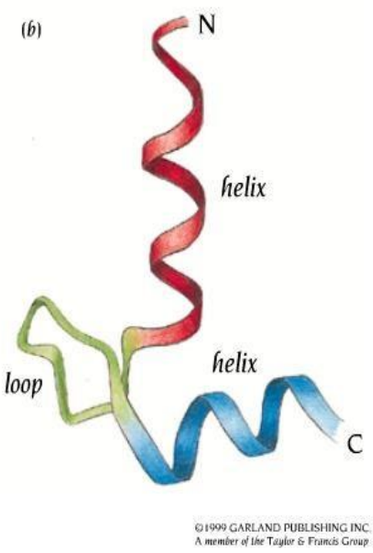


Figure 35 A Special Secondary Structure

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.

172 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.

173 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.

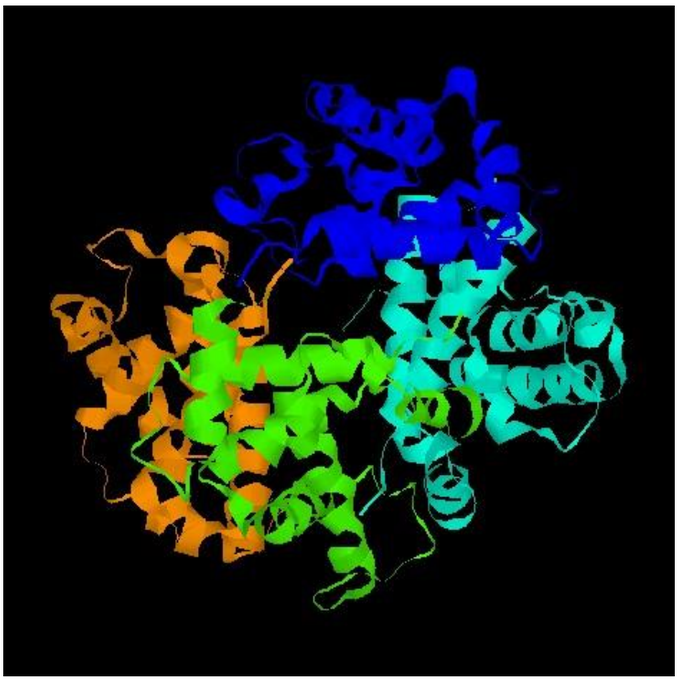


Figure 37 Example of Quaternary Structure See how 2' and 3' structures come together

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'

174 Introduction to Protein Bond Angles

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)

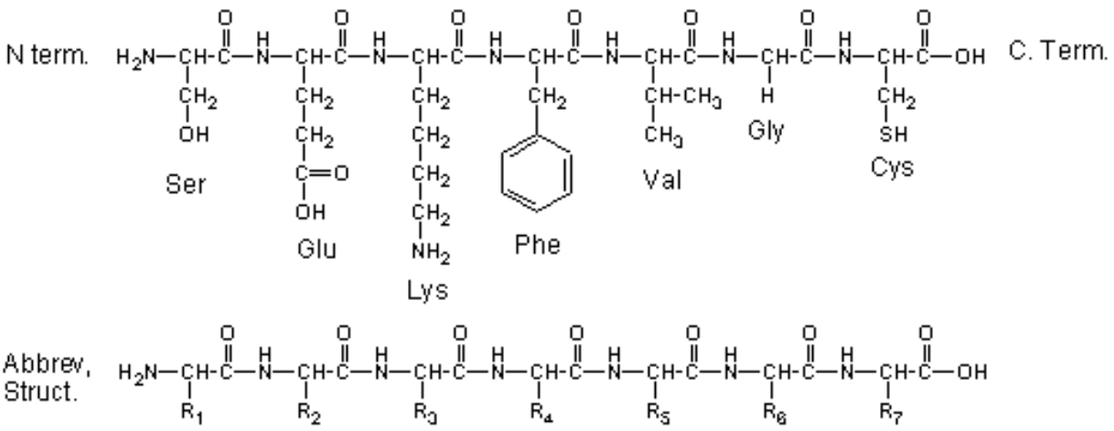


Figure 38 Linear Protein

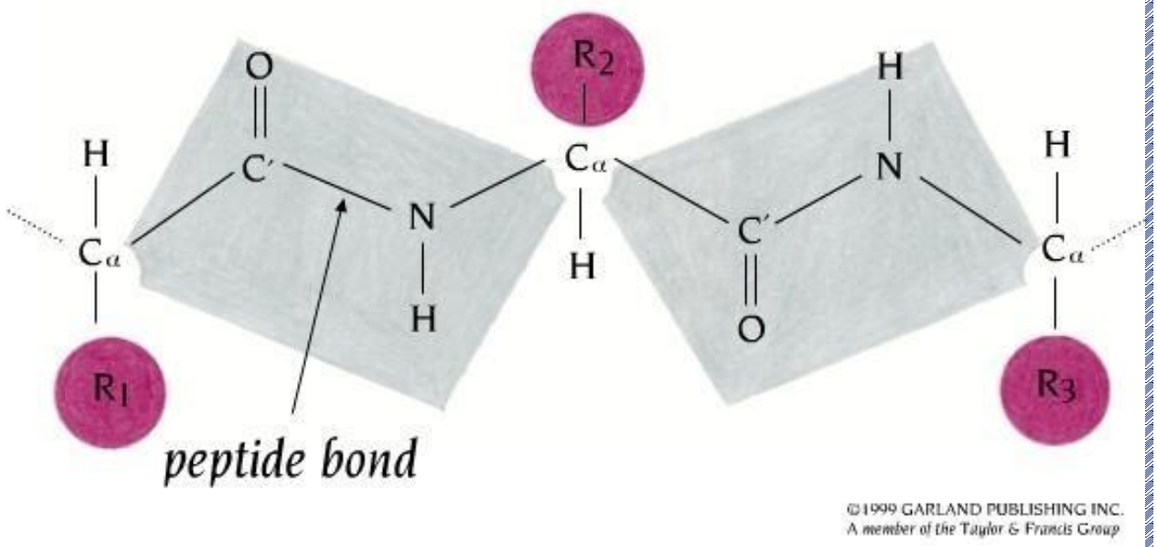


Figure 39 Formation of Planar Peptide Bond

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.

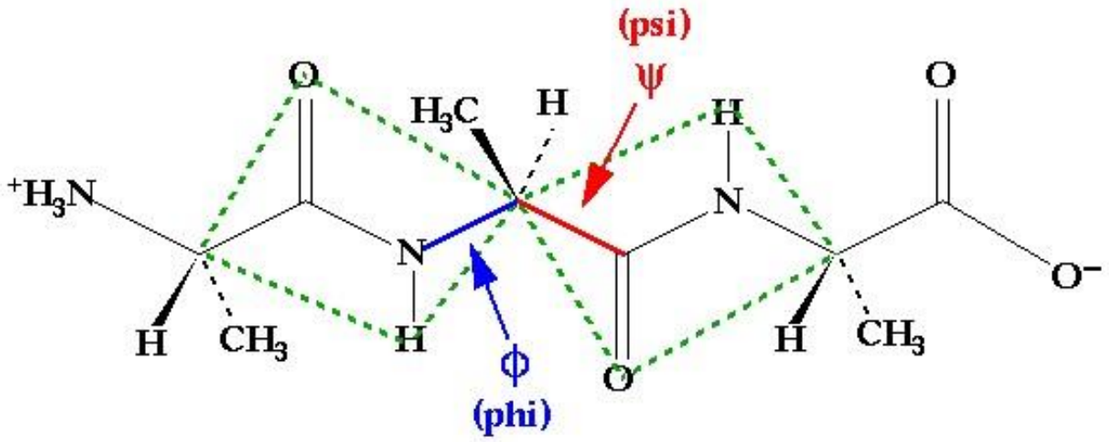


Figure 40 Protein after Folding: Phi and Psi Angles

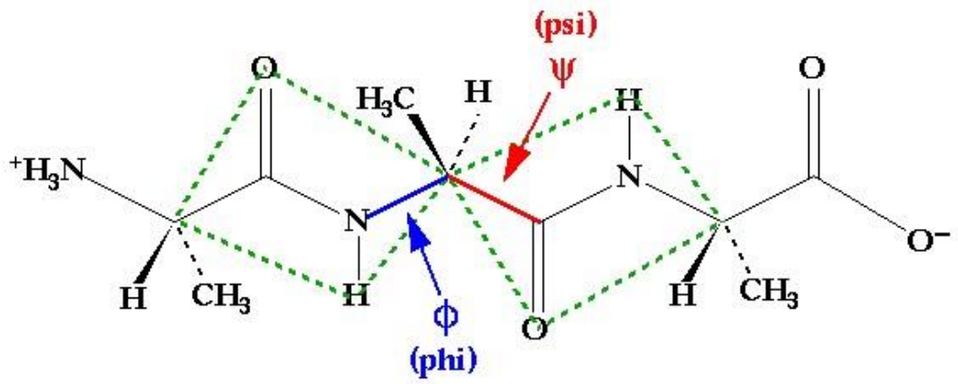


Figure 0.171 Protein after Folding: Phi and Psi Angles

Φ (*phi*, involving C'-N-C$^{\alpha}$-C') ψ (*psi*, involving N-C$^{\alpha}$-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.

175 Ramachandran Plot

Phi and Psi angles can be measured with in the folded structures like:

- φ - *phi*
- Involves C'-N-Cᵅ-C‘
- ψ – *psi*
- Involves N-Cᵅ-C'-N

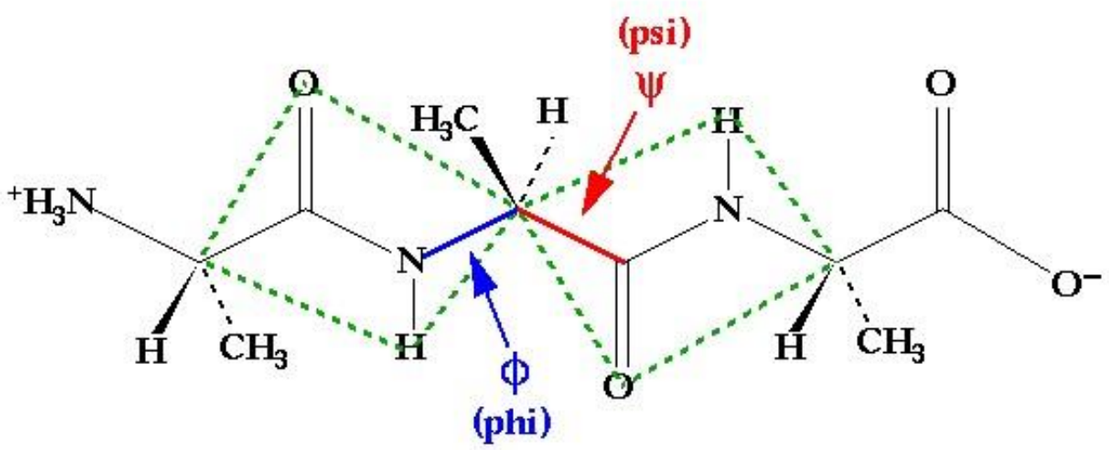


Figure 42 Phi and Psi Angles

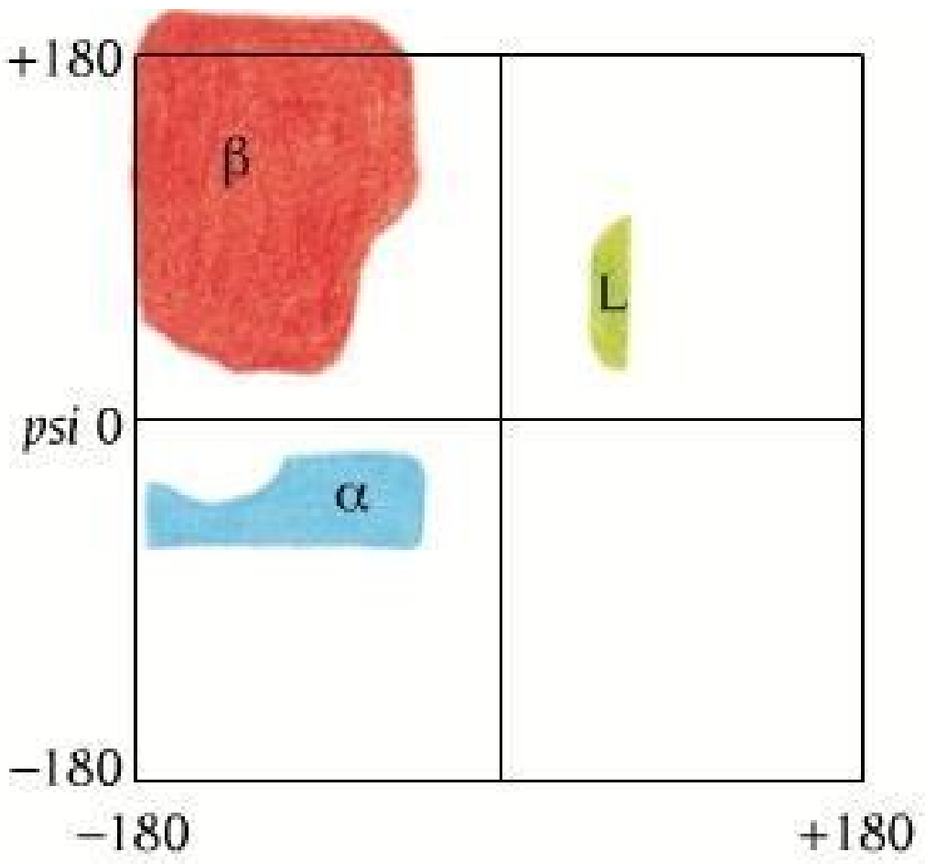
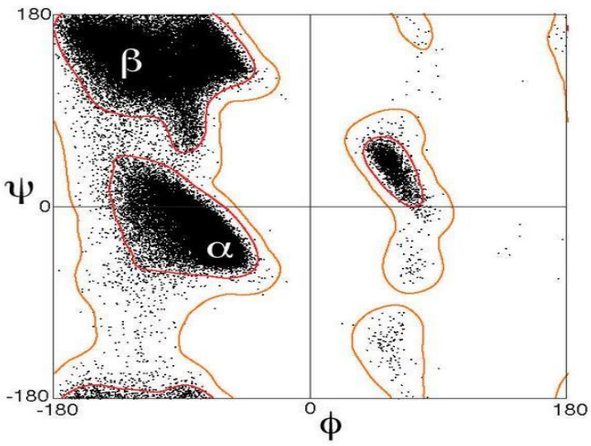


Figure 43 Allowable Phi and Psi Angles



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.

176 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

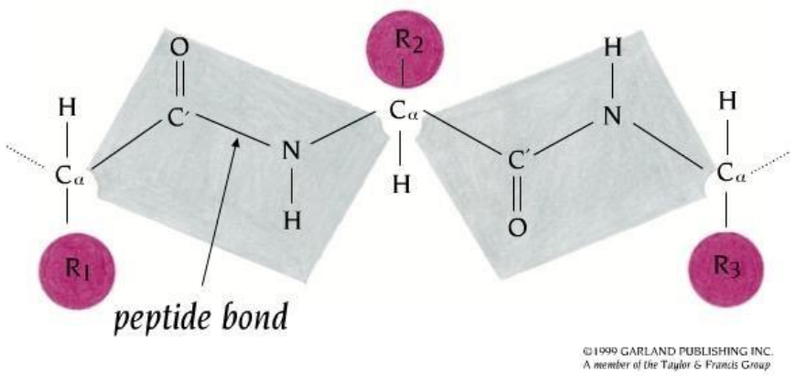


Figure 45 Omitting Planar bonds and Tracing C-Alpha atoms in backbone

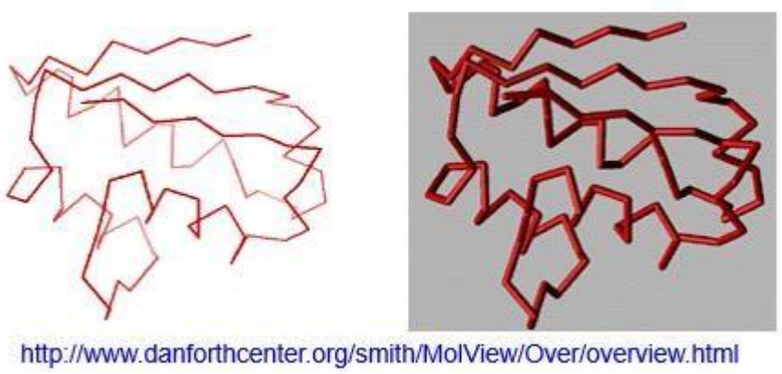


Figure 46 C-Alpha Backbone visualization

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.

177 Structure Visualization – II

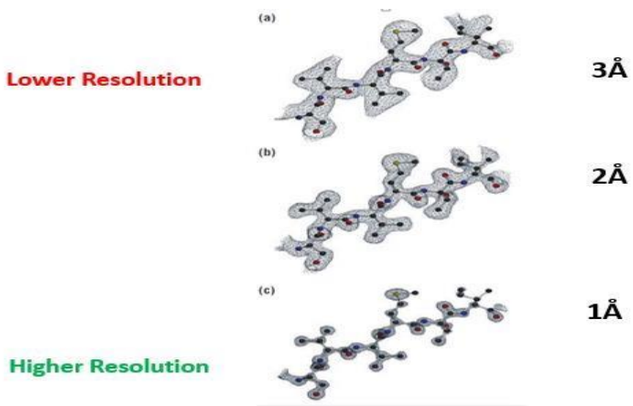
C-Alphas can be used to construct the backbone of a protein towards its visualization. We also need a representation of measurements for assigning the atomic distances. The ångström is used to express the size of atoms, molecules and extremely small biological structures, the lengths of chemical bonds, 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.

178 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

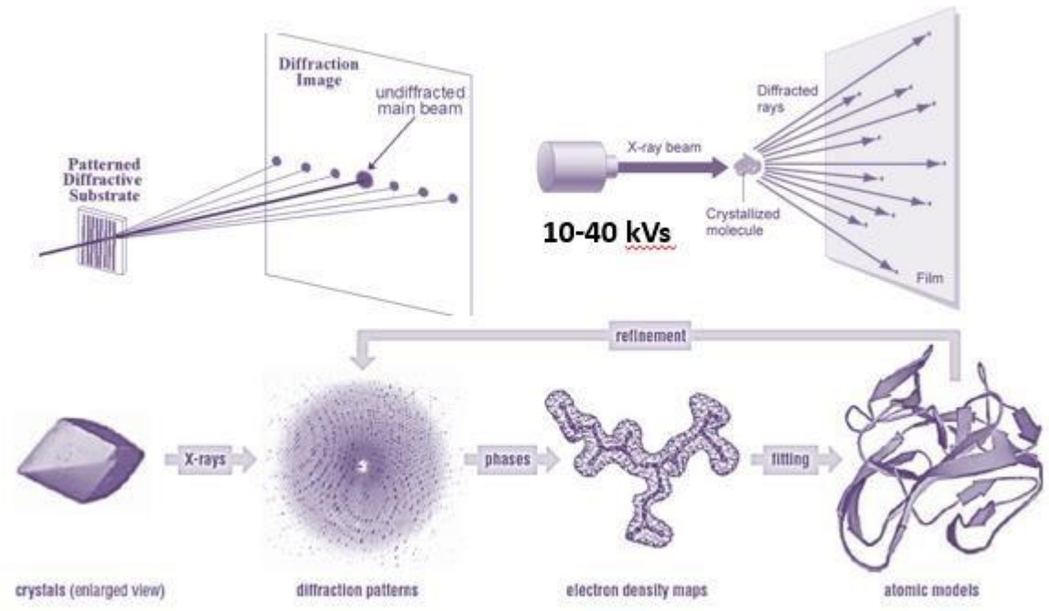
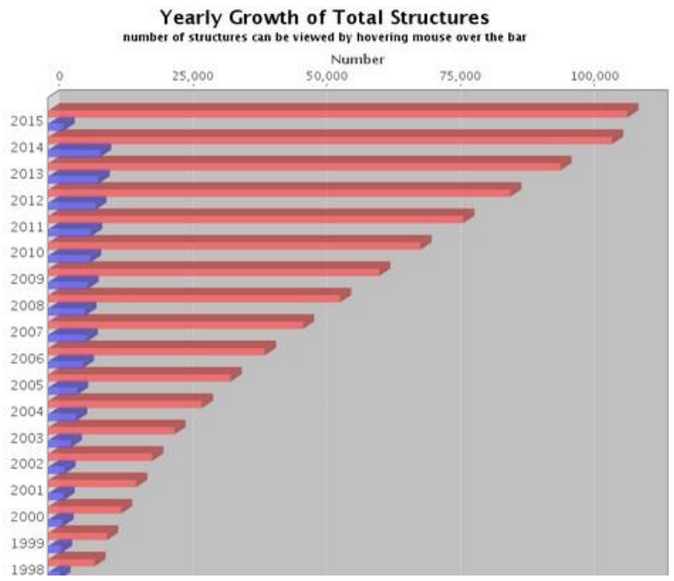


Figure 48 x-ray crystallography

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.

179 Protein Data Bank 2

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).

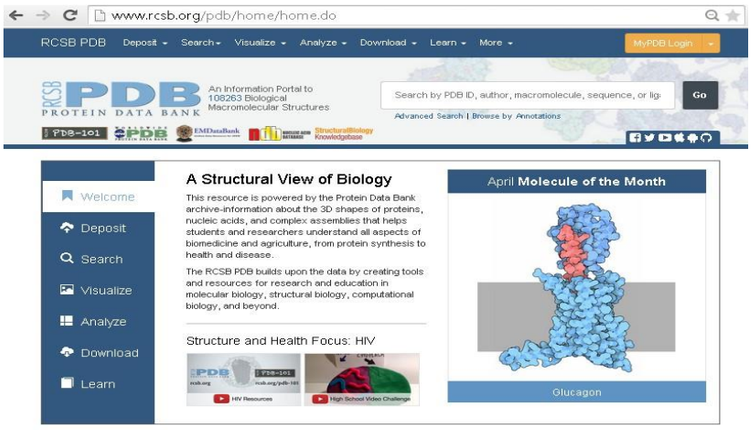
DBREF – Cross links to other databases.

SEQRES – The official amino acid sequence (protein, RNA or DNA) of the structure.

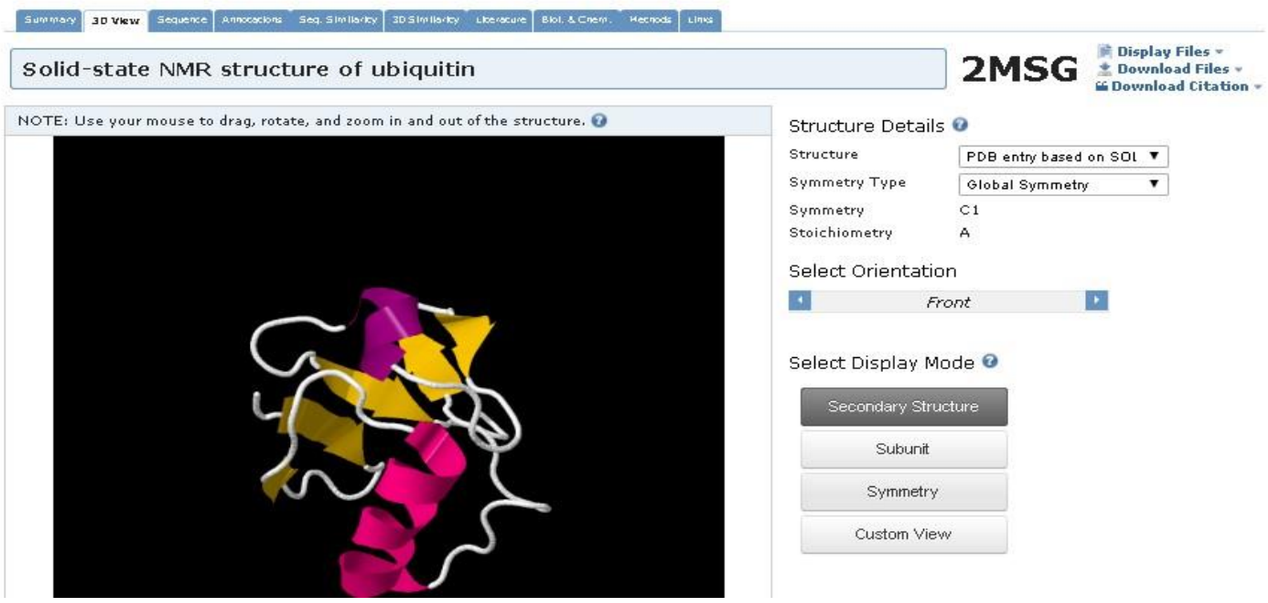
HELIX/SHEET – Details of the regions of secondary structure found in the protein.

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
```







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.

180 Visualization Techniques

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

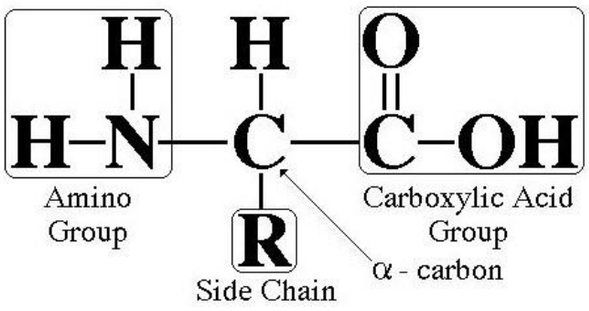
- To trace the backbone of a protein, CA atoms trace can be used
- Note that CA atoms have the side chains attached to them
- A coordinates can be found in the PDB file

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.

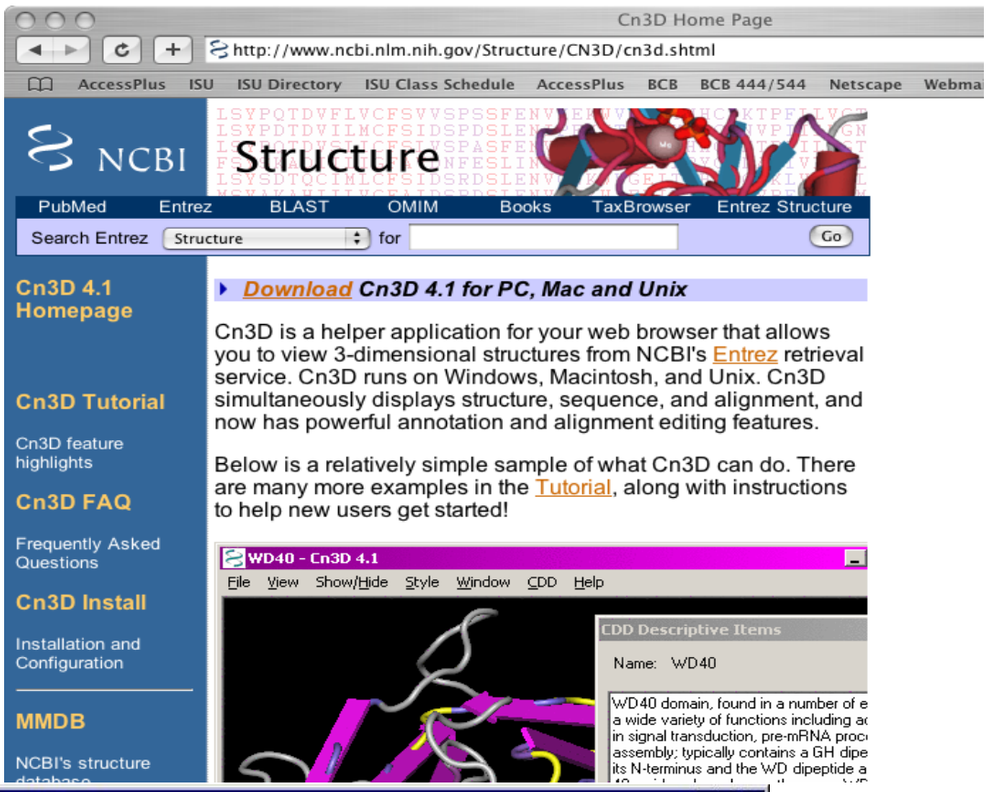
181 Online Resources for Protein Visualization

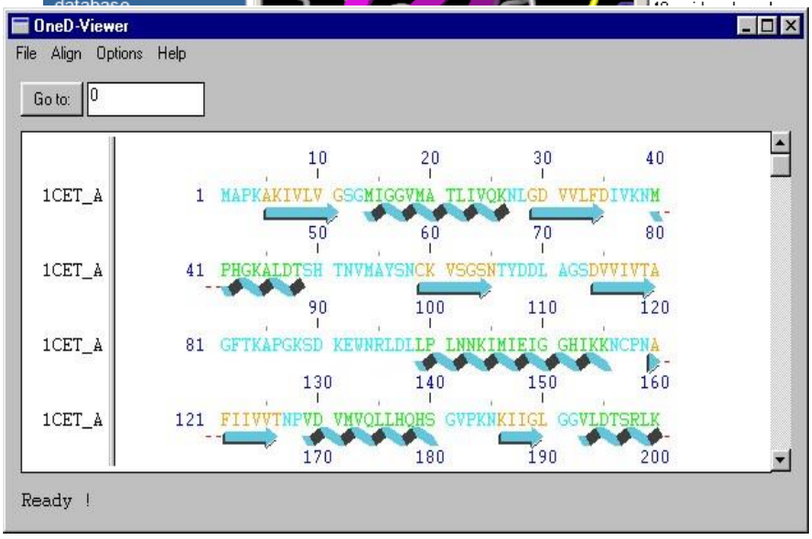
Protein structures can be visualized by tracing the CA atoms. CA Coordinates can be taken from PDB.

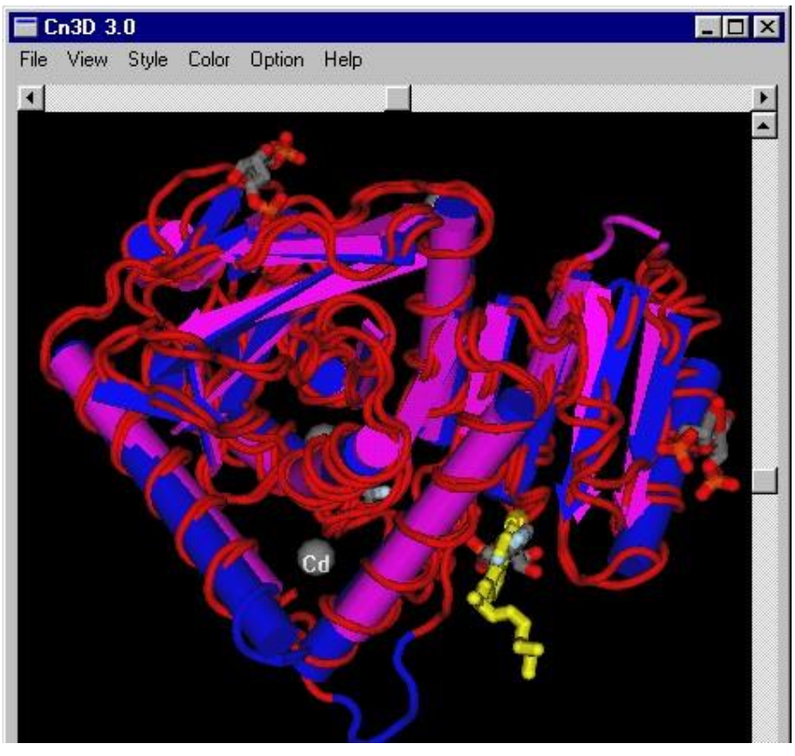


Online Tools

- 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


Cn3D Home Page
http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
AccessPlus ISU ISU Directory ISU Class Schedule AccessPlus BCB BCB 444/544 Netscape Webma
NCBI Structure
PubMed Entrez BLAST OMIM Books TaxBrowser Entrez Structure
Search Entrez Structure for Go
Cn3D 4.1 Homepage
Cn3D Tutorial
Cn3D feature highlights
Cn3D FAQ
Frequently Asked Questions
Cn3D Install
Installation and Configuration
MMDB
NCBI's structure
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!
WD40 - Cn3D 4.1
File View Show/Hide Style Window CDD Help
CDD Descriptive Items
Name: WD40
WD40 domain, found in a number of e
a wide variety of functions including ac
in signal transduction, pre-mRNA proc
assembly; typically contains a GH dipe
its N-terminus and the WD dipeptide a


OneD-Viewer
File Align Options Help
Go to: 0
1CET_A 1 MAPKAKIVLV GSGMIGGVMA TLIVQKNLGD VVLFDIVKNM
1CET_A 41 PHGKALDTSH TNVMAYSNCK VSGSNTYDDL AGSDVVIVTA
1CET_A 81 GFTKAPGKSD KEWNRLDLLP LNNKIMIEIG GHIKKNCPNA
1CET_A 121 FIIVVTNPVD VMVQLLHQHS GVPKNKIIGL GGVLDTSRLK
Ready !


Cn3D 3.0
File View Style Color Option Help
Cd

Protein structures are visualized using several online tools. These tools include Rasmol, CHIME, Swiss PDB Viewer and Cn3D.

182 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).

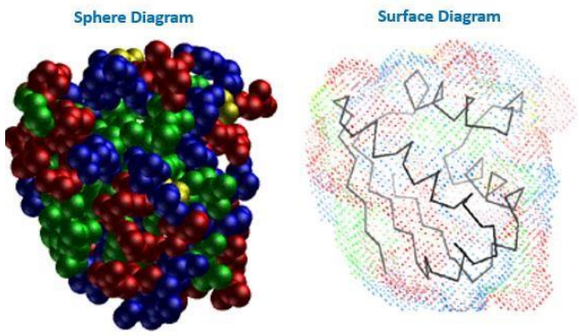


Figure 50 sphere and surface diagrams of protein

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.

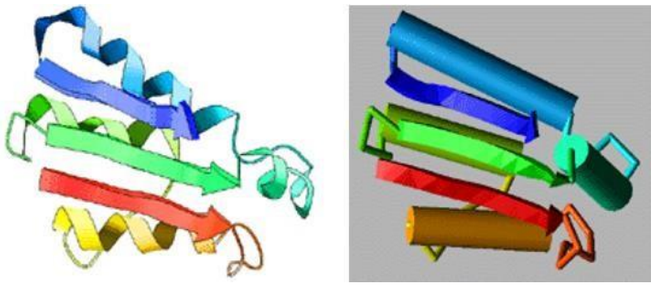


Figure 51 ribbon diagrams

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.

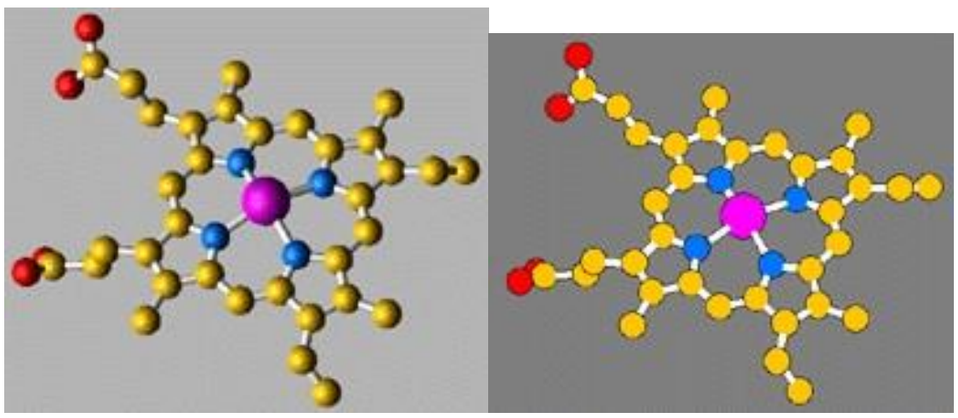


Figure 52 Balls and sticks model

http://www.danforthcenter.org/smith/MolView/Over/overview.html

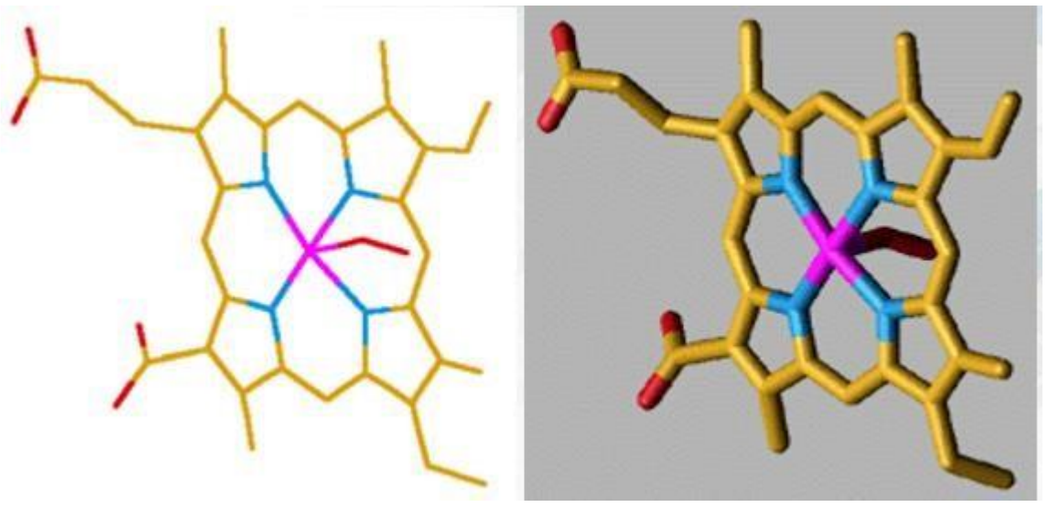


Figure 53 Colored Sticks Models

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.

183 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

184 Calculating Energies of Protein Structures

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.

$$\begin{aligned} Energy_{TOTAL} = {} & Atoms_{VWF} \times Energy_{VWF} + \\ & Atoms_{HB} \times Energy_{HB} + \\ & Atoms_{IonicInteraction} \times Energy_{IonicInteraction} \end{aligned}$$

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.

185 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 XRay or NMR.

186 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

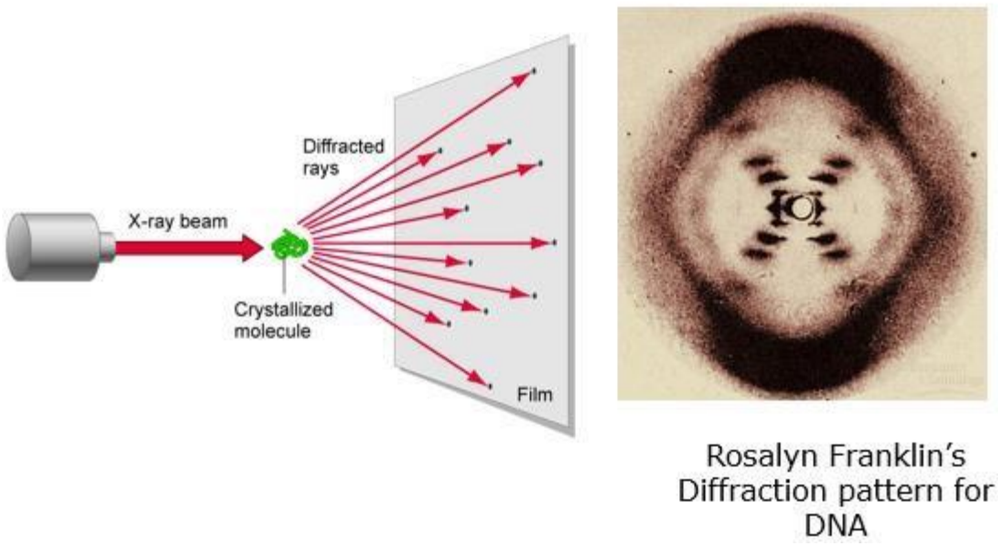


Figure 56 Principle of X-Ray Crystallography

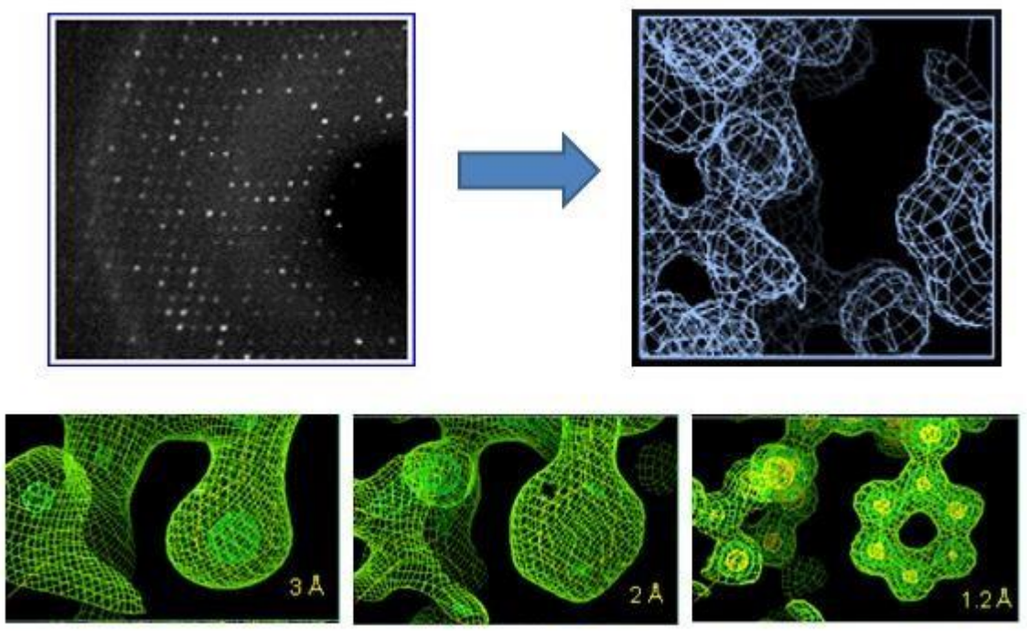


Figure 57 from Diffraction Patterns to Atomic Positions

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.

187 Protein Structures - 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.

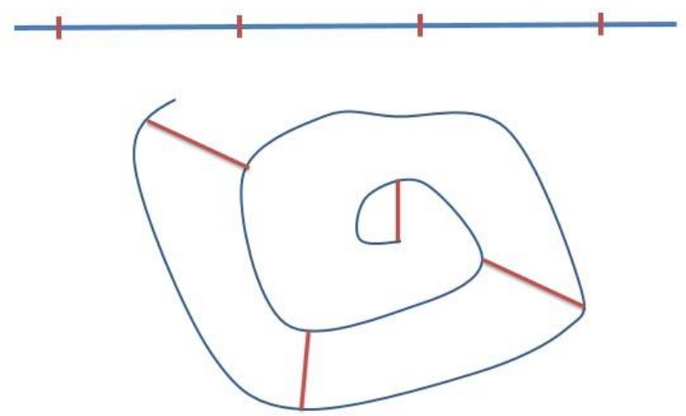
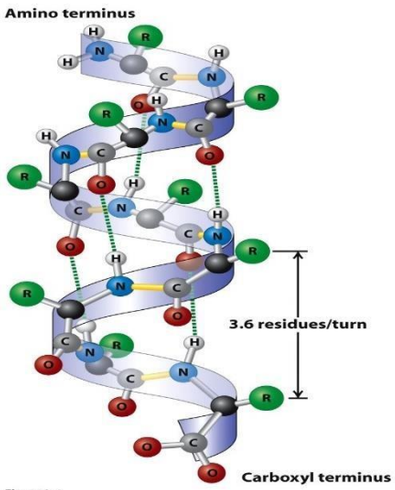


Figure 58 Hydrogen Bonds to Fold an Amino Acid Chain



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.

188 Protein Structures - 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.

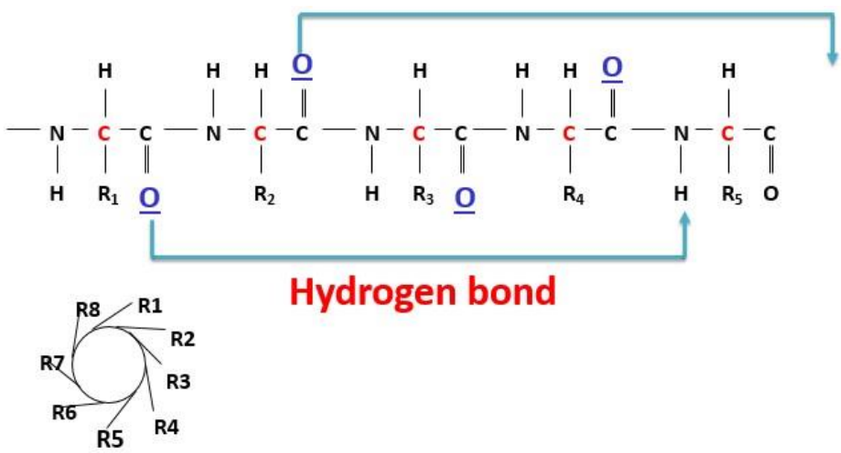
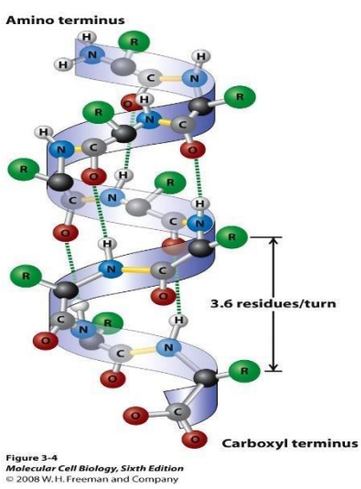


Figure 59 Forming Alpha Helix



Every Oxygen bound to 4th neighboring Amino Group' Hydrogen.

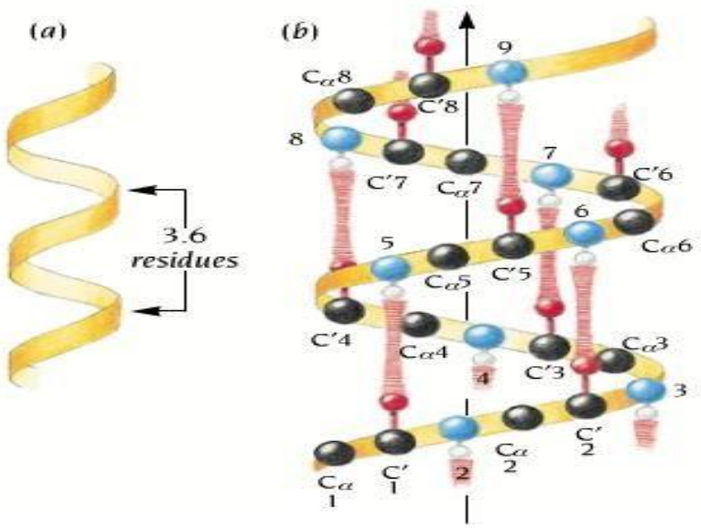


Figure 60 Carbons (Black) & Nitrogen's (Blue): 1-5, 2-6, 3-7…

Amino Acid	Preference			Properties
	Helix	Strand	Turn	
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

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

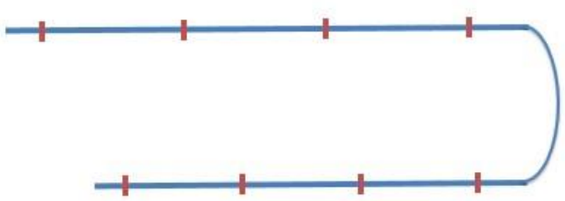
Helix Formers

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.

189 Protein Structures - Beta Sheets I

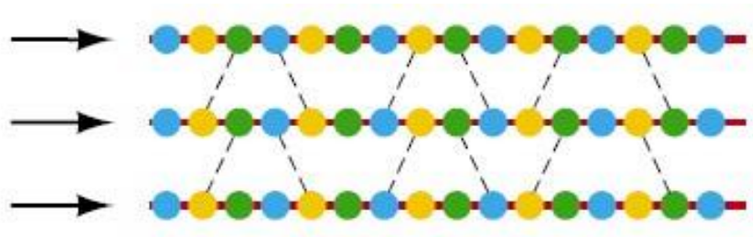
Alpha Helices are formed by hydrogen bonding (O-H) between C_i and N_{i+4} atoms in the protein backbone. Beta Sheets are another common secondary structure. They are constituted by several Beta Strands which come together. 5 to 10 resides are needed to make a Beta Strand, typically.

Hydrogen Bonds to make in Beta Strands

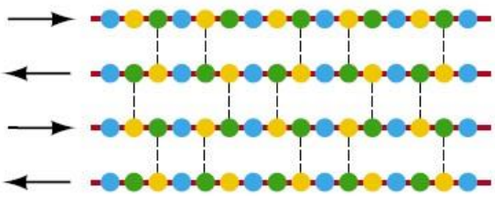


The Beta Sheet is made up of several Beta Strands

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.

190 Protein Structures - 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**

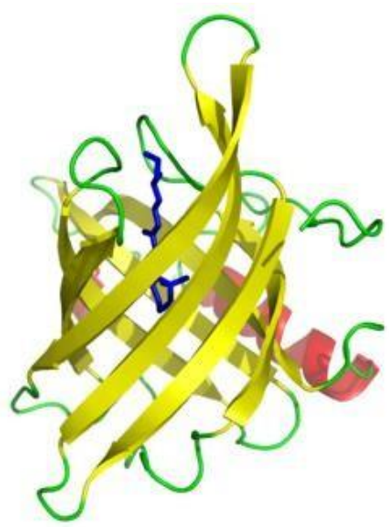


Figure 62 beta barrel

Beta Sandwiches

Beta Sandwiches are made of two beta sheets which are usually twisted and packed so their strands are aligned.

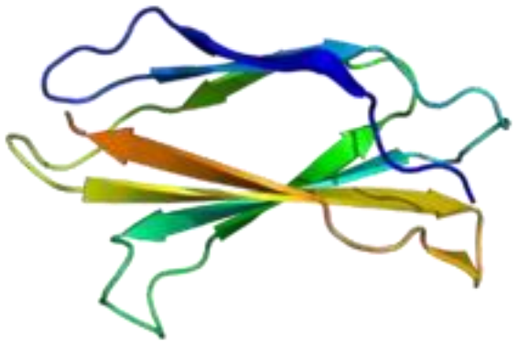


Figure 63 Illustration of the β-sandwich from Tenascin C (PDB entry: 1TEN).

Amino Acid	Preference: Helix	Preference: Strand	Preference: 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	

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

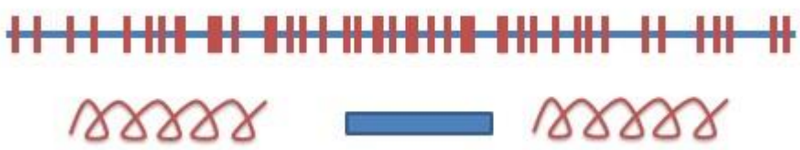
Figure 64 Preference of Amino Acids for making Beta strands

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.

191 Protein Structures - 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.

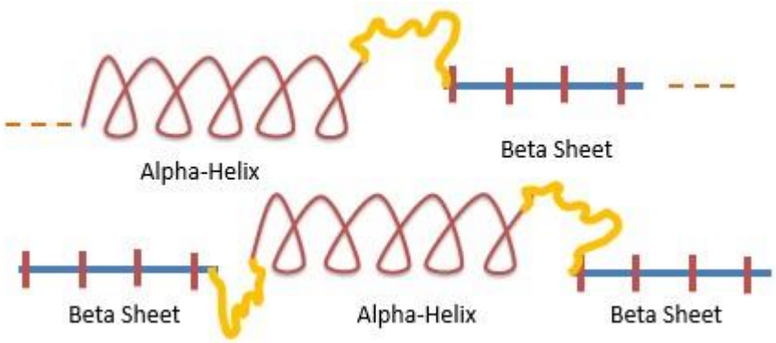
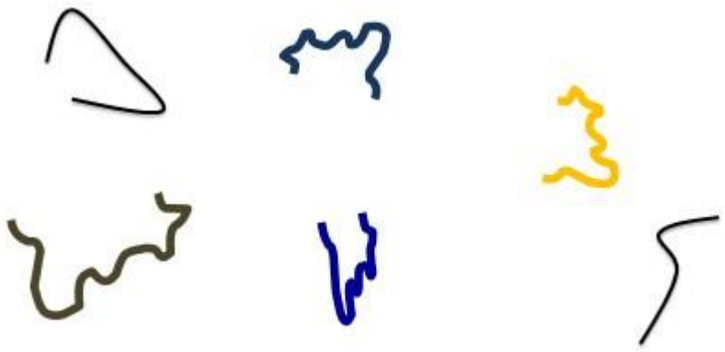


Figure 65 Joining Alpha Helices and Beta Sheets in a Protein Backbone

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

192 Protein Structures - 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.

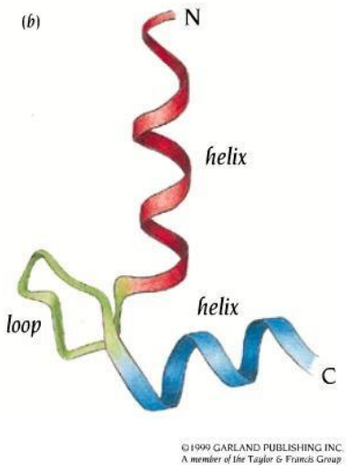


Figure 66 Loops in 3D Conformation

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.

193 Protein Structures – 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

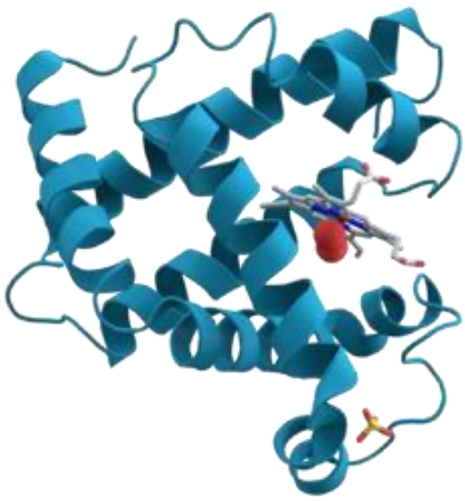


Figure 68 Coils in Myoglobin

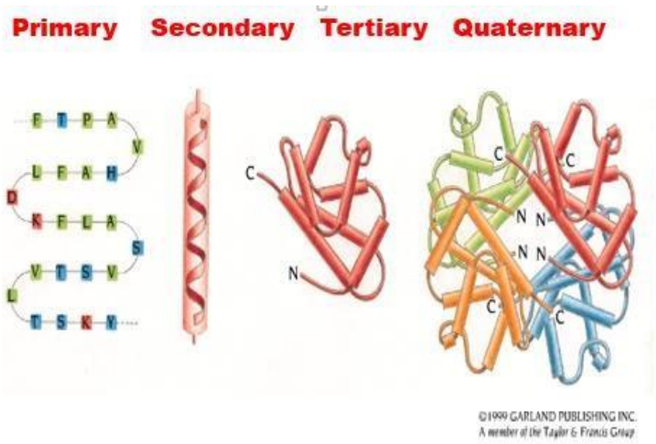
Coils are those secondary structure formed by the protein backbone which are neither helices, sheets nor loops. In fact, coils do not have a consistent classifiable structure. Hence, coils are random structure and random length.

194 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



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)

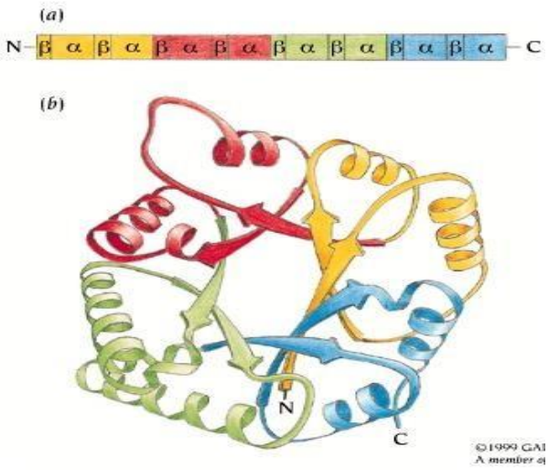


Figure 69 Domain (Functionally Complete)

Domains are semi-independent functional structures in a protein. Have a stable structure. Over ~40 residues. Protein may contain multiple domains.

195 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.

196 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)

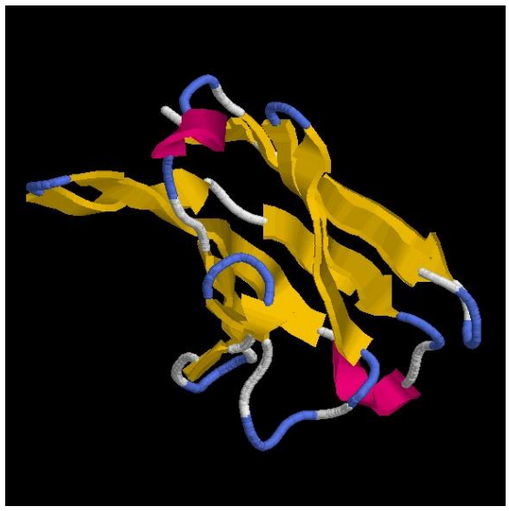


Figure 71 Immunoglobulin (8fab)

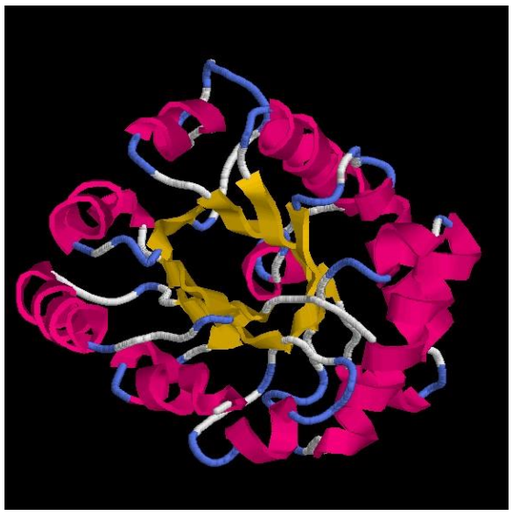


Figure 72 Alpha / Beta: Triosephosphate isomerase (1hti)



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.

197 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.

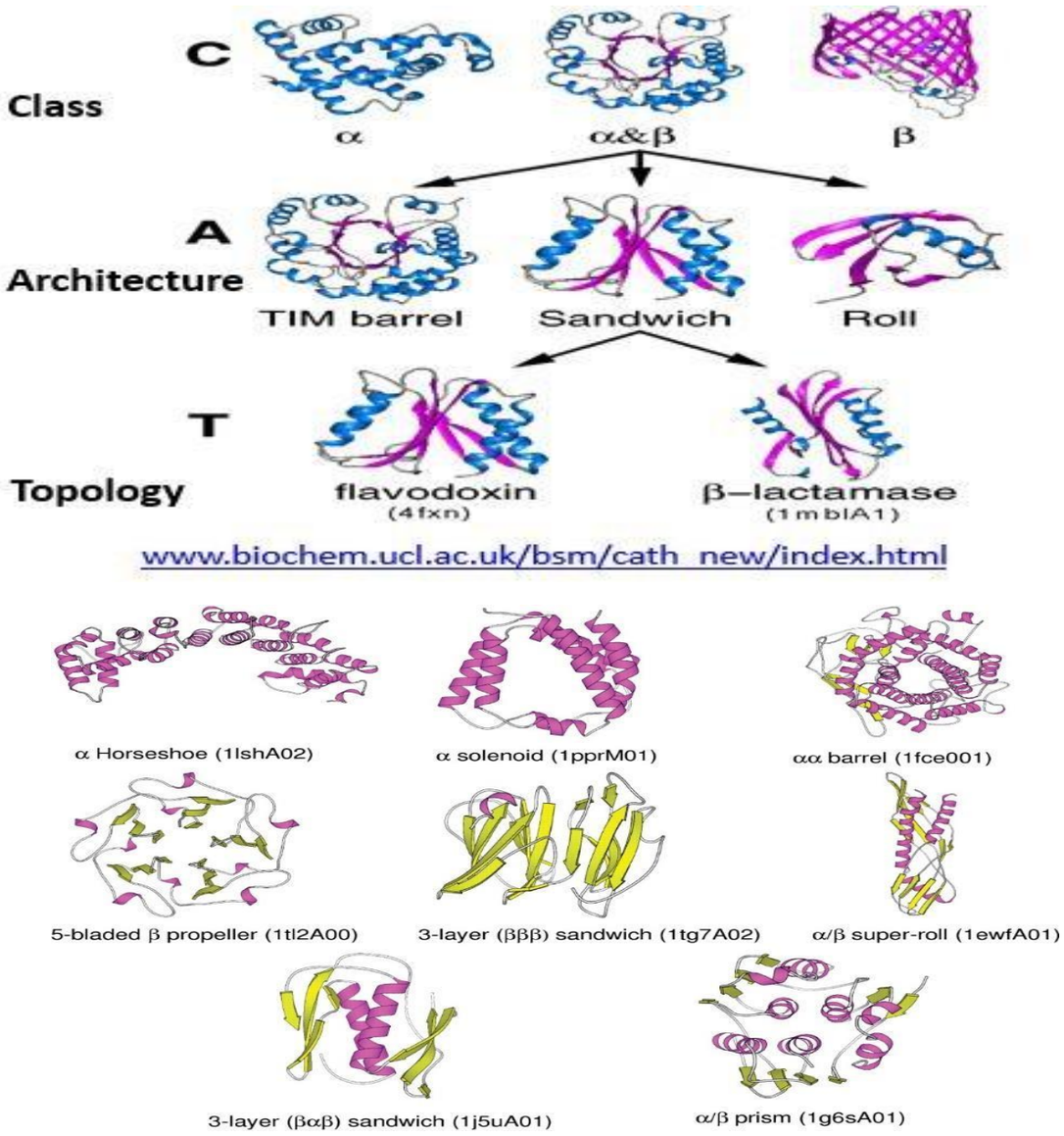


Figure 74 Structural Classes

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

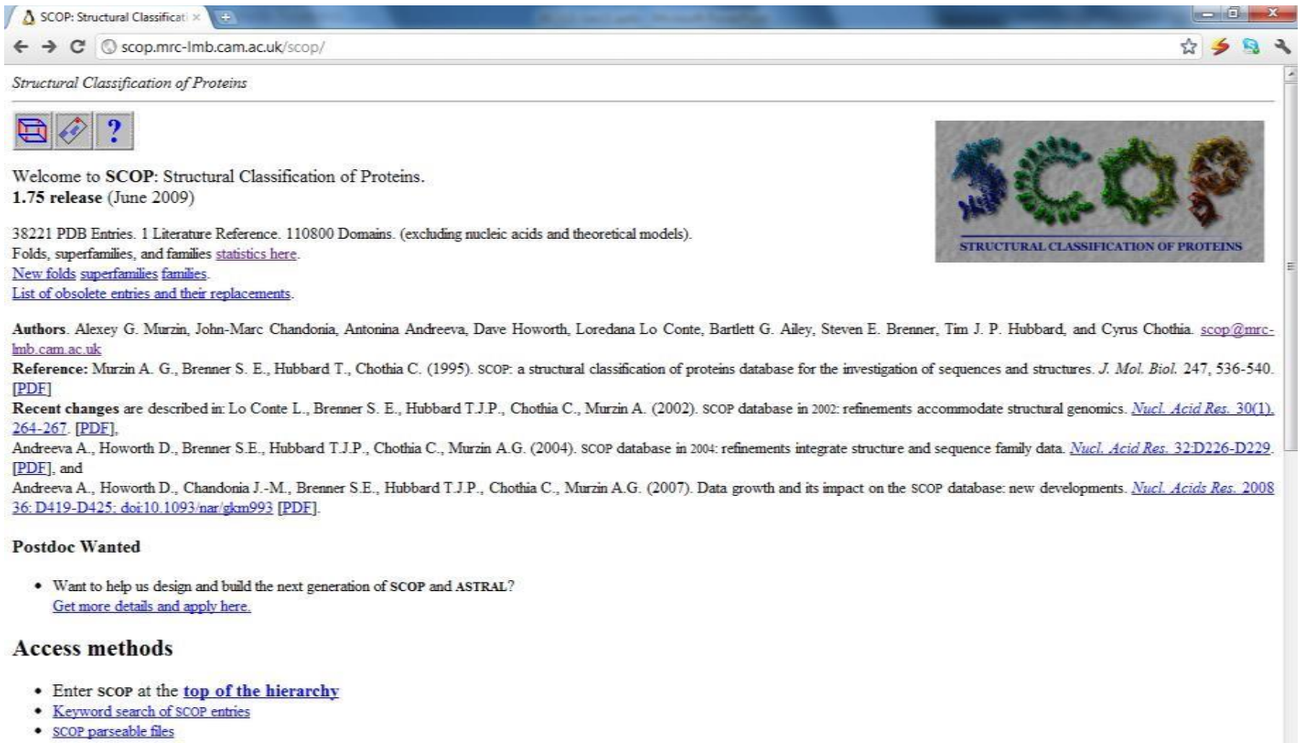
Homology

- Same Family
- Clear evolutionary relationship
- Pairwise sequence similarity > 30%

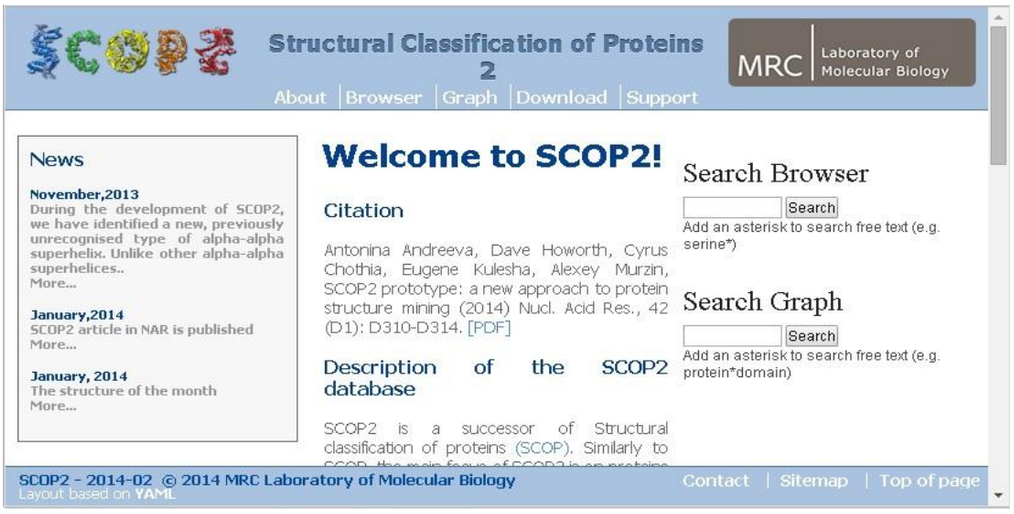
CATH classifies proteins by their structural similarity. It also considers the internal organization of the structural components in proteins.

198 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.



http://scop.mrc-lmb.cam.ac.uk/scop/



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.

199 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.

200 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.

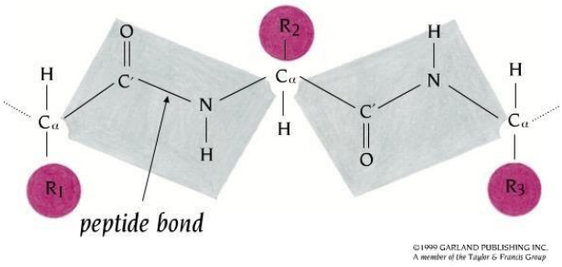


Figure 74 C-Alpha atoms in backbone

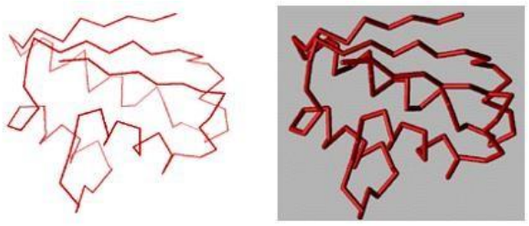


Figure 76 Tracing and Visualizing C-Alpha Backbone

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.

201 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.

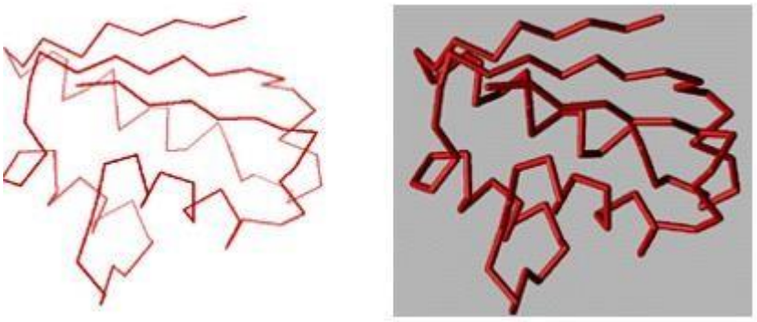


Figure 77 Tracing and Visualizing C-Alpha Backbone

http://www.danforthcenter.org/smith/MolView/Over/overview.html

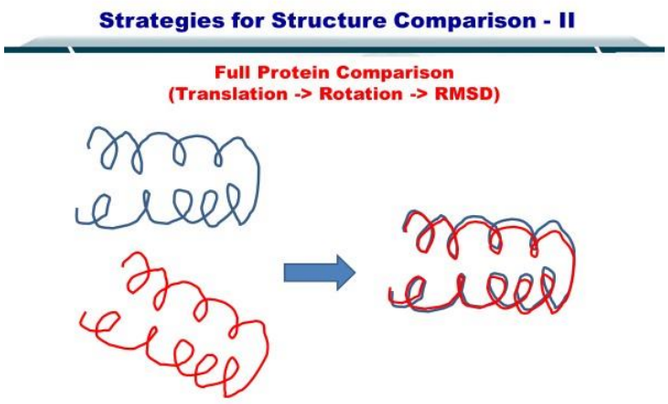
Strategy # 1 – Whole Protein Structure Comparison by Intermolecular distances

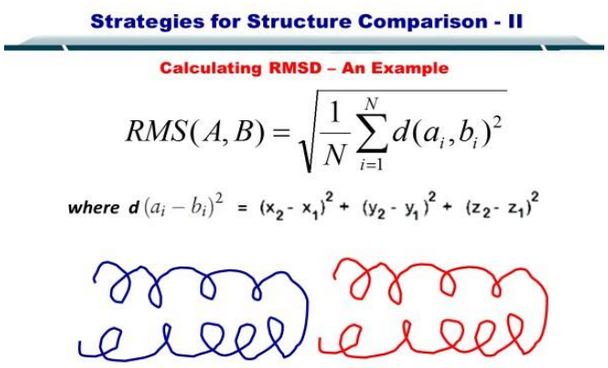
- Two protein sequences are pair-wise aligned with each other
- Corresponding Alpha Carbons are identified
- Coordinates of corresponding Alpha Carbons are retrieved from PDB
- Their individual differences calculated
- Root Mean Square Distance is computed to assess the similarity

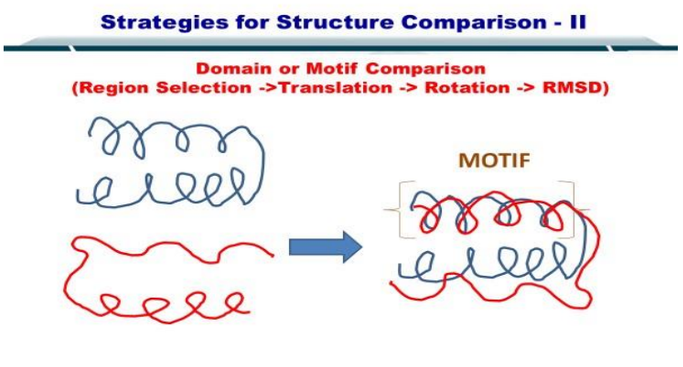
Whole protein structures can be compared by calculating the root mean squared difference (RMSD) between their Alpha Carbons positions. The lower the RMSD, the similar are the proteins.

202 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.



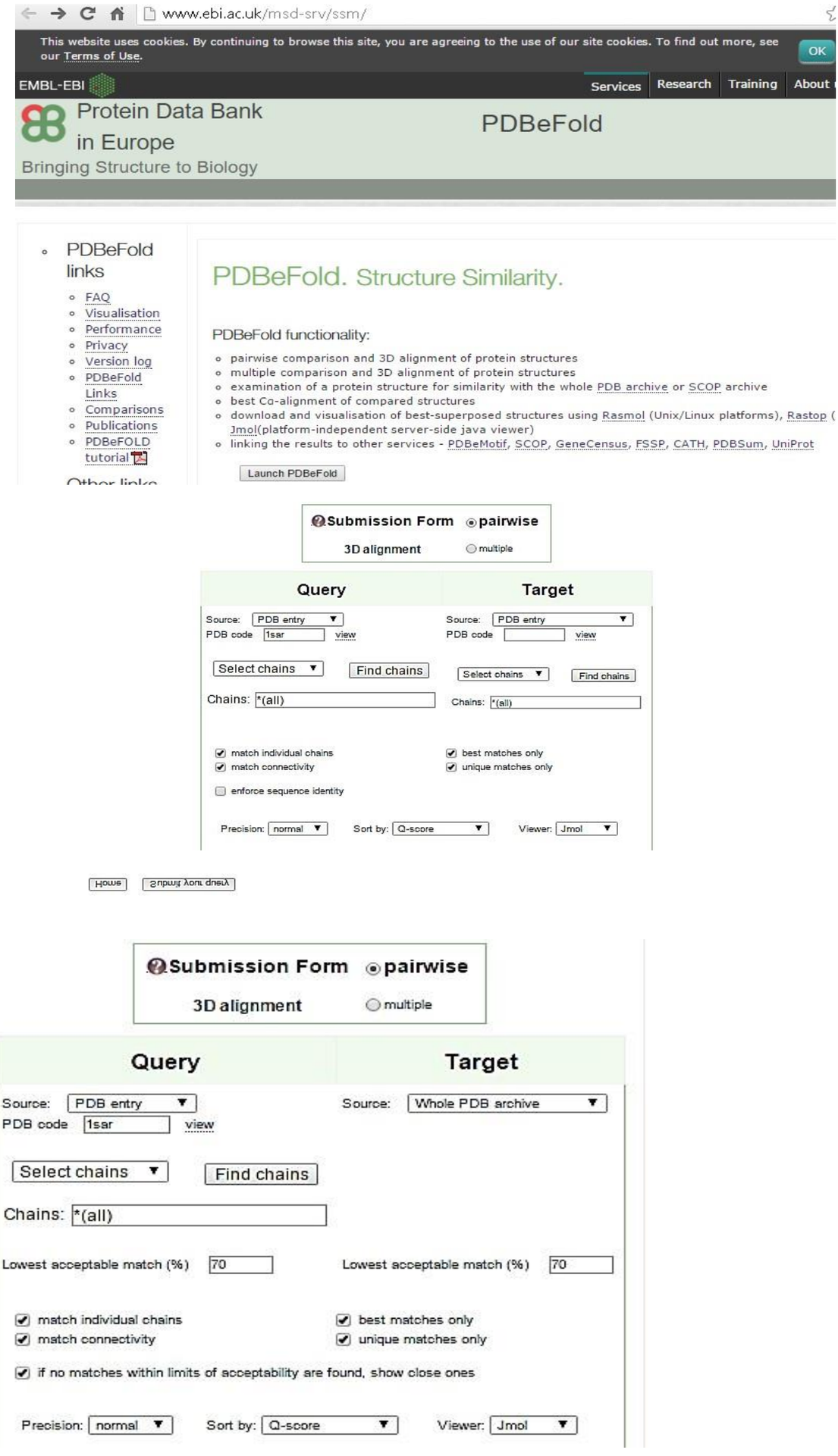




Motifs, Domains and Full Proteins can be compared by using the rigid body super-positioning. Depending on the RMSD, proteins, their motifs and domains can be selectively compared.

203 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.



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.

204 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

205 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.

206 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)

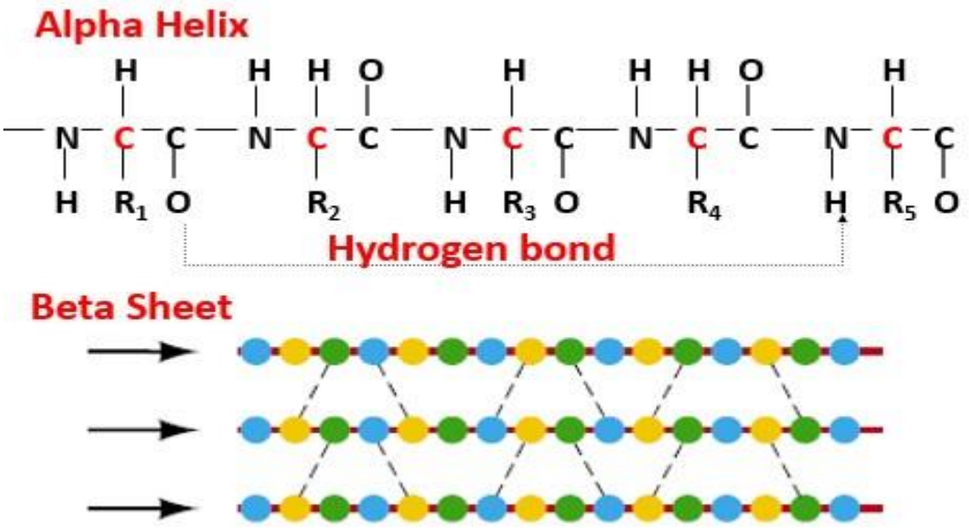
The propensity for forming "ααα β β β T T" is = 29.47

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.

207 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.

208 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.

209 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.

210 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.

211 Chou Fasman Algorithm – IV

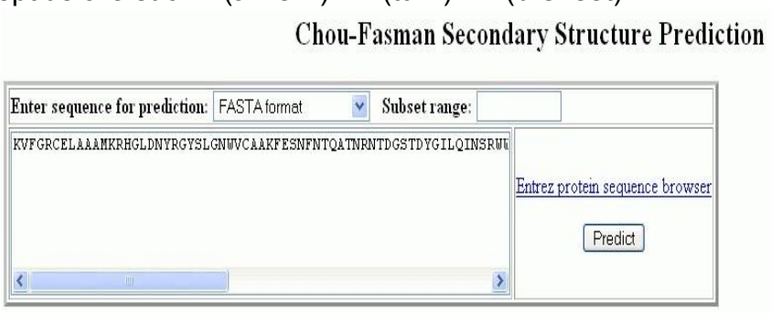
After computing the propensity of alpha helices and beta sheets, we need to settle for loops. Let's see how we can find out the loops using Chou Fasman Algorithm.

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)

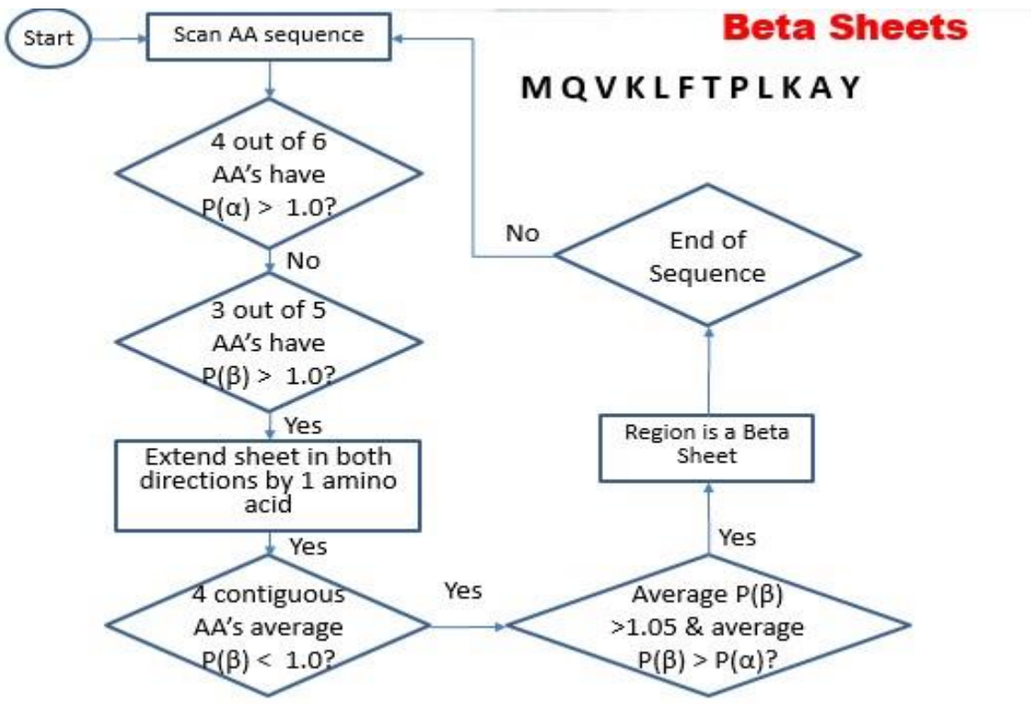


http://fasta.bioch.virginia.edu/fasta_www/chofas.htm

Chou Fasman Algorithm helps predict Alpha Helices, Beta Sheets and Turns. The algorithm is based on statistical occurrence of Amino Acids in known structures.

212 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 can be predicted from primary amino acid sequences. Next, we will see the flowchart of Alpha Helices and Beta Turns.

213 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.

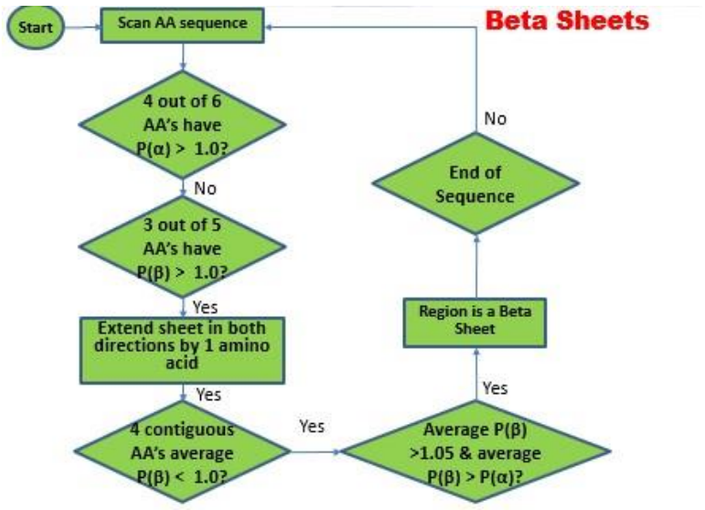


Figure 80 Beta sheet flowchart

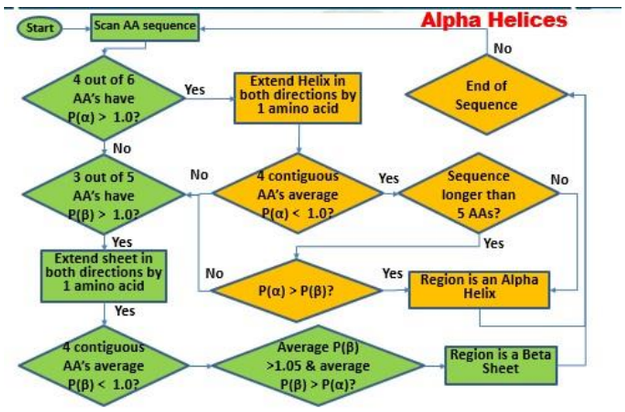


Figure 81 Alpha helices flowchart

Now we have reviewed flowcharts for Alpha Helices and Beta Sheets. Next up is the flow chart for Beta Turns.

214 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.

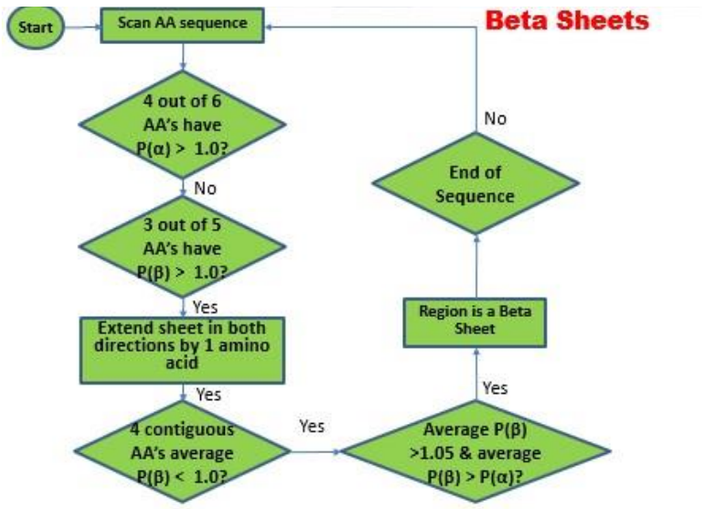


Figure 82 Beta sheet flowchart

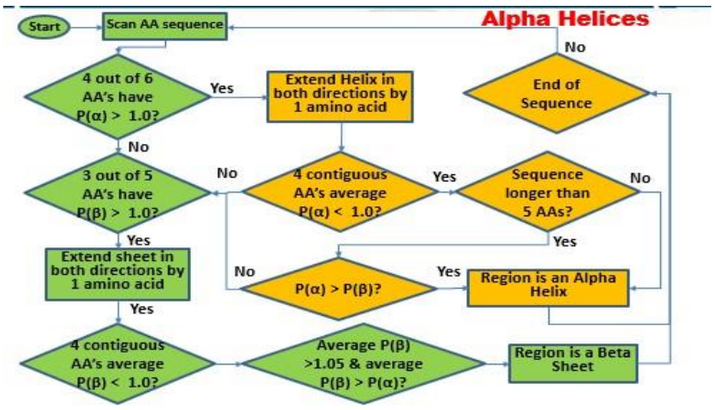


Figure 83 Alpha helices flowchart

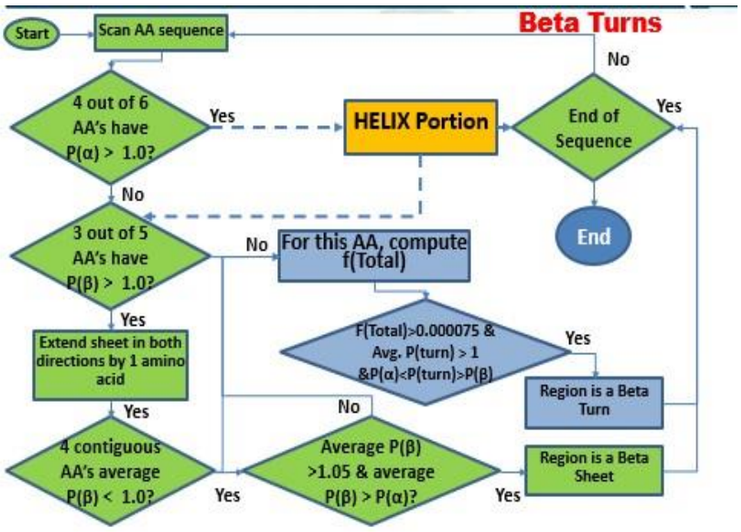


Figure 84 Beta turn

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.

215 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!

216 Summary of Visualization, Classification and 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.

217 Introduction to Homology modelling

Background

- Proteins are 3D molecules with their own unique structures
- Protein structure is reflective of the protein function ➢ Protein structure includes 1', 2', 3' and 4' structures
- 1' structure of proteins is the sequence of proteins and can be obtained by mass spectrometry
- 2' structures formed by proteins are the helices, beta sheets, loops and coils ➢ 3' structure of proteins is the combination of 2' structures such that the overall protein structure is formed
- 4' protein structure is formed when two or more proteins complex together
- X-Ray Crystallography and NMR Spectroscopy are used to find the structures of proteins
- However, these methods are difficult and expensive ➢ Solution: Prediction of structures

Introduction

- Protein sequence gives rise to its structure
- If another protein which has a similar sequence also has its structure known, the structure of an unknown protein can be predicted based on that similar protein
- So, it is then possible to identify unknown protein structures by just examining the homologous protein sequences Conclusions
  - Sequence Identity
  - Alignment Length

Which combination of identity and alignment length is suitable for best for structure prediction?

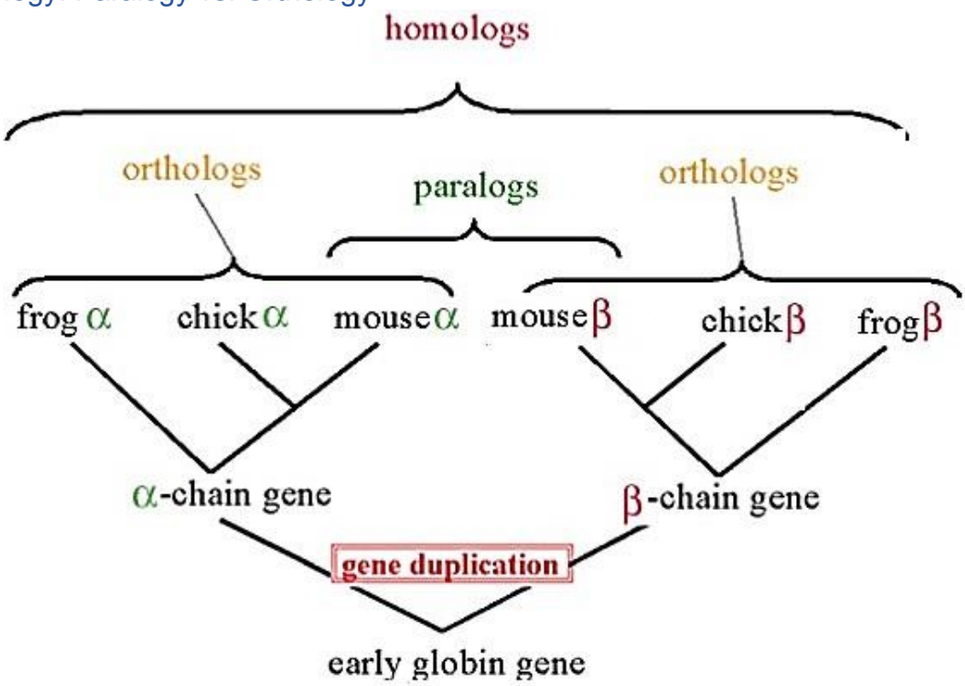
218 Homology, Paralogy and Orthology

Background

In homology modelling, proteins with similar 1' sequences are considered

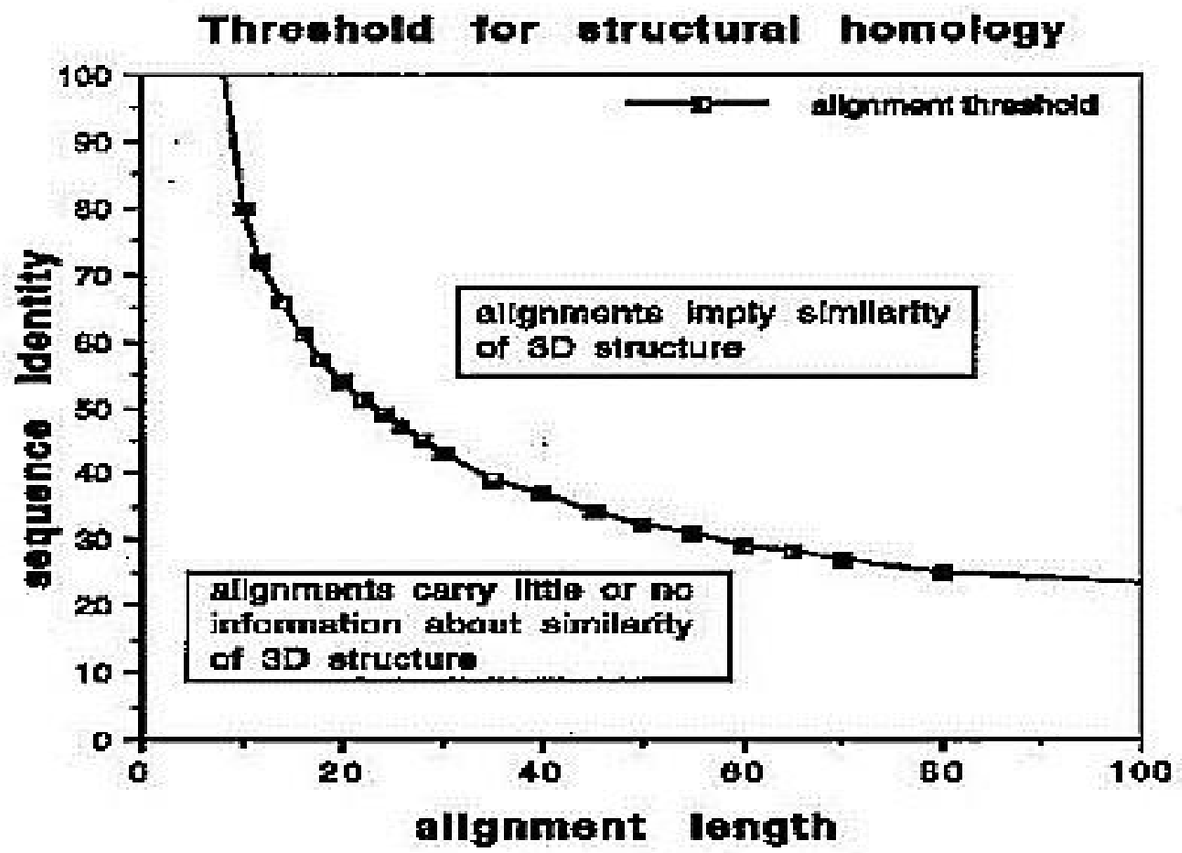
Given that one of them has its 3' structure known, then the 3' structure of other protein can be predicted

Homology: Paralogy vs. Orthology



http://bioweb.uwlax.edu/GenWeb/Molecular/Bioinformatics/Unit_4/Lab_4-2/lab_4-2.htm

How much homology is required or better?



Conclusions

Good sequence alignment and identity ensures that homology modelling will give accurate results

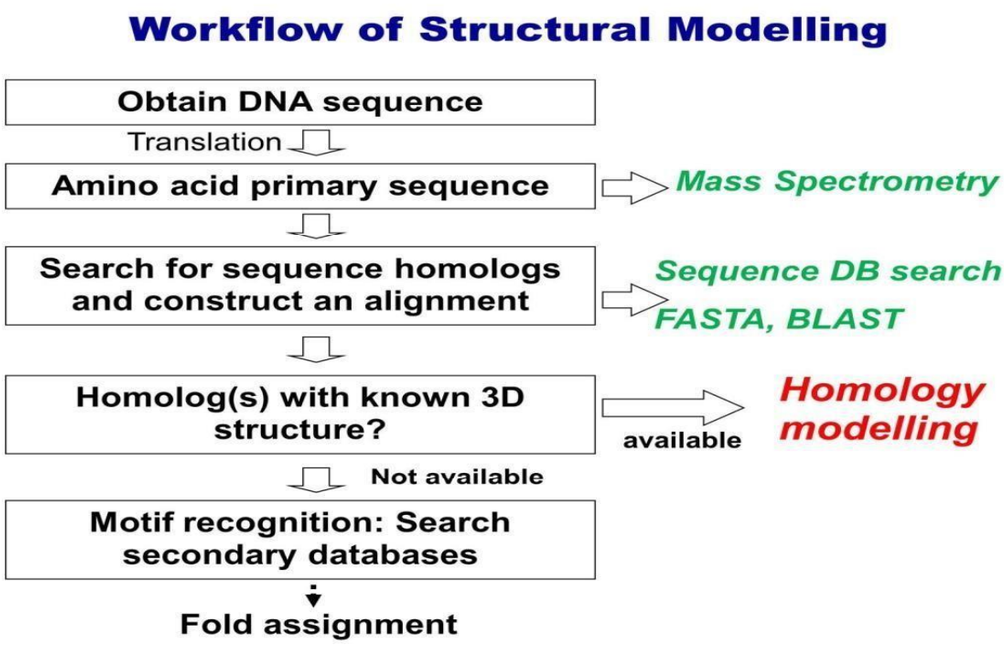
Next, what is the workflow for homology modelling?

219 Workflow for Structural Modelling

Background

Homology modelling is used to predict structures of proteins having high sequence similarity with other proteins with known structures!

Let's consider the workflow of homology modelling



Introduction

Overall, there are three different strategies for structure prediction

1. Homology Modelling
2. Threading/Fold Recognition
3. Ab Initio Modelling

Conclusions

Next, we will proceed to perform homology modelling
For that there is a seven step procedure which we will see in the next module.

220 Seven Steps to Homology Modelling – I

Background
Protein structure can be predicted by 3 methods:

1. Homology Modelling
2. Fold Recognition / Threading
3. Ab Initio Modelling

Introduction

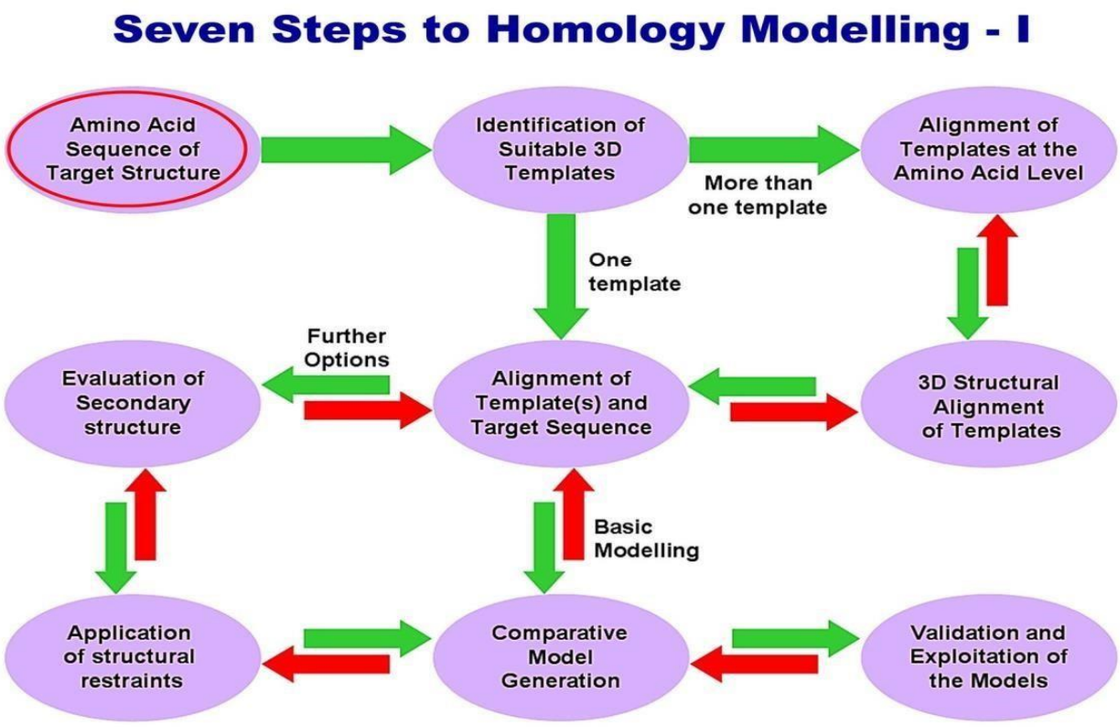
Let's start by looking at Homology Modelling
There are seven salient steps in any Homology Modelling pipeline
Definition of Template (known) & Target (unknown) Homology modeling of the

target structure can be done as follows:

1. Template recognition and initial alignment
2. Alignment correction

Backbone generation

3. Loop modeling
4. Side-chain modeling
5. Model optimization
6. Model validation



Conclusions

Homology modelling works in seven steps
It is a repetitive process
Next, we will look at each step in detail!

221 Seven Steps to Homology Modelling – II

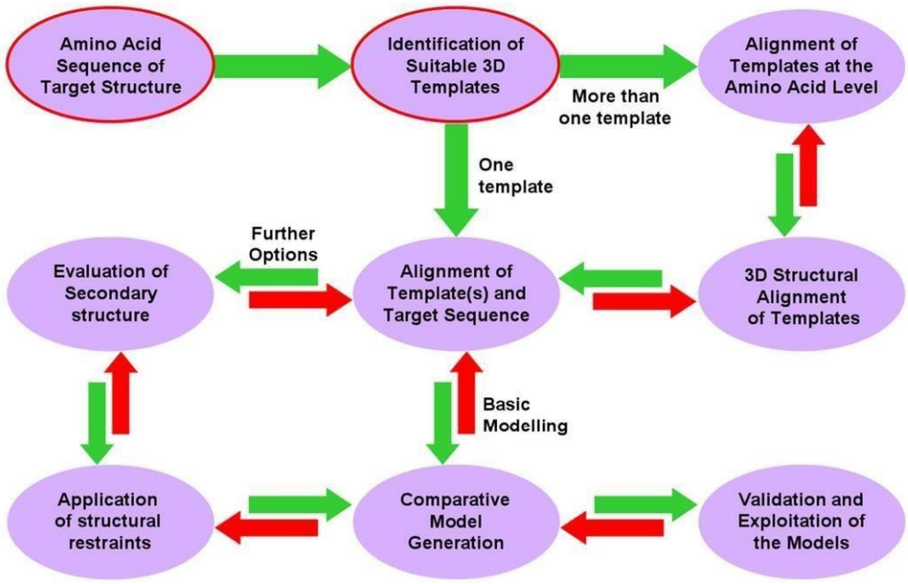
Background

Template recognition and initial alignment
Alignment correction

Backbone generation

Loop modeling
Side-chain modeling
Model optimization
Model validation

Seven Steps to Homology Modelling - II

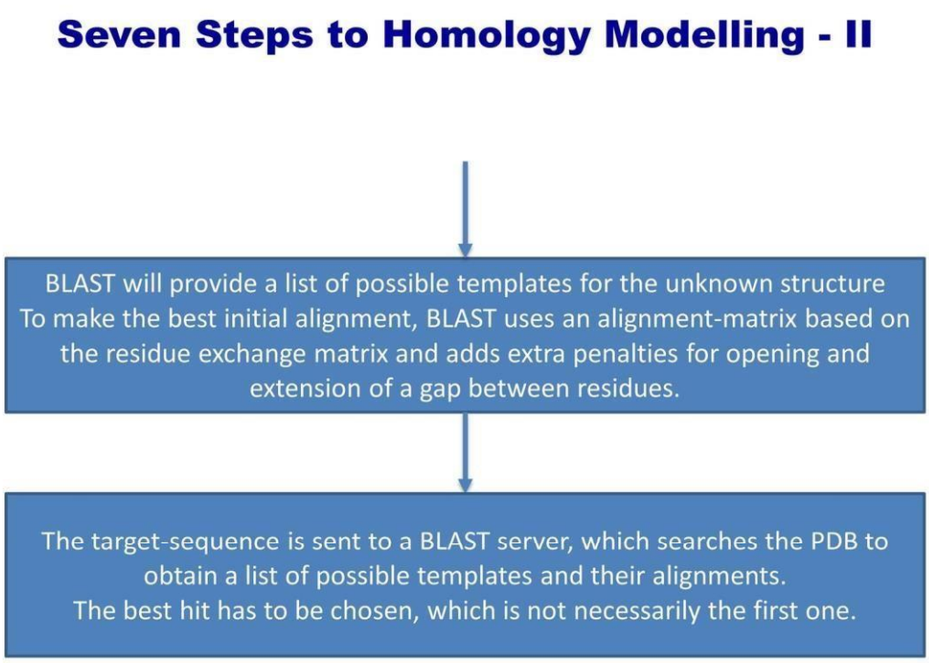


Seven Steps to Homology Modelling - II

Compare the sequence of the unknown protein with all the sequences of known structures stored in the Protein Data Bank (PDB).

BLAST this sequence against PDB sequences – Obtain a list of known protein structures that match the sequence.

BLAST uses a residue exchange scoring matrix.
Residues that are easily exchanged (e.g. Ile to Leu) get a better score than residues that have different properties (for example Glu to Trp).
Function specific conserved residues get the best score (e.g. Cys to Cys).



Conclusions

Now the template and target are selected
Next, we perform fine-tuning of alignment and introduce corrections to ready the mismatches and gaps

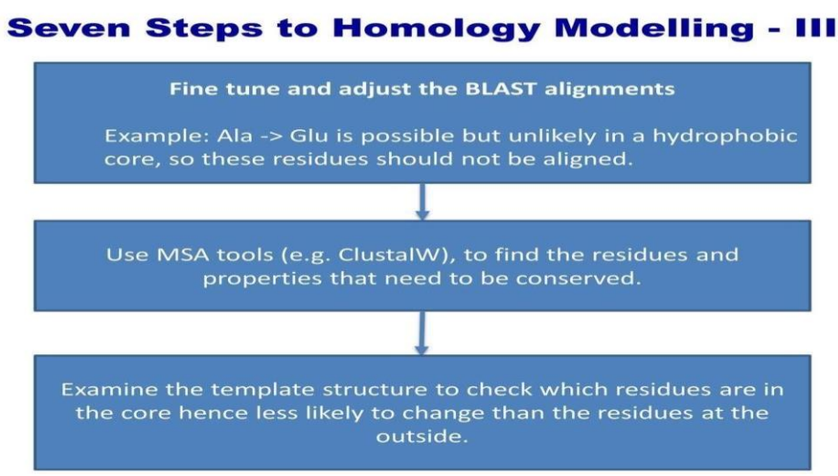
222 Seven Steps to Homology Modelling – III

Background

Template recognition and initial alignment
Alignment correction

Backbone generation

Loop modeling
Side-chain modeling
Model optimization
Model validation



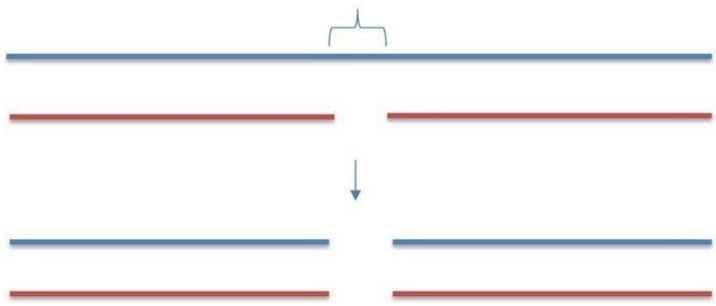
Seven Steps to Homology Modelling - III

Insertions and deletions can be made in those parts of the sequence which are highly variable .

Note: MSA can be helpful to find these places.

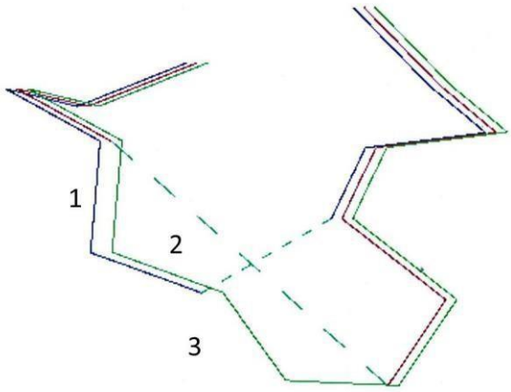
Gaps have to be shifted around until they are as small as possible.

Seven Steps to Homology Modelling - III



Note: After a deletion of 3 residues a big gap occurs in the RED structure, which was the best alignment.

Seven Steps to Homology Modelling - III



Remember that deletion of 3 residues left a big gap in the RED structure, which was the best alignment.

After shifting several residues, the gap is much smaller (blue structure) and more likely to be correct.

Conclusions

The alignment now stands fine tuned and corrected
Gaps and mismatches have been evaluated and adjusted

Next step, using this alignment, assemble the backbone

223 Seven Steps to Homology Modelling – IV

Background

Template recognition and initial alignment
Alignment correction

Backbone generation

Loop modeling
Side-chain modeling
Model optimization
Model validation

Seven Steps to Homology Modelling - IV

The coordinates of the template-backbone are copied to the target.

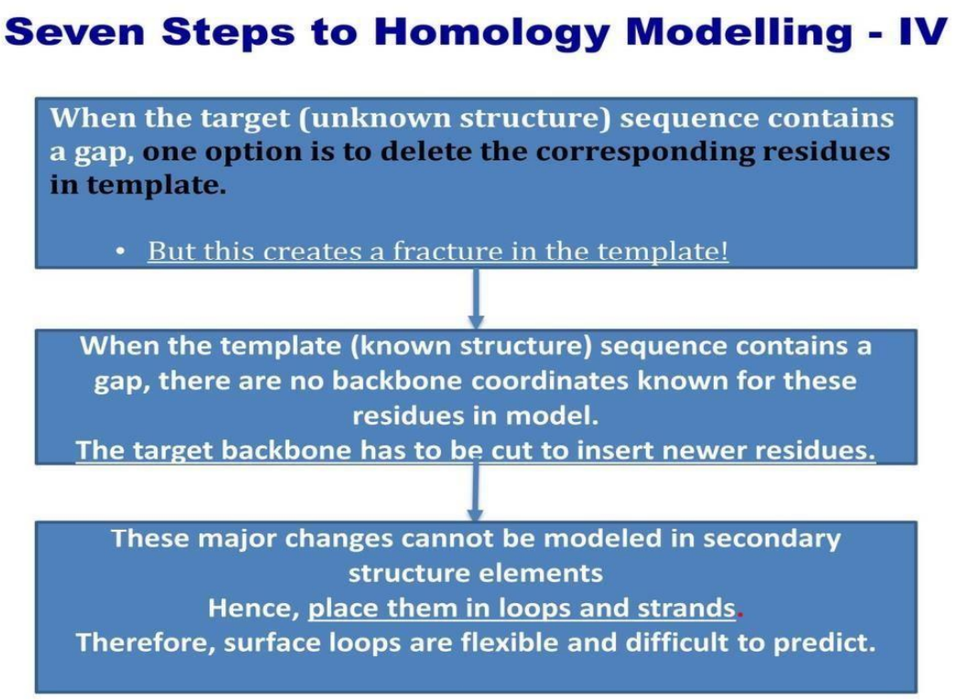
When the residues are identical, the side-chain coordinates are also copied.

Note that PDB-file may contain small offsets / errors, so try to use multiple similar templates.

Template recognition and initial alignment ✓ Alignment correction

Backbone generation

Loop modeling
Side-chain modeling
Model optimization
Model validation



Conclusions

The protein backbone is ready!
Next, loops were modelled and used to bridge gaps
Next step, using this backbone and loop choices, place the side-chains

224 Seven Steps to Homology Modelling – V

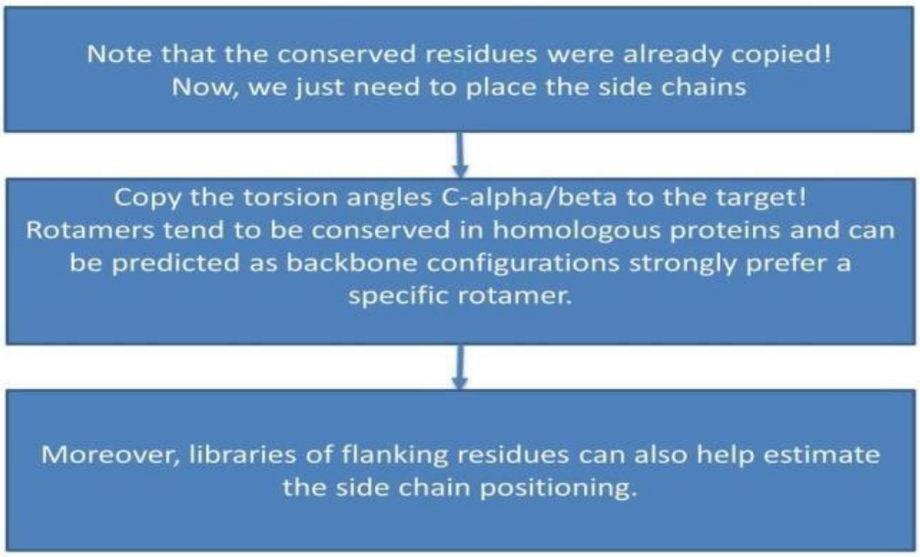
Background

Template recognition and initial alignment
Alignment correction

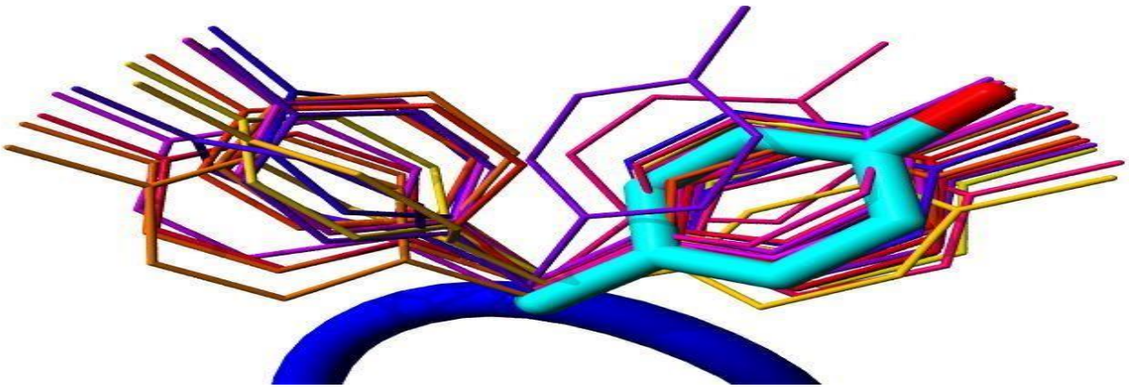
Backbone generation

Loop modeling
Side-chain modeling
Model optimization
Model validation

Seven Steps to Homology Modelling - V



The backbone of tyrosine strongly prefers two rotamers and the real side-chain may fit one of them!



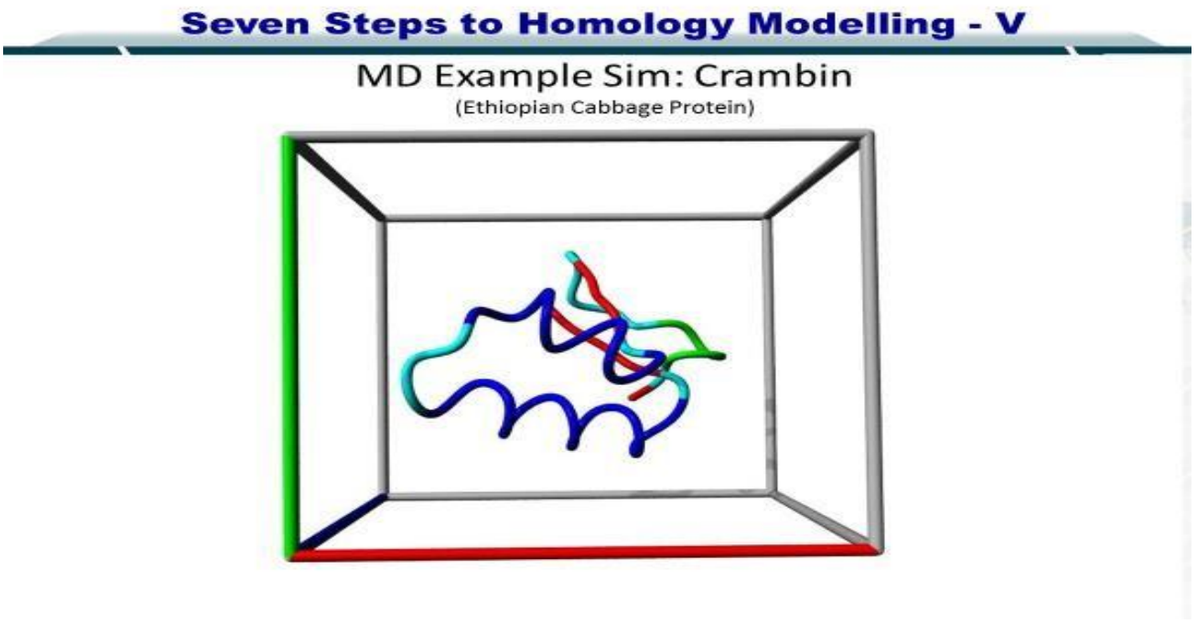
Next….

Template recognition and initial alignment
Alignment correction

Backbone generation

Loop modeling
Side-chain modeling
Model optimization ✓ Model validation

The backbone of tyrosine strongly prefers two rotamers and the real side-chain may fit one of them!



Conclusions

Now we have minimized large errors
However, smaller errors may still exist
Next step, validate the model that we have constructed!

225 Seven Steps to Homology Modelling – VI

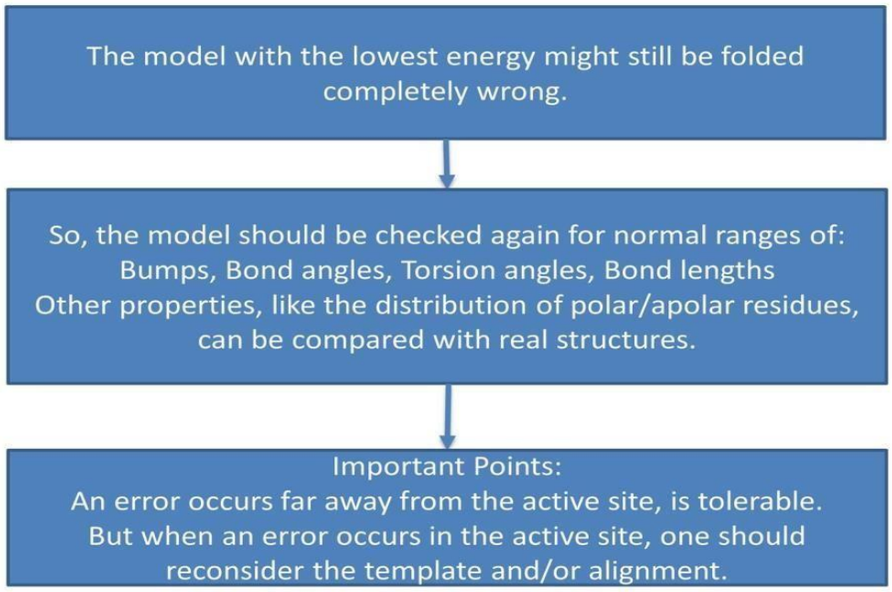
Background

Template recognition and initial alignment
Alignment correction

Backbone generation

Loop modeling
Side-chain modeling
Model optimization
Model validation

Seven Steps to Homology Modelling - VI



Limitations of Homology Modelling
Large Bias towards structure of template
Cannot study conformational changes
Cannot elicit new catalytic/binding sites

Conclusions

So how can we overcome such limitations?
Other strategies include: Threading, and Ab Initio Modelling
We will also examine online tools for each

226 Modeller for Homology Modelling

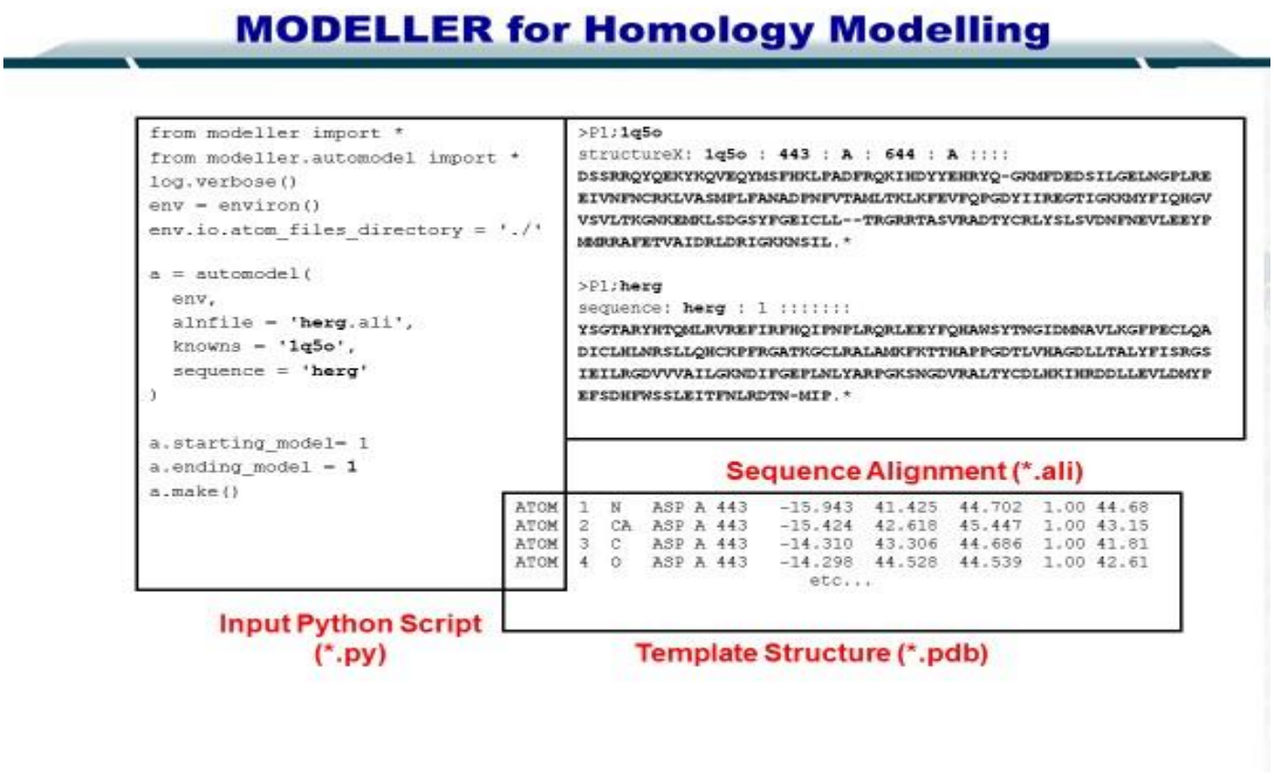
Background

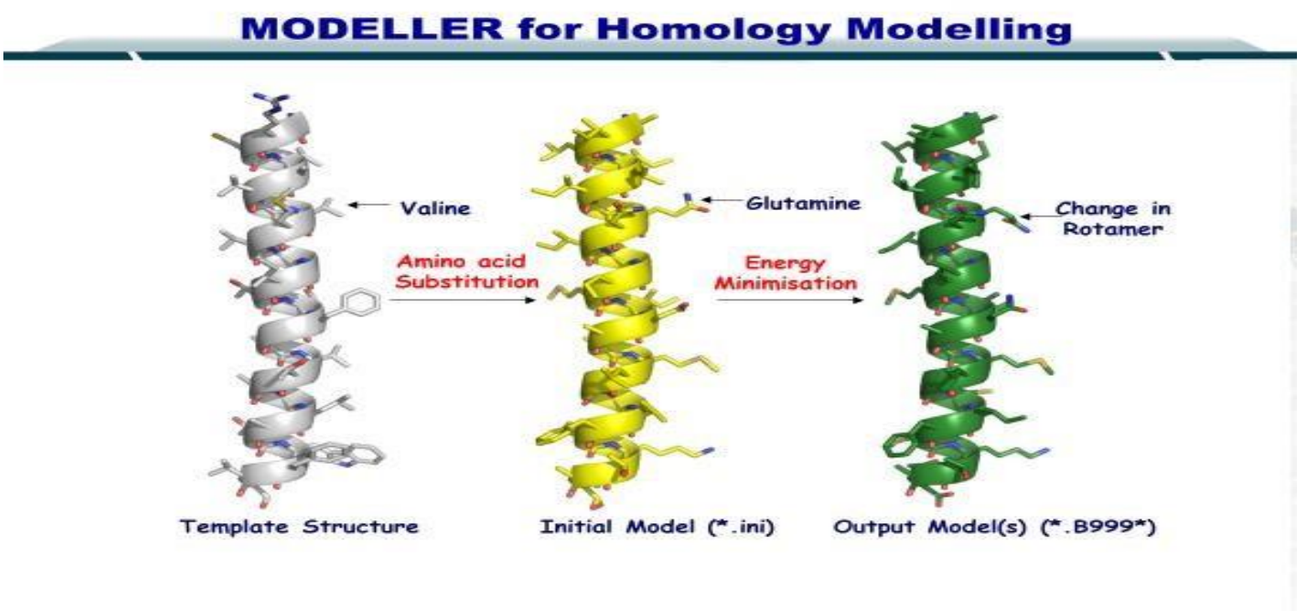
Template recognition and initial alignment
Alignment correction

Backbone generation
Loop modeling
Side-chain modeling
Model optimization
Model validation

Background

Modeller is a software for homology modelling
salilab.org/modeller
Inputs: Python script file, Sequence alignment & Template (PDB)





.log : log output from the run.
.B* : model generated in the PDB format.
.D* : progress of optimisation.
.V* : violation profile.
.ini : initial model that is generated.
.rsr : restraints in user format.
.sch : schedule file for the optimisation process.

Automated Modelling Servers Swiss Model

http://swissmodel.expasy.org//SWISS-MODEL.html

Robetta http://robetta.bakerlab.org/

3D Jigsaw

http://www.bmm.icnet.uk/servers/3djigsaw/

Phyre

http://www.sbg.bio.ic.ac.uk/phyre/

Conclusions

Homology modelling helps predict protein structures by using prior structural information
Several tools are available to perform homology modelling in a programmatic or automated way!

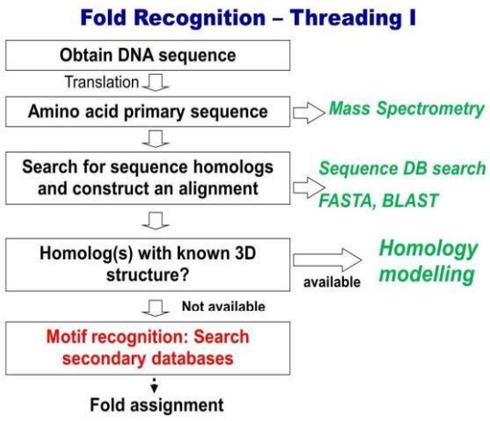
227 Online Tools for Fold Recognition

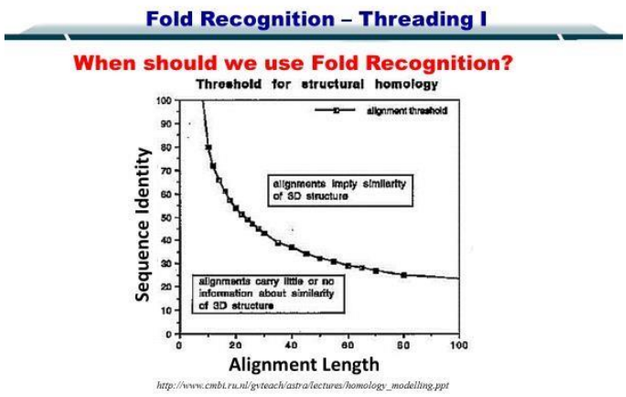
Background

Template recognition and initial alignment
Alignment correction

Backbone generation

Loop modeling
Side-chain modeling
Model optimization
Model validation





Introduction

A protein fold is defined by the way the secondary structure elements of the structure are arranged relative to each other in space.
Common folds include 4-helix bundle and the TIM barrel. Introduction
5,000 stable folds in nature
Fold recognition: Finding the best fit of a sequence to a set of candidate folds

Conclusions

Fold recognition or Threading is a technique for predicting protein structures
It is useful in cases where homology modelling fails to predict quality structures

228 iTASSER

Background

Fold recognition is also called Threading

Technique for predicting protein structures
Employed when homology modelling cannot predict quality structures

Fold Recognition – Threading II

Find the best way to

"mount" the residue sequence of one protein

onto a known protein structure!

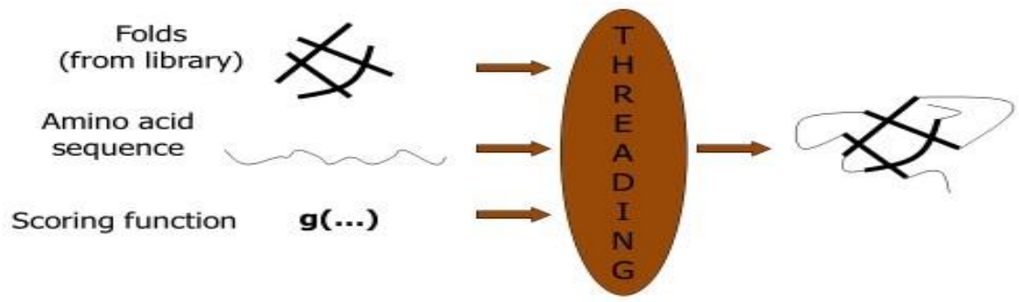


The process of threading

In the process of "Threading", we mount an amino acid sequence on to the backbone of template structures in a folds library
Each step is "drag" along the sequence (MQVKLFTY...) through each location of each template fold
Then, for each fold, we must compute the fitness of sequence matching that fold!

Fold Recognition – Threading II

Inputs and outputs of threading



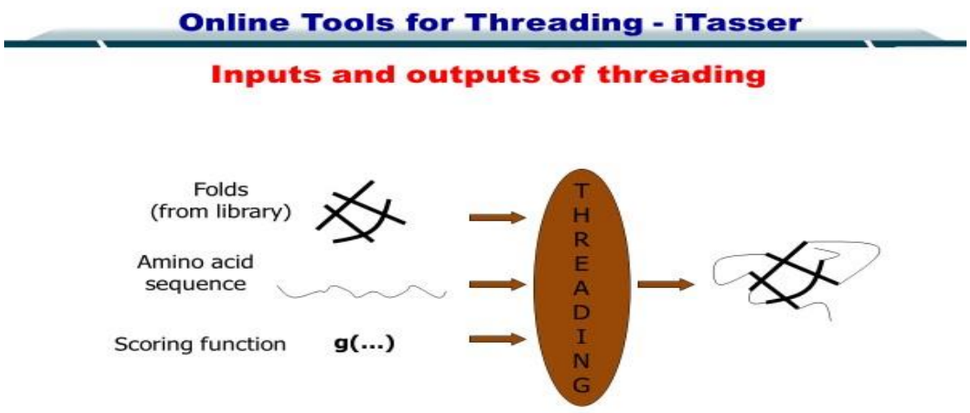
Conclusions

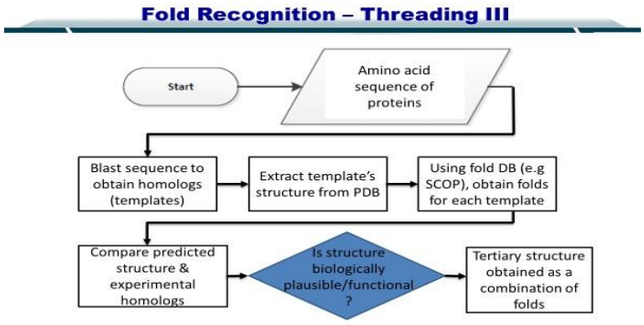
Threading involves "passing" the amino acid sequence through each fold in the database
The best match is computed using a scoring function

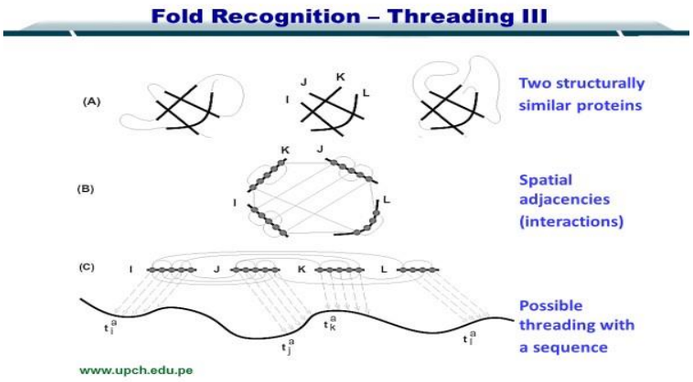
229 GOR Algorithm

Background

Threading involves "passing" the amino acid sequence through each fold in the database ✓ The best match is computed using a scoring function ✓ Flowchart







Conclusions

Combinations of secondary structures come together to form the best prediction
Scoring typically involves using a Z-Score function based on energy of a molecule

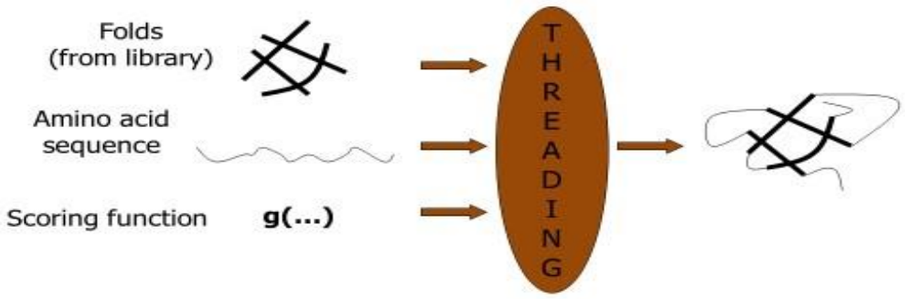
230 Online Tools for Threading – iTasser

Background

Threading involves "passing" the amino acid sequence through each fold in the database
The best match is computed using a scoring function ✓ iTASSER

Online Tools for Threading - iTasser

Inputs and outputs of threading

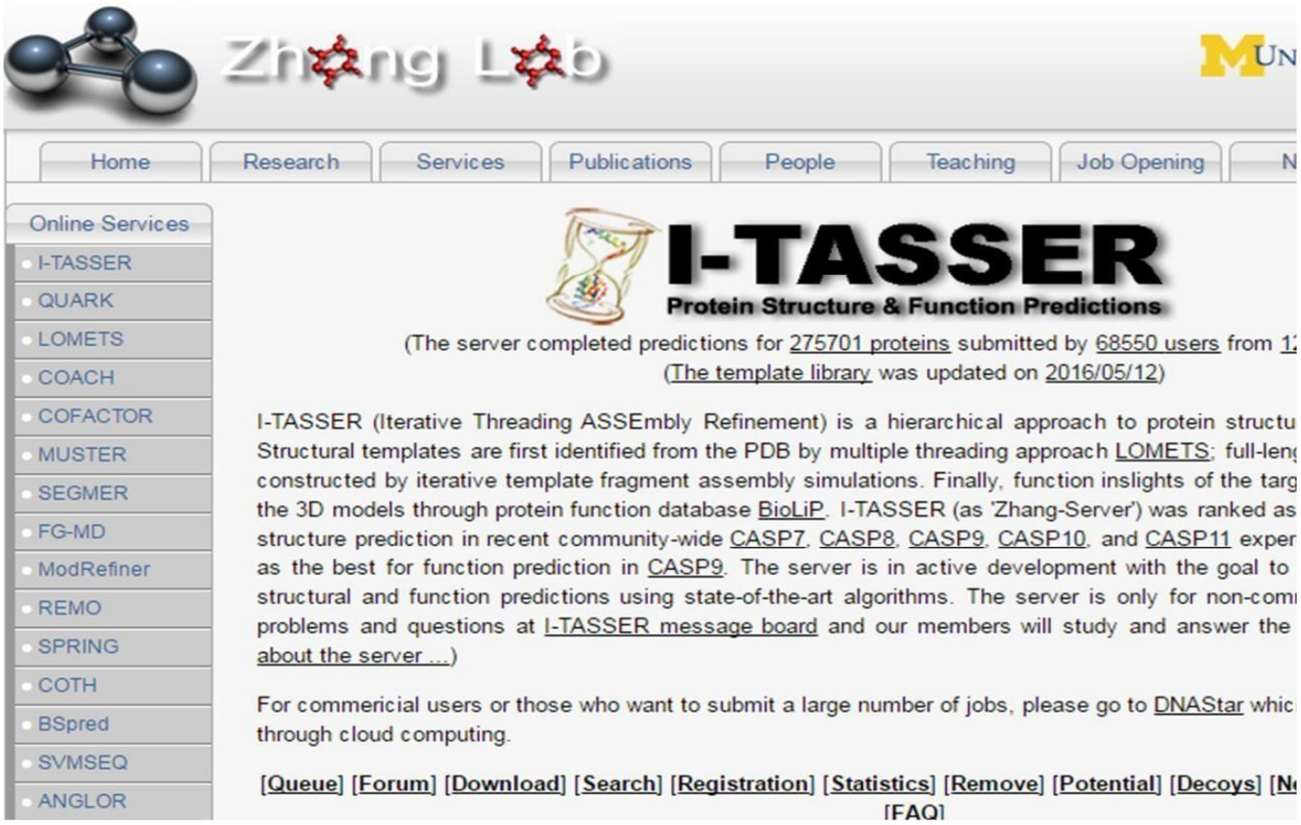


Iterative threading assembly refinement (I-TASSER) server

Software for automated protein structure &function prediction based on the sequencetostructure-to-function.

Steps:

1. Starts from amino acid sequence
2. i-TASSER first generates 3D atomic models from multiple threading alignments and iterative structural assembly simulations.
    i. The function of the protein is then inferred by structurally matching the 3D models with other known proteins.
3. Outputs full-length secondary & tertiary structures and functional annotations on ligandbinding sites
4. An estimate of accuracy of the predictions is provided based on the confidence score of the modeling



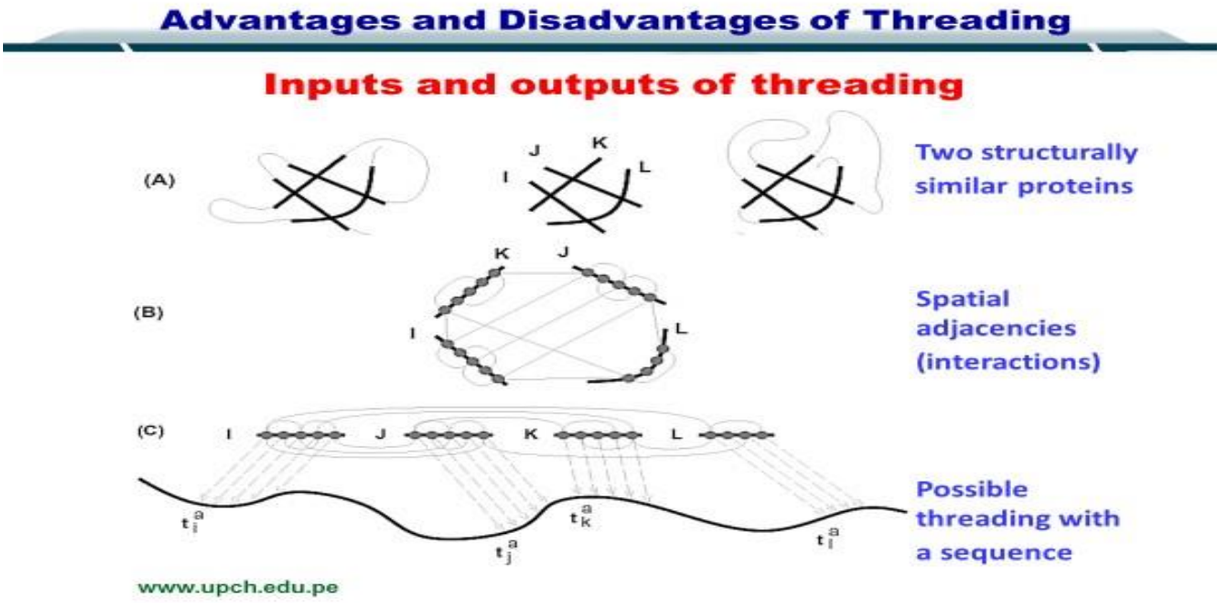
Conclusions

✓ iTASSER helps thread amino acid sequences on fold and secondary structure databases ✓ It also helps predict function of structures output.

231 Advantages and Disadvantages of Threading

Background

Fold recognition or Threading is a technique for predicting protein structures



▪ It is useful in cases where homology modelling fails to predict quality

structures Advantages

Threading helps predict secondary structures of proteins towards tertiary structure prediction ✓ For the "Twilight Zone" with low alignment quality and identity, threading is useful

Disadvantages

Novel proteins cannot be predicted using threading
Fewer than 30% of the predicted first hits are true remote homologues

✓ Validation of each result is necessary

232 Machine Learning Approaches to Structure Prediction

Introduction

- Proposed by Bowie et al in 1991
- Converts 3D structure into a 1-D string profile for each structure in the fold library
- Align the target sequence to these profiles

3D-1D Bowie Algorithm

Inputs and outputs of 3D-1D

- Identify amino acids based on:
  protein core, side chain positioning, solubility
  etc. (6 in all)
- Part of secondary structure including α-helix, β-sheet etc (3 in all)
- Total of 3 x 6 = 18 distinct states

3D-1D Bowie Algorithm

Inputs and outputs of 3D-1D

$P_{a:j}$ = prob. of finding amino acid (a) in environment (j)

P_a = probability of finding (a) anywhere

Maximize sum of scores for the fold:

$$s_{aj} = \log\left(\frac{P_{a:j}}{P_a}\right)$$

Conclusion

- 3D-1D methods convert structure and environment information into "profiles"
- Score for each amino acid is computed for each profile

233 Introduction to Ab-Initio Modelling

Background

Ab initio methods have Anfinsen's thermodynamic hypothesis at the center
These methods attempt to identify the structure with minimum free energy

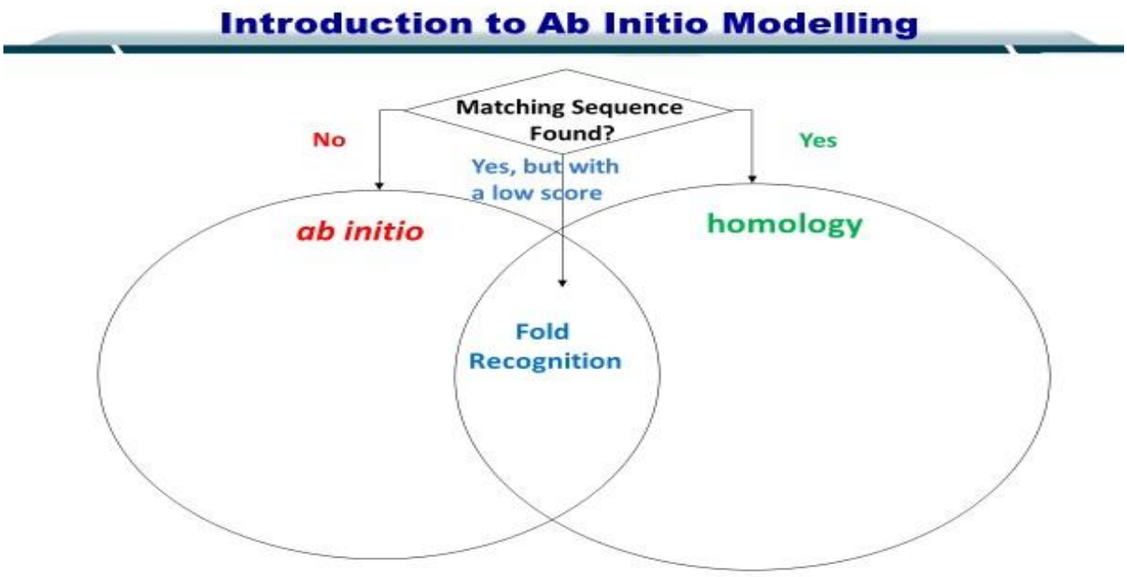
Need for Ab Initio Modelling

Applicable to any sequence
Not very accurate biologically
Accuracy and applicability are limited by our understanding of the protein folding problem

Limitation

Computationally expensive

Suitable for proteins with less than 100 residues



Conclusion

Ab initio methods rely on computing the energies of folded proteins
The protein structures with the lowest energy are deemed as plausible predictions

234 Rationale of Ab Initio Modelling

Background

Rationale

Ab initio methods rely on computing the energies of folded proteins
The protein structures with the lowest energy are declared as plausible predictions ✓
Sometimes it so happens that even slightly homologous proteins may not be available.
This renders homology modelling and threading/fold recognition as futile
Also, newer protein structures continue to be discovered every day
These could not have been identified by methods which only rely on matching with available structures

Lastly, homology / fold recognition predict protein structures without computing fundamental physical/chemical properties of the mechanisms and driving forces in structure formation

Conclusion

Ab initio methods, in contrast, base their predictions on physical models for these mechanisms
Energy released during the folding process is computed for predicting structure

235 Strategies for Ab Initio Modelling

Background

✓ Ab initio methods base their predictions on physical models of folding mechanisms
✓ Stabilization is measured by energy released during the folding process

Energy Optimization in Ab Initio Modelling

1. Start with a rough initial model.

2. Define an energy function mapping structures to energy values. We have to minimize this later!!
3. Solve the computational problem of finding the global minimum.

Simulation of the Folding Process

1. Build an accurate initial model (including energy and forces).
2. Accurately simulate the dynamics of the protein folding process.
3. The native structure will steadily emerge.

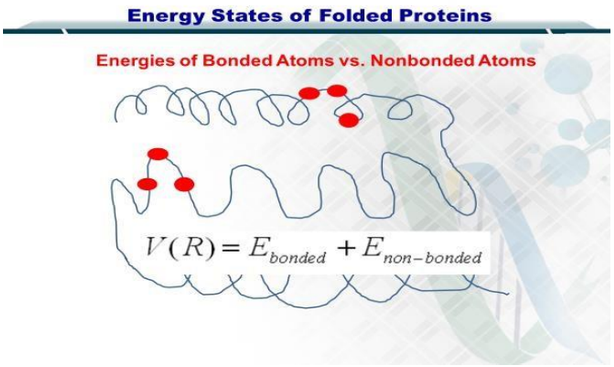
Conclusion

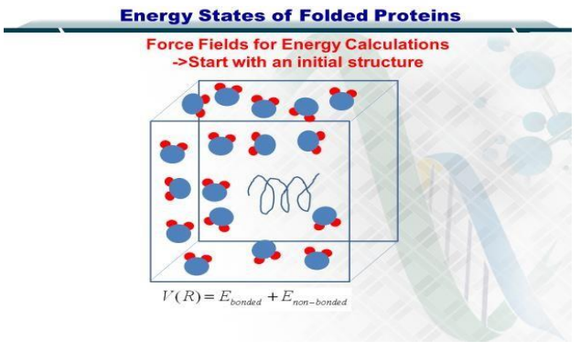
✓ Start with an energy function Fold structures in order to obtain the most stable structure This structure will have the minimum energy

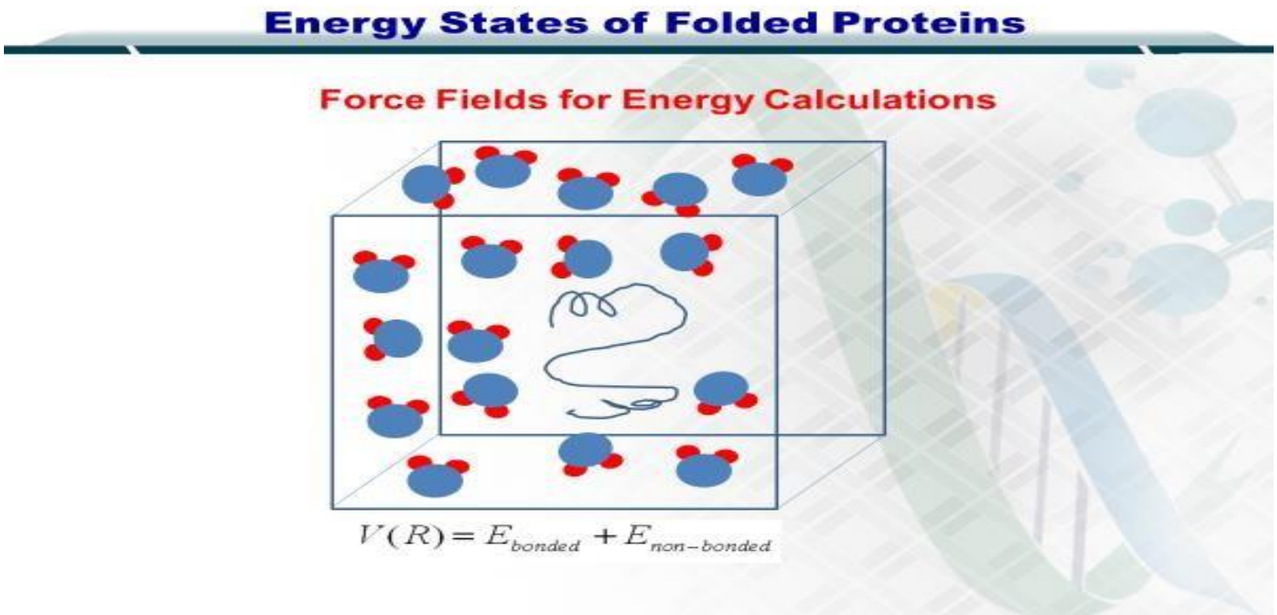
236 Energy States of Folded Proteins

Background

Ab initio methods predict protein structures by folding proteins based on each constituent atom's volume, charge, mass etc.







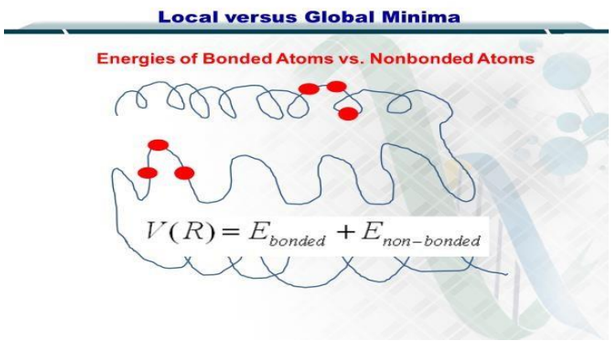
Conclusion

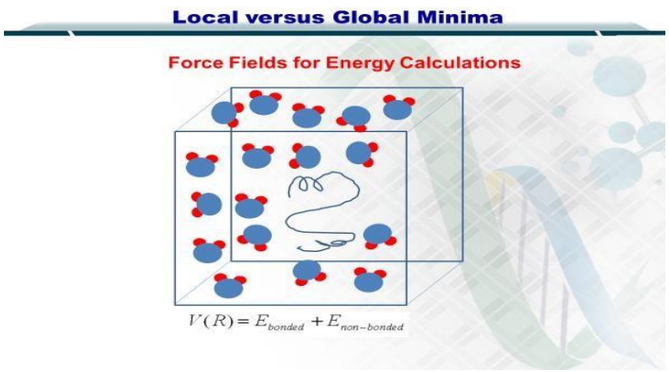
The protein structure reporting lowest energy is selected to be the optimal structure ✓
How easy is it to compute the "really" lowest energy of a folded protein?

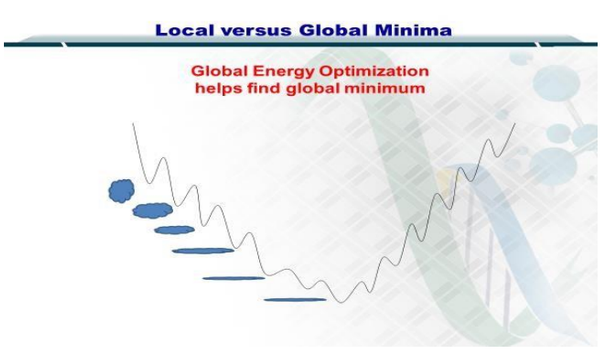
237 Local versus Global Minima

Background

The protein structure reporting lowest energy is selected to be the optimal structure ✓
How easy is it to compute the "really" lowest energy of a folded protein?









Best Case Energy Function

Clear energy minimum in the native structure
Viable path towards this minimum
Global optimization finds the most stable structure

Optimal Energy Function

Easier to design and compute
Native structure not always at the global minimum
No clear way of choosing among alternative structures that are generated

238 Pros and Cons of Ab Initio Modelling

Background

Native structure not always at the global minimum
No clear way of choosing among alternative structures that are generated

Advantages

a. Ab Initio methods can fold any target sequence using only physical atomic properties

b. Predictions are mostly accurate and correctly describe the natural folding process

Disadvantages

1. Ab initio methods are the very difficult to design (energy function)
2. These methods are slow due to the huge possibilities
3. An order of 10^{12} steps are needed to simulate protein folding for medium sized protein structures

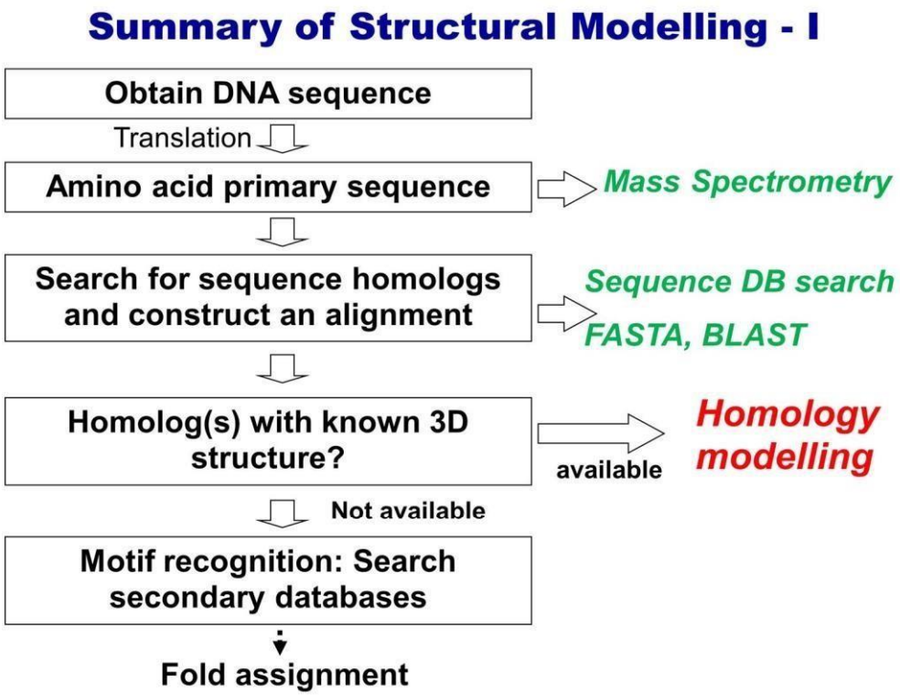
Challenges in Ab Initio Modelling

Very hard to accurately describe energy functions that can reliably discriminate native and non-native structures.
Enormous amount of computations.

239 Summary of Structural Modelling – I

Strategies for Structural Modelling

1. Homology Modelling
2. Fold Recognition
3. Ab Initio Modelling



Homology modeling of the target structure can be done as follows:

1. Template recognition and initial alignment
2. Alignment correction
3. Backbone generation
4. Loop modeling
5. Side-chain modeling
6. Model optimization
7. Model validation

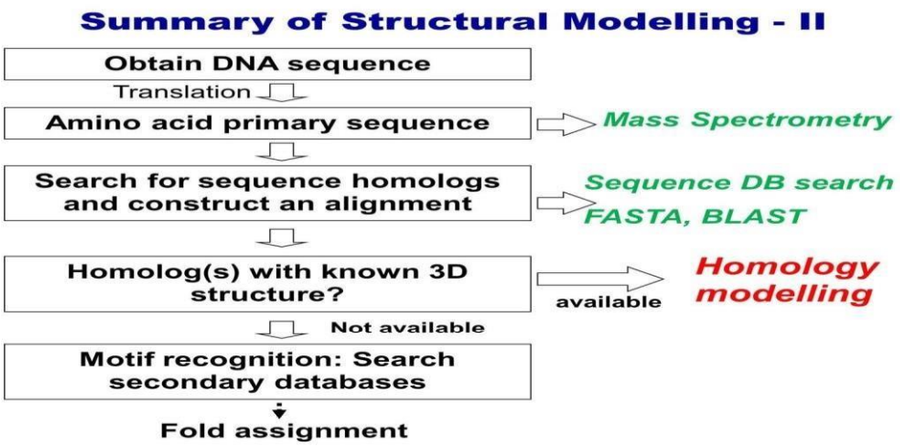
Conclusion

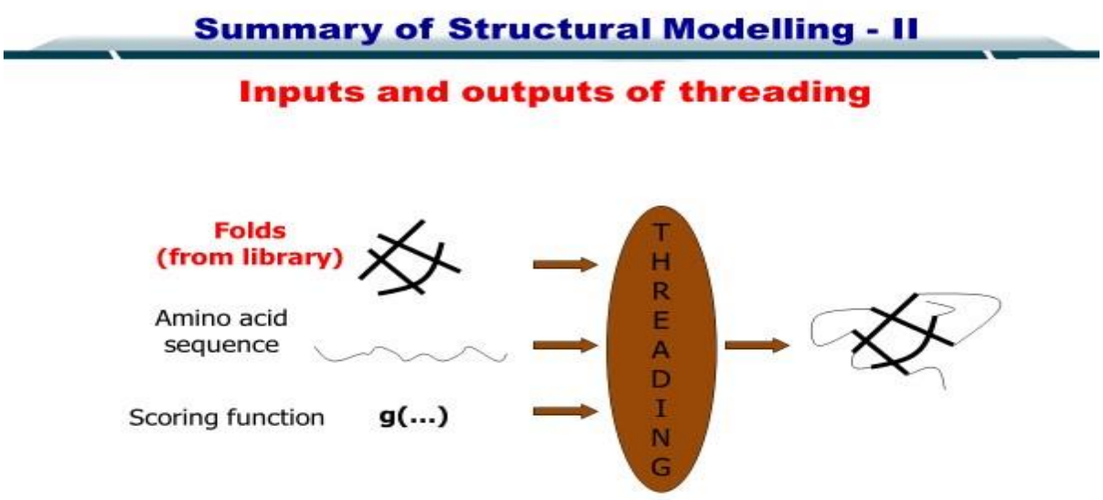
Homology modelling is performed in cases of high identity and alignment score
For the “Twilight zone”, other strategies are employed

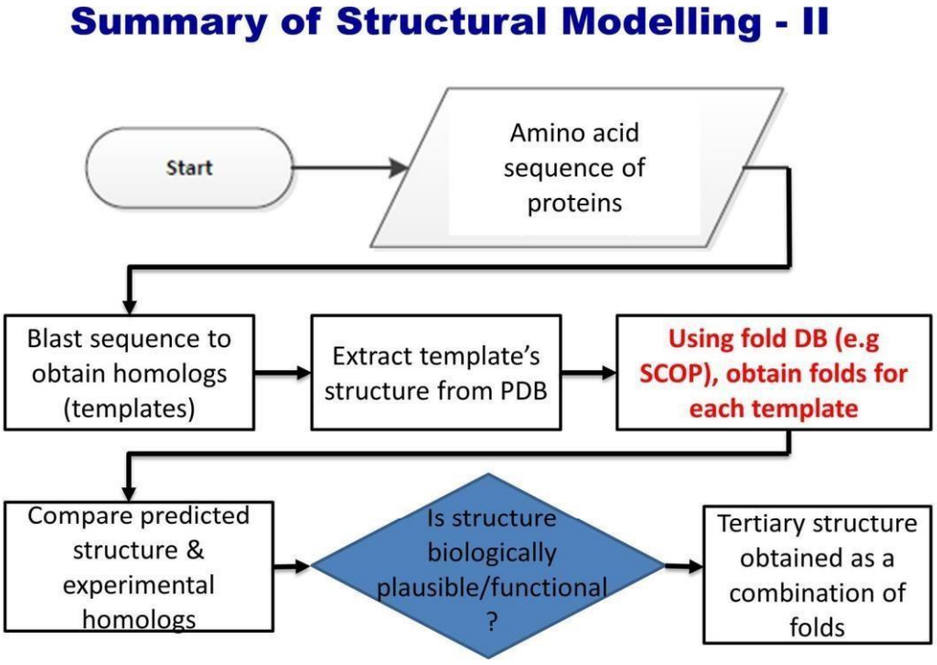
240 Summary of Structural Modelling – II

Strategies for Structural Modelling

1. Homology Modelling
2. Fold Recognition
3. Ab Initio Modellin







Conclusion

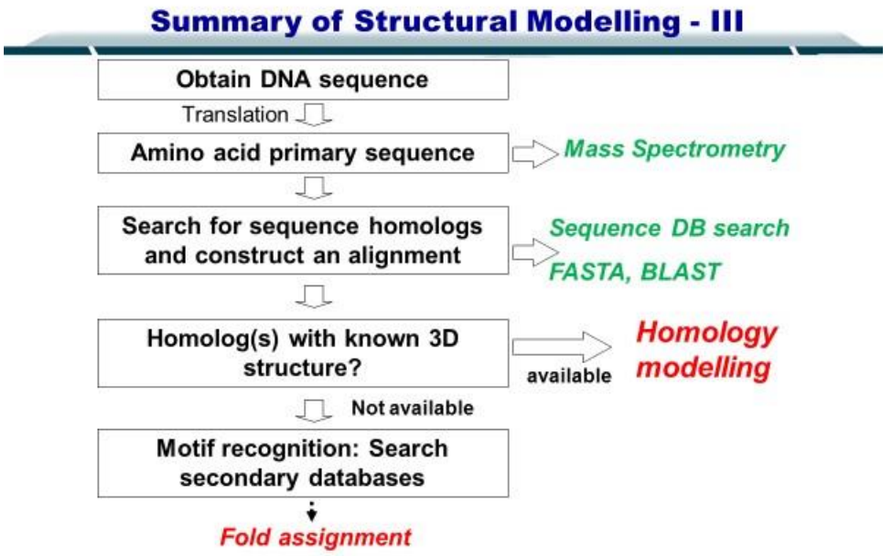
1. For low identity and alignment scores, a "Twilight zone" for structure prediction exists

2. Fold recognition / threading is useful in such cases

241 Summary of Structural Modelling – III

Strategies for Structural Modelling

1. Homology Modelling
2. Fold Recognition
3. Ab Initio Modelling Energy Optimization in Ab Initio Modelling



1) Start with a rough initial model.
2) Define an energy function mapping structures to energy values. We have to minimize this later!!
3) Solve the computational problem of finding the global minimum.

Simulation of the Folding Process

1. Build an accurate initial model (including energy and forces).
2. Accurately simulate the dynamics of the protein folding process.
3. The native structure will steadily emerge.

Conclusion

For cases where even the fold libraries do not give any high scoring matches, Ab Initio strategies can help model the structure
However, this is a complex and computationally expensive process

242 Review of Sequence Analysis

Important Concepts

Types of Alignments:

Global Alignment (Needle Wunsch)
Local Alignment (Smith Waterman)

Advanced Tools:

Fast Alignment (FASTA)
Basic Local Alignment Search Tool (BLAST)

Databases:

GenBank

UniProt

Online Portals:

Ensemble Expasy

UniProtKB

243 Review of Phylogenetics

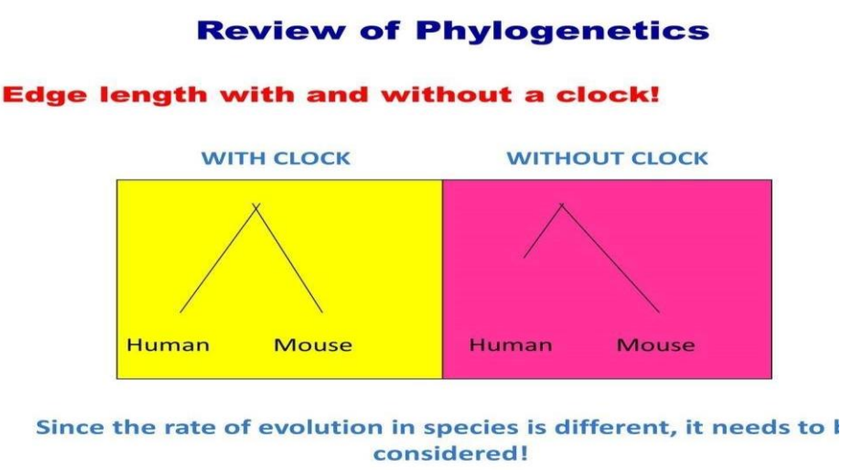
Important Concepts

Molecular Evolution

1. Insertions
2. Deletions
3. Substitutions

Phylogenetic Trees

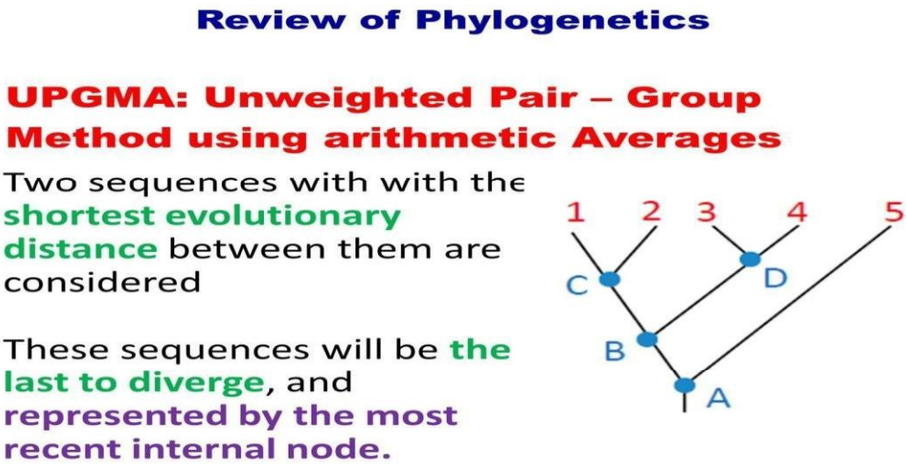
1. Scaled Trees
2. Unscaled Trees



Phylogenetic Trees

Rooted Trees

Unrooted Trees



Clustering Vs. Non-clustering Methods:

UPGMA is a clustering method

Maximum Parsimony etc are non-clustering methods (not included in this course).

244 Review of Protein Sequencing

Important Concepts

Techniques of protein sequencing
Edman Degradation Mass Spectrometry

Protein Ionization
Mass Analysis
Protein Fragmentation
MS1
MS2
Estimating and scoring whole protein mass
Extracting & Scoring Peptide Sequence Tags
Searching Post-translational Modifications

Composite Scoring Schemes

Online tools:

Mascot
Sequest
Prosight PC

245 Review of RNA Structure Prediction

Important Concepts

Role of RNA in biological processes
Atomic Force Microscopy
Need for Structure Prediction

RNA Secondary Structures

1. Hairpin Loops
2. Bulges
3. Helices
4. Intersection

Conceptual basis for structure prediction

Energy is released as a result of nucleotide coupling
Folded RNA is more stable in terms of energy

Algorithms for predicting RNA Structure

Dot plot
Zuker's Algorithm
Martinez Algorithm
Nussinov Jacob Algorithm

- RNA Structure Databases
- Online tools for predicting structures given a sequence

246 Review of Protein Structures

Important Concepts

Protein Structures are generally of four types:

i. Primary
ii. Secondary
iii. Tertiary
iv. Quaternary

Techniques for determining protein structures

- X-Ray Crystallography
- NMR Spectroscopy
- Why number of known protein sequences is much larger as compared to known proteins structures?

Types of Protein Secondary Structures

- Helices
- Beta Sheets
- Coils
- Loops

Foundation of structure prediction algorithms

Propensities of certain amino acids to form specific secondary structures
Algorithm for predicting protein structures
Chou Fasman Algorithm
Protein Structure Database - PDB
Online tools for predicting structures by using proteins sequences

247 Review of Homology Modelling

Important Concepts

Four Strata of Protein Structures

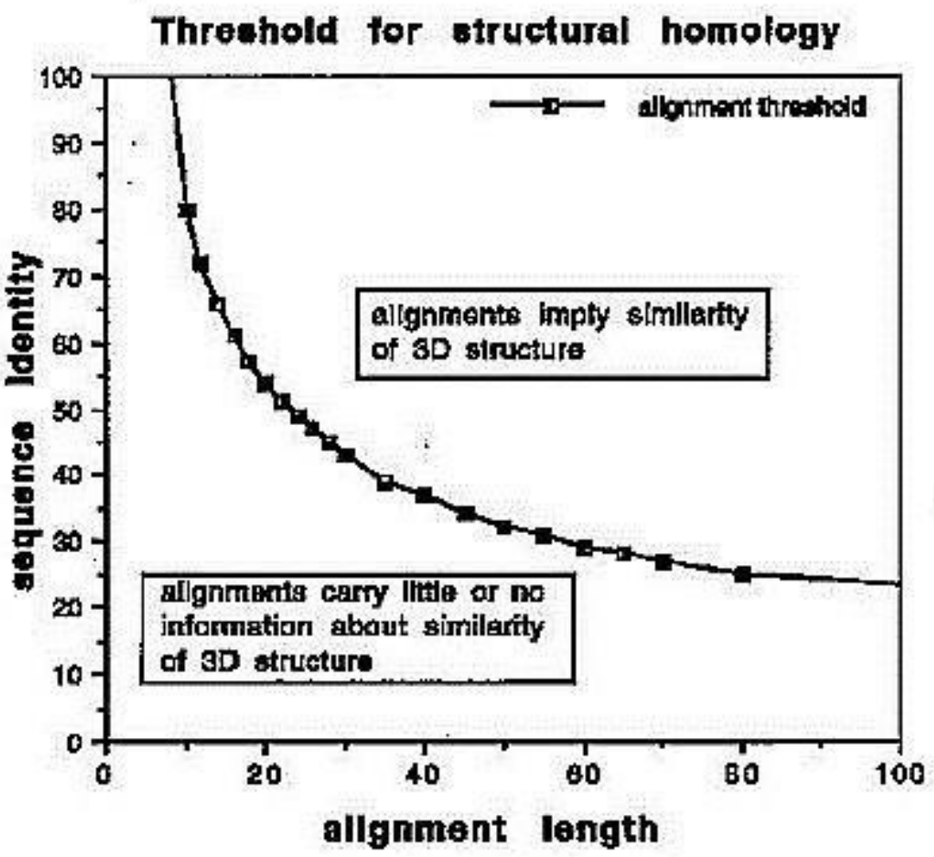
- Primary
- Secondary
- Tertiary
- Quaternary

Justification for homology modelling

Number of known protein sequences is much larger as compared to known proteins structures

Three Strategies for Structure Prediction

- Homology Modelling
- Fold Recognition
- Ab Initio Modelling



Protein Structure Database - PDB
Online tools for predicting structures such as MODELLER and iTASSER

248 Conclusions from this Course

Important Concepts

- Definition of Bioinformatics
- Need for Bioinformatics Areas within Bioinformatics
- Bioinformatics as an interdisciplinary area
- Need to store, process and analyze biological data
- Requirement of newer faster algorithms

Specific areas focused were:

- Comparing sequences
- Comparing structures
- Predicting structures
- We studied the basic algorithms for each topic

With evolution and growth of Bioinformatics, newer and better algorithms are now also available!

249 Advanced Follow-up Courses

Important Concepts

- We looked into the foundations of Bioinformatics
- However, each topic that was studied has a undergone a lot of development
- For advanced study in Genomics, you may take “Computational Genomics” course
- Genome Assembly, Gene Finding, Annotation, GWAS etc

- For advanced study in Proteomics, you may take “Computational Proteomics” course.

Protein Sequencing, PTM search, Structure Modelling and PPI studies
For advanced study in Integrative Biology, you may take “Systems Biology” course.

Metabolomics, Transcriptomics, Network Biology etc

Also, now there are cutting edge courses on:

Nano-Bio-IT
Computational Drug Design

Personalized Medicine

250 Careers in Bioinformatics

- Pakistan as an infrastructure-limited country
- The onset of digital revolution
- Emergence of data as the most precious commodity, globally
- Specifically, health data as a key commodity of the future
- Health and disease as the primordial challenge of mankind
- Unique opportunity for us in Pakistan
- Bioinformatics requires two things
- Smart mind
- Internet connected computer

- You can take public databases and design drugs!

One man vs. Roche?

BIGDATA

- You can make a startup company which manages and process health BIGDATA!
- All it needs is basic software development skills coupled with

The next disruption

- The next Google, Facebook and Uber is going to emerge from Health and Bioinformatics
- Pharmaceutical companies are investing into bioinformatics human resource development

Jobs Market

- Pharmaceutical Giants
- Research Centers & Universities Hospital & Diagnostic IT departments

- Your own startup company