

AN INTRODUCTORY COURSE
BIOINFORMATICS-I

A STUDENT HANDOUT

2016

VER 1.0.0.0

LAST REVISED ON May 2, 2016

Contents

Chapter 1- Introduction to Bioinformatics	1
Module 001: INTRODUCTION TO BIOINFORMATICS	1
Module 002: INTRODUCTION TO BIOINFORMATICS	2
Module 003: INTRODUCTION TO BIOINFORMATICS	3
Module 004: NEED FOR BIOINFORMATICS –II	4
Module 005: APPLICATIONS OF BIOINFORMATICS – I.....	5
Module 006: APPLICATIONS OF BIOINFORMATICS - II	6
Module 007: FRONTIERS IN BIOINFORMATICS - I.....	7
Module 008: FRONTIER IN BIOINFORMATICS-II	9
Module 009: Overview of Course Contents - I.....	10
Module 010: Overview of Course Contents - II.....	12
Module 011: Overview of Course Contents – III.....	15
Chapter 2 - Sequence Analysis.....	18
Module 001: Gene, mRNA and Protein Sequences	18
Module 002: TRANSCRIPTION.....	20
Module 003: NUCLEOTIDES	21
Module 004: TRANSLATION	23
Module 005: AMINO ACIDS	25
Module 006: STORAGE OF BIOLOGICAL SEQUENCE INFORMATION	27
Module 007: USING GENBANK	28
Module 008: USING UNIPROT.....	30
Module 009: COMPARING SEQUENCES.....	32
Module 010: SIMILARITIES & DIFFERENCES IN SEQUENCES.....	33
Module 011: SIMILARITIES & DIFFERENCES IN SEQUENCES.....	34
Module 012: PAIR WISE ALIGNMENT –II	35
Module 013: PAIR WISE SEQUENCE ALIGNMENT –III.....	36
Module 014: DOT PLOTS.....	37
Module 015: EXEMPLE OF DOT PLOTS.....	39
Module 016: IDENTITY VS SIMILARITY	40
Module 017: INTRODUCTION TO ALIGNMENT APPROACHES	41
Module 018: WHY LOCAL ALIGNMENT.....	42

Module 019: ALIGNING, INSERTION & DELETION	43
Module 020: ALIGNING MUTATION IN SEQUENCES.....	44
Module 021: INTRODUCTION TO DYNAMIC PROGRAMMING	45
Module 022: DYNAMIC PROGRAMMING ESSENTIALS.....	46
Module 023: DYNAMIC PROGRAMMING METHODOLOGY	47
Module 024: NEEDLEMAN WUNSCH ALGORITHM-I.....	48
Module 025: NEEDLEMAN WUNSCH ALGORITHM-II.....	50
Module 026: NEEDLEMAN WUNSCH ALGORITHM-III.....	52
Module 027: NEEDLEMAN WUNSCH ALGORITHM EXAMPLE.....	53
Module 028: BACKTRACKING ALIGNMENTS.....	55
Module 029: REVISITING LOCAL AND GLOBAL ALIGNMENTS.....	56
Module 030: OVERLAP MATCHES.....	57
Module 031: EXAMPLE	58
Module 032: MOVING FROM GLOBAL TO LOCAL ALIGNMENT	59
Module 033: SMITH WATERMAN ALGORITHM	60
Module 034: EXAMPLE OF SMITH WATERMAN ALGORITHM	61
Module 035: REPEATED ALIGNMENTS	62
Module 036: EXAMPLES OF REPEATED ALIGNMENTS.....	63
Module 037: INTRODUCTION TO SCORING ALIGNMENTS	64
Module 038: MEASURING ALIGNMENTS SCORES	65
Module 039: SCORING MATRICES	66
Module 040: DERIVING SCORING MATRICES.....	67
Module 041: PAM MATRICES.....	69
Module 042: BLOSUM MATRICES.....	70
Module 043: MULTIPLE SEQUENCE ALIGNMENT	71
Module 044: MORE ON MULTIPLE SEQUENCE ALIGNMENT	72
Module 045: PROGRESSIVE ALIGNMENT FOR MSA.....	74
Module 046: MSA-EXAMPLE.....	76
Module 047: CLUSTALW	79
Module 048: INTRODUCTION TO BLAST-I.....	81
Module 049: INTRODUCTION TO BLAST-II.....	83
Module 050: BLAST ALGORITHM	85
Module 051: TYPES OF BLAST	86

Module 052: SUMMERY OF BLAST	87
Module 053: INTRODUCTION TO FASTA.....	91
Module 054: INTRODUCTION TO FASTA-II	94
Module 055: FASTA ALGORITHM.....	96
Module 056: TYPES OF FASTA.....	98
Module 057: SUMMERY OF FASTA	99
Module 058: BIOLOGICAL DATABASE AND ONLINE TOOLS.....	103
Module 059: EXPASY	104
Module 060: UNIPROT AND SWISSPROT	110
Module 061: PROTEIN DATA BANK.....	112
Module 062: REVIEW OF SEQUENCE ALIGNMENT	114
Module 063: GENBANK.....	115
Module 064: ENSEMBLE	116
Chapter 3 - Molecular Evolution	117
Module 001: Molecular Evolution & Phylogeny.....	117
Module 002: Evolution of Sequences	119
Module 003: Concepts and Terminologies - I	120
Module 004: Concepts and Terminologies - II	122
Module 005: Algorithms and Techniques	123
Module 006: Introduction to UPGMA.....	124
Module 007: UPGMA-I.....	126
Module 008: UPGMA-II	128
Module 009: UPGMA-III.....	131
Module 010: UPGMA-IV.....	133
Module 011: UPGMA-V.....	135
Chapter 4 - RNA Secondary Structure Prediction	137
Module 001: DNA TO RNA SEQUENCES.....	137
Module 002: TYPES OF RNA &THEIR FUNCTIONS.....	138
Module 003: SIGNIFICANCES OF RNA STRUCTURE.....	140
Module 004: RNA FOLDING AND ENERGY FOLDING	142
Module 005: CALCULATING ENERGIES OF FOLDING-AN EXAMPLE	143
Module 006: TYPES OF RNA SECONDARY STRUCTURES-I	144
Module 007: TYPES OF RNA SECONDARY STRUCTURES-II.....	146

Module 008: TYPES OF RNA SECONDARY STRUCTURES-III	147
Module 009: RNA TERTIARY STRUCTURES.....	148
Module 010: CIRCULAR REPRESENTATION OF STRUCTURES.....	149
Module 011: EXPERIMENTAL METHODS TO DETERMINE RNA STRUCTURES.....	150
Module 012: Strategies for RNA Structure Prediction	151
Module 013: Dot Plots for RNA 2' Structure Prediction	152
Module 014: ENERGY BASES METHODS	153
Module 015: Zuker's Algorithm	154
Module 016: Zuker's Algorithm EXAMPLE	155
Module 017: Zuker's Algorithm – A Flow Chart.....	156
Module 018: Martinez Algorithm	158
Module 019: Dynamic Programming Approaches	158
Module 020: Nussinov-Jacobson Algorithm – An Overview.....	160
Module 021: Nussinov-Jacobson Algorithm – The Flowchart	161
Module 022: Nussinov-Jacobson Algorithm – EXAMPLE	162
Module 023: Score Calculations & Traceback.....	163
Module 024: Comparison of Algorithms.....	164
Module 025: WEB RESOURCES-I	165
Module 026: WEB RESOURCES-II	167
Chapter 5 - Protein Sequences	168
Module 001: FROM DNA/RNA SEQUENCES TO PROTEIN	168
Module 002: CODING OF AMINO ACIDS.....	170
Module 003: OPEN READING FRAMES	171
Module 004: ORF Extraction – A Flowchart.....	172
Module 005: SEQUENCING PROTEINS	173
Module 006: Application of Mass Spectrometry in Protein Sequencing	175
Module 007: Techniques for MS Proteomics.....	177
Module 008: Types of MS-based proteomics	178
Module 009: BOTTOM UP PROTEOMICS	179
Module 010: Two Approaches for Bottom up Proteomics (BUP).....	180
Module 011: TOP DOWN PROTEOMICS.....	181
Module 012: PROTEIN SEQUENCE IDENTIFICATION.....	182
Module 013: PROTEIN IONIZATION TECHNIQUES	183

Module 014: MS1 & INTACT PROTEIN MASS.....	184
Module 015: SCORING INTACT PROTEIN MASS.....	185
Module 016: PROTEIN FRAGMENTATION TECHNIQUES.....	186
Module 017: TANDEM MS	187
Module 018: MEASURING EXPERIMENTAL FRAGMENT'S MASS	188
Module 019: Calculating Theoretical Fragment's Mass.....	189
Module 020: PEPTIDE SEQUENCE TAG	191
Module 021: Extracting Peptide Sequence Tags	192
Module 022: Using Peptide Sequence Tags in Protein Search	193
Module 023: Scoring Peptide Sequence Tags.....	194
Module 024: In silico Fragment Comparison	195
Module 025: In silico Fragment Comparison and scoring	196
Module 026: Protein Sequence Database Search Algorithm.....	197
Module 027: Integrative Scoring Schemes	198
Module 028: LARGE SCALE PROTEOMICS	199
Module 029: PROTEOMICS DATA FILE FORMATS.....	200
Module 030: RAW FILE FORMAT	202
Module 031: MGF FILE FORMAT.....	203
Module 032: OPEN MS DATA FORMAT	206
Module 033: Online Proteomics Tools - MASCOT	209
Module 034: Online Proteomics Tools – ProSight PTM.....	211
Module 035: Example Case Study - I.....	214
Module 036: Example Case Study – II.....	216
Module 037: Example Case Study – III.....	217
Module 038: Example Case Study – IV.....	218
Module 039: Example Case Study – V.....	219
Module 040: Example Case Study – VI.....	220
Chapter 6 - Protein Structures	221
Module 001: PROPERTIES OF AMINO ACIDS-I	221
Module 002: PROPERTIES OF AMINO ACIDS-II	223
Module 003: PROPERTIES OF AMINO ACIDS-III	225
Module 004: PROPERTIES OF AMINO ACIDS-IV	226
Module 005: PROPERTIES OF AMINO ACIDS-V	228

Module 006: STRUCTURAL TRAIT OF AMINO ACID-I	229
Module 007: STRUCTURAL TRAIT OF AMINO ACID-II	231
Module 008: STRUCTURAL TRAIT OF AMINO ACID-III	232
Module 09: STRUCTURAL TRAIT OF AMINO ACID-IV	233
Module 010: INTRODUCTION TO PROTEIN FOLDING	234
Module 011: IMPORTANCE OF PROTEIN FOLDING.....	235
Module 012: COMPUTING PROTEIN FOLDING POSSIBILITIES	236
Module 013: PROCESSING OF PROTEIN FOLDING	237
Module 014: MODELS OF PROTEIN FOLDING.....	239
Module 015: PROTEIN STRUCTURE	241
Module 016: Primary, Secondary, Tertiary and Quaternary Structures.....	243
Module 017: Primary Structures of Protein.....	244
Module 018: Secondary Structures of Proteins - I.....	246
Module 019: Secondary Structures of Proteins – II	248
Module 020: Tertiary Structures of Proteins	250
Module 021: Quaternary Structures of Proteins	251
Module 022: Introduction to Bond Angles in Proteins	252
Module 023: Ramachandran Plot	254
Module 024: Structure Visualization - I	256
Module 025: Structure Visualization – II.....	257
Module 026: Experimental Determination of Protein Structure.....	258
Module 027: Protein Databank.....	259
Module 028: Visualization Technique.....	262
Module 029: Online Resources for Protein Visualization	263
Module 030: Types of Protein Visualizations.....	265
Module 031: Introduction to Energy of Protein Structures.....	267
Module 032: Calculating Energy of a Protein Structure	268
Module 033: Structure Determination for Energy Calculations	269
Module 034: Review of Experimental Structure Determination	270
Module 035: Alpha Helices - I	272
Module 036: Alpha Helices – II	273
Module 037: Beta Sheets - I.....	275
Module 038: Beta Sheets - II.....	276

Module 039: LOOPS-I.....	278
Module 040: LOOPS-II.....	280
Module 041: COILS.....	282
Module 042: Structure Classification - I.....	283
Module 043: Structure Classification - II.....	285
Module 044: Examples of Protein Domains.....	286
Module 045: CATH Classification	288
Module 046: Classification Databases	290
Module 047: Algorithms for Structure Classification.....	292
Module 048: Protein Structure Comparison.....	293
Module 049: Strategies for Structure Comparison - I.....	294
Module 050: Strategies for Structure Comparison - II.....	295
Module 051: Online Resources for Structure Comparison.....	296
Module 052: Protein Structure Prediction.....	298
Module 053: Predicting Secondary Structures	299
Module 054: Introduction to Chou Fasman Algorithm.....	300
Module 055: 2' Structures in Chou Fasman Algorithm.....	301
Module 056: Chou Fasman Algorithm - I.....	302
Module 057: Chou Fasman Algorithm - II	303
Module 058: Chou Fasman Algorithm - III	304
Module 059: Chou Fasman Algorithm - IV.....	305
Module 060: Chou Fasman Algorithm – Flowchart I	307
Module 061: Chou Fasman Algorithm – Flowchart II	308
Module 063: Chou Fasman Algorithm – Improvements.....	311
Module 064: Summary of Visualization, Classification & Prediction.....	312

Chapter 1- Introduction to Bioinformatics

Module 001: INTRODUCTION TO BIOINFORMATICS

- **BACKGROUND**

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

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

- **EXPERIMENTS IN BIOLOGY**

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

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

- **DIGITALIZATION OF BIOLOGY**

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

- **SPEED OF DATA GROWTH**

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

- **CONCLUSION**

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

Module 002: INTRODUCTION TO BIOINFORMATICS

• MOTIVATION

Bioinformatics is becoming a popular science due to several reasons.

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

• SCOPE OF BIOINFORMATICS

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

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

• ACTIVITIES

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

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

• CONCLUSION

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

Module 003: INTRODUCTION TO BIOINFORMATICS

- **NEED FOR BIOINFORMATICS –I**

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

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

- **CONCLUSION**

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

Module 004: NEED FOR BIOINFORMATICS –II

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

- Bacteria
- Archaea
- Eucarya

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

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

The simple answer to that bioinformatics is:

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

- **POSSIBLE CONTRIBUTIONS**

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

- **CONCLUSION**

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

Module 005: APPLICATIONS OF BIOINFORMATICS – I

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

The answers to these questions are categorized as follows:

- **GENOMICS**

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

- **EVOLUTIONARY STUDIES**

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

- **PROTEOMICS**

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

- **SYSTEMS BIOLOGY**

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

- **CONCLUSION**

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

Module 006: APPLICATIONS OF BIOINFORMATICS - II

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

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

- **SMALL TO BIG**

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

- **CONCLUSION**

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

Module 007: FRONTIERS IN BIOINFORMATICS - I

- **INTROCDUCTION**

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

- **FRONTIER IN GENOMICS**

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

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

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

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

- **FRONTIER IN TRANSCRIPTOMICS**

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

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

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

- **FRONTIER IN PROTEOMICS**

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

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

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

- **CONCLUSION**

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

Module 008: FRONTIER IN BIOINFORMATICS-II

Frontier in Bioinformatics includes

- Next generation genomics
- Transcriptomics
- Proteomics

- **FRONTIER IN PROTEIN STURCTURE**

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

- **FRONTIER IN SYSTEM BIOLOGY**

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

- **FRONTIER IN PERSONALIZED MEDICINE**

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

- **CONCLUSION**

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

Module 009: Overview of Course Contents - I

Philosophy behind the Course Outlay

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

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

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

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

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

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

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

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

Module 010: Overview of Course Contents - II

Chapter	Contents
Protein Sequences	From DNA/RNA Sequences to Proteins
	Coding of Amino Acids
	Open Reading Frames
	Sequencing Proteins
	Application of MS in sequencing
	Bottom Up Proteomics
	Top Down Proteomics
	Protein Ionization Techniques
	MS1 and Intact Protein Mass
Chapter	Contents
Protein Sequences	Scoring Intact Protein Mass
	Protein Fragmentation Techniques
	Tandem MS
	Measuring Experimental and Theoretical Fragment's Mass
	Peptide Sequence Tags
	Scoring Peptide Sequence Tags
	In silico Protein Fragmentation
	In silico Fragment Comparison and Scoring

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

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

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

Chapter	Contents
Protein Structures	Calculating Energies of Protein Structures
	Structure Determination for Energy Calculations
	Protein Structures - Alpha Helices, Beta Sheets, Loops, Coils
	Protein Domains
	Classification Databases
	Algorithms for Structure Classification
	Protein Structure Comparison
Chapter	Contents
Protein Structures	Online Resources for Structure Comparison
	Protein Structure Prediction
	Predicting Secondary Structures
	Introduction to Chou Fasman Algorithm

Summary

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

Module 011: Overview of Course Contents – III

Chapter	Contents
Homology Modelling	Introduction to Homology modelling
	Need for Homology Modelling
	Seven Steps to Homology Modelling
	Algorithm And Examples
	Fold Recognition/Threading
	Online Tools for Fold Recognition
	GOR Algorithm
Chapter	Contents
Homology Modelling	3D-1D Bowie Algorithm
	Machine Learning Approaches to Structure Prediction
	Neural Networks for Structure Prediction
	PSIPRED
	Introduction to Hidden Markov Models
	Ab initio modelling
Chapter	Contents
Homology Modelling	Hinds and Levitt Algorithm
	Computational Assessment of Structure Prediction (CASP)
	Online Tools for Homology Modelling
	Databases for Structure Modelling
	Example of Hepatitis C Virus Modelling

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

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

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

Conclusion

- These contents will give you an initial exposure to the variety of topics in bioinformatics
- After covering these topics, you should have a basic conceptual foundation for further studies into Bioinformatic

Chapter 2 - Sequence Analysis

Module 001: Gene, mRNA and Protein Sequences

• INTRODUCTION

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

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

DNA → RNA → Protein.

Proteins are than use in constructing the cell.

• DNA

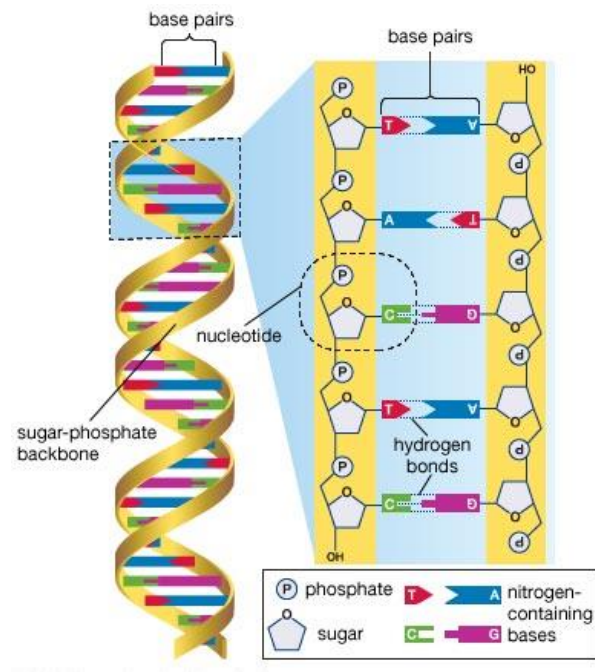


Figure 0.1 DNA Double helix

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

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

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

- **CONCLUSION**

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

Module 002: TRANSCRIPTION

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

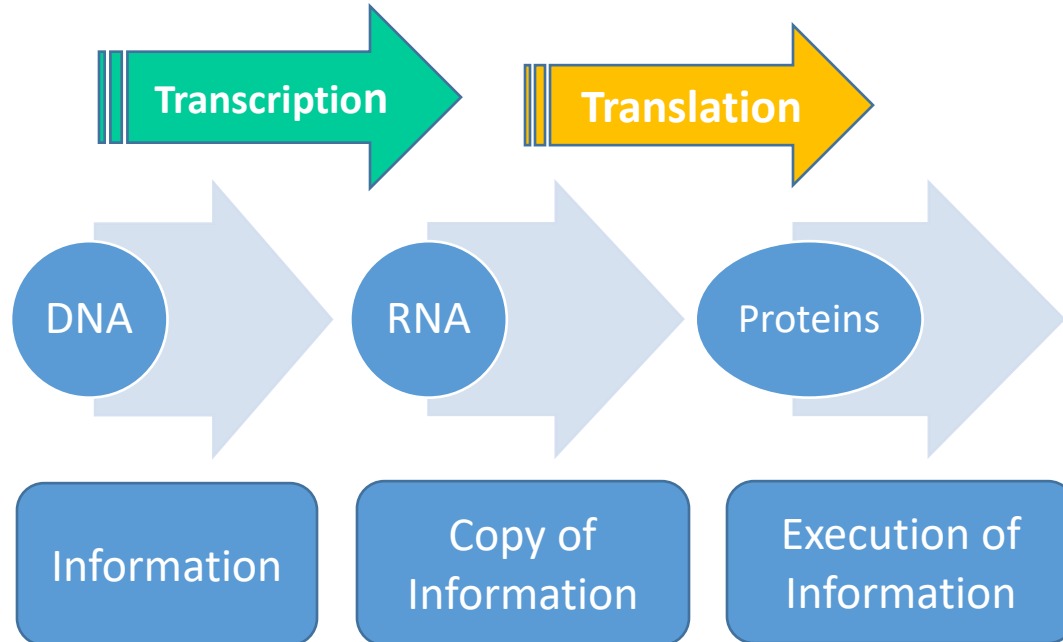


Figure 0.2 Flow of information from DNA to Proteins

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

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

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

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

Module 003: NUCLEOTIDES

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

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

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

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

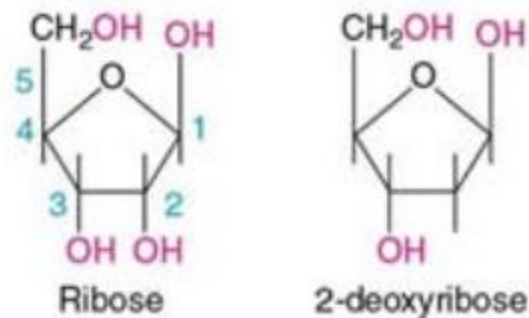


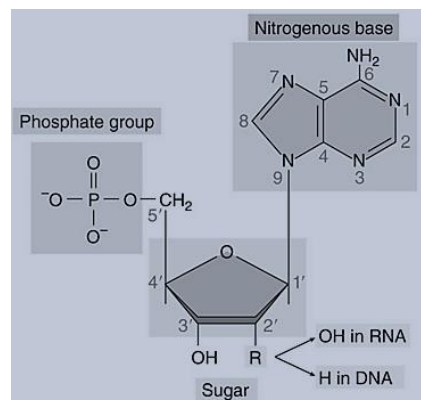
Figure 0.3 Difference between RNA and DNA sugar

RNA

DNA

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

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



- **CONCLUSION**

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

Module 004: TRANSLATION

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

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

- MECHANISM**

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

Amino Acid	3- Letters	1- Letter
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V
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

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

Figure 0.4 sixty four codons combinations

- **CONCLUSION**

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

Module 005: AMINO ACIDS

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

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

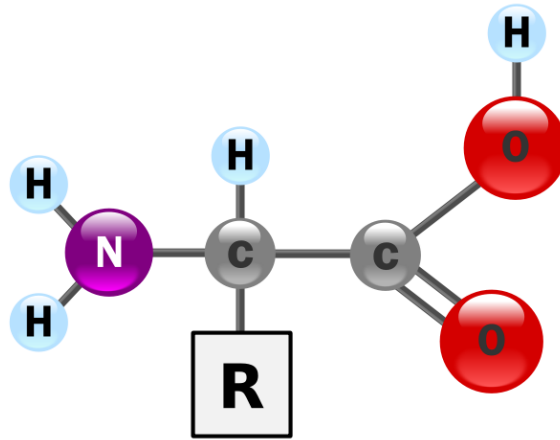
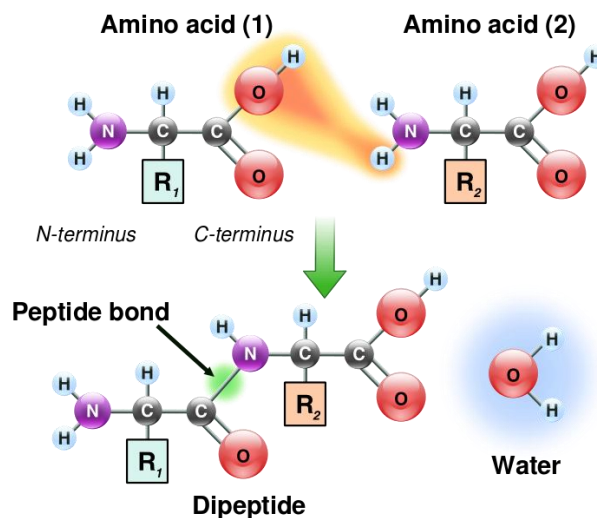


Figure 0.5 structure of amino acid

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



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

Module 006: STORAGE OF BIOLOGICAL SEQUENCE INFORMATION

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

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

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

- **SOLUTIONS DATABASES**

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

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

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

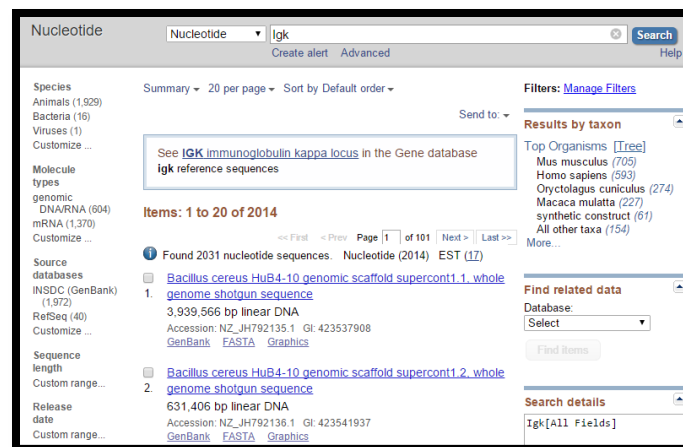
Module 007: USING GENBANK

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

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

www.ncbi.nlm.nih.gov/genbank

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



```

DDDYIEALEYGMPTTGGIGIGIDRLVMLLTNAPSIRDVLLFPAMRHKQD"
repeat region 631283..631406
ORIGIN
  1 catattatat ggtaaagtc tcgatctatt agtattcgtc agctccacat gtcaccatgc
 61 ttccacctcg aacctatcaa cctgatcatc ttccagggat cttactagct tacgctatgg
121 gaaatctcat cttgaggggg gcttcatgct tagatgcttt cagcacttat ccttccgca
181 catagctacc cagctatgcc cttggcagaa caactgggtac accagcgggtg cgtccatccc
241 ggtcctctcg tactaaggac agtcctcttc aaatttccta cggccacgac ggaataggac
301 cgaactgtct cagcagcttc tgaaccagc tcgctgaccg ctttaatggg cgaacagccc
361 aacccttggg accgactaca gcccaggat gcgatgagcc gacatcgagg tgccaaacct
421 ccccgctgat gtggactctt gggggagata agcctgttat cccggggta gcttttatcc
481 gttgagcgat ggcccttcca tgcggaacca ccggatcact aagcccgaat ttcgtccctg
541 ctgacttgtt aggtctcgca gtcaagctcc cttatgcctt tgcaacttac gaatgatttc
601 caaccattct gagggaaact ttgggcgcct ccgttacact ttaggaggcg accgcccag
661 tcaaaactgcc cactgacac tgtctcccgg gtcgataaga ccgtaggtt agaatttcaa
721 tacagtcagg gcggtatccc accagcgctt ccaccgaagc tagcgctccg gtttcaatgg
781 ctcccgccta tcctgtacaa actgtaccaa aattcaatat caggctacag taaagctcca
841 cggggtcttt ccgtcctgtc gcgggtaacc tgcacttcca caggacttat aatttcaccg
901 agtctctggt tgagacagtg cccaaatcgt tacacctttc gtgagggtcg gaacttacc
961 gacaaggaat ttgctacct taggaccgtt atagttacgg ccgcccgtta ctggggcttc
1021 agttcagagc ttcgcttacg ctaacccttc tccttaacct tccagcaccg ggcagggtgc

```

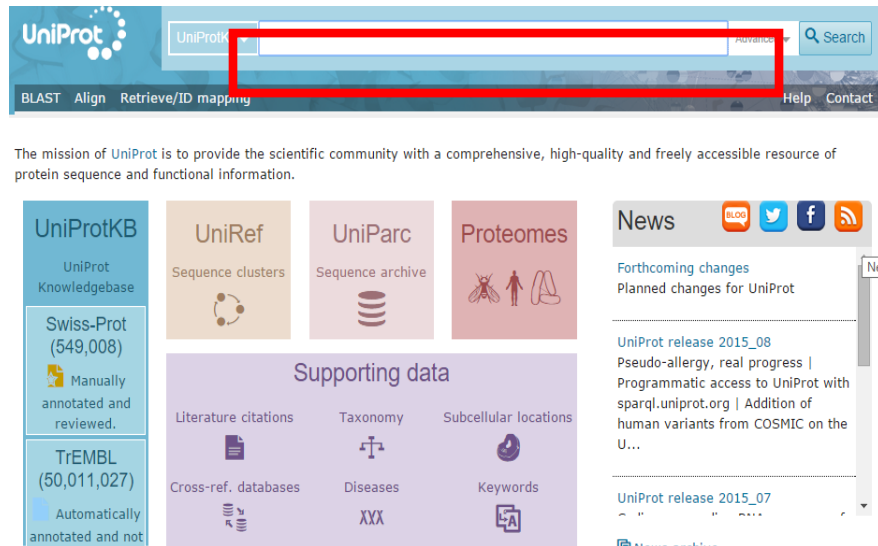
Sequences can be searched from GenBank by typing;

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

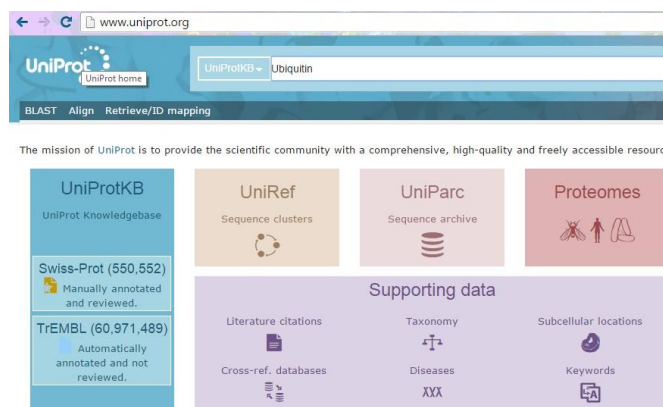
Module 008: USING UNIPROT

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

www.Uniprot.org



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.

UniProtKB results

About UniProtKB

Basket

Filter by

Reviewed (15,494)
Swiss-ProtUnreviewed (270,177)
TrEMBL

Popular organisms

Human (5,428)

Mouse (3,888)

A. thaliana (2,173)

Rat (1,999)

Zebrafish (1,554)

Other organisms

 Go

Search terms

BLAST Align Download Add to basket Columns

1 to 25 of 285,671 Show 25

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

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

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

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

Module 009: COMPARING SEQUENCES

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

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

As there nucleotides can be similar or differ from each other

Figure 0.6 BLAST is used to compare the nucleotides sequences

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

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

Module 010: SIMILARITIES & DIFFERENCES IN SEQUENCES

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

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

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

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

- **CONCLUSION**

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

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

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

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

Module 011: SIMILARITIES & DIFFERENCES IN SEQUENCES

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

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

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

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

• CONCLUSION

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

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

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

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

Module 012: PAIR WISE ALIGNMENT –II

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

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

There are two types of pair alignments.

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

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

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

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

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

Module 013: PAIR WISE SEQUENCE ALIGNMENT –III

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

Which are

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

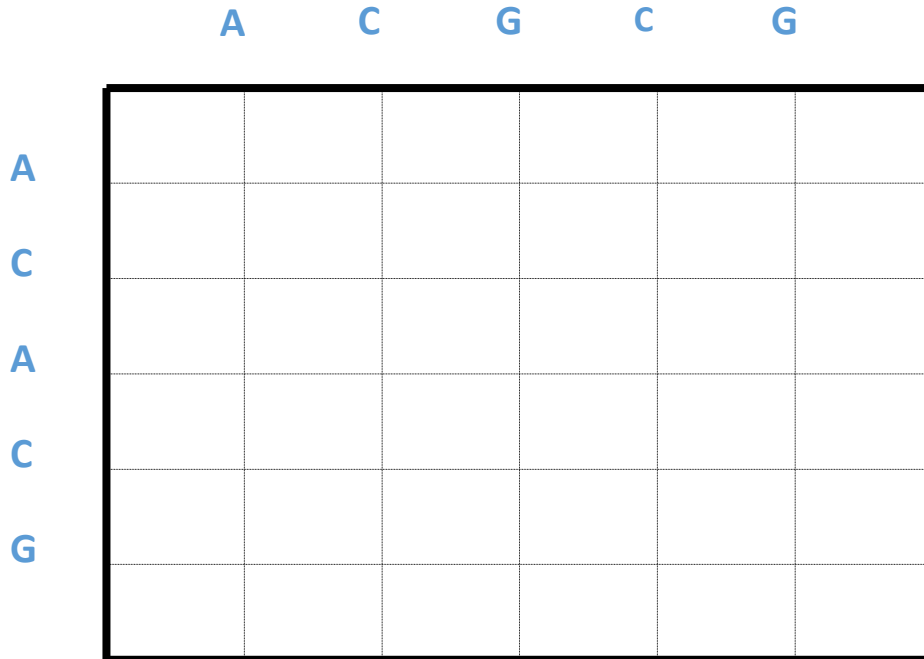
By applying all above ways to any sequence the matching and mismatching can be increased or decreased between to different comparing sequencing.

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

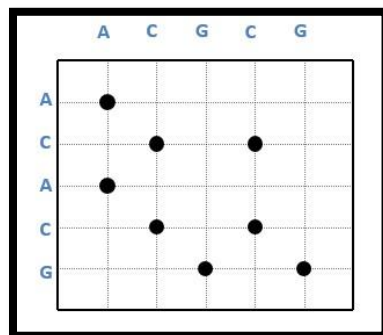
But among above Substitution increases mismatch of sequence.

Module 014: DOT PLOTS

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



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



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

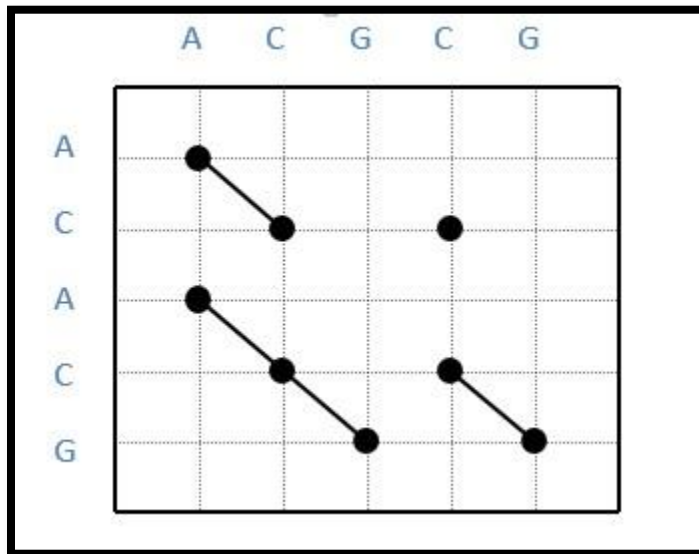


Figure 0.7 dot plot diagonal pattern

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

Module 015: 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.

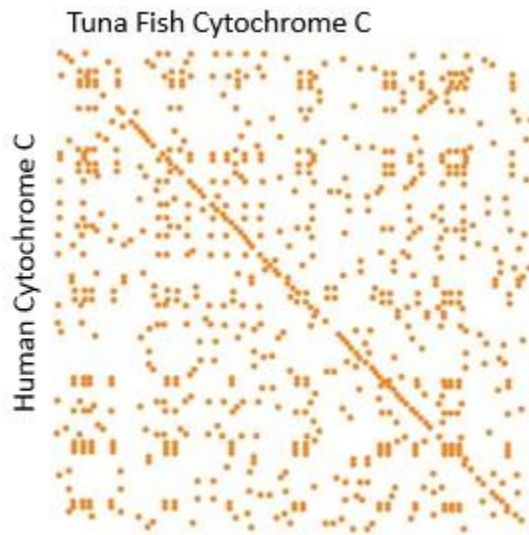


Figure 0.8 tuna fish vs Human

- **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.

Module 016: 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.

Module 017: INTRODUCTION TO ALIGNMENT APPROACHES

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

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

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

```

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

Figure 0.9 local alignment

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

```

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

Figure 0.10 Global alignment

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

Module 018: WHY LOCAL ALIGNMENT

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

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

- **DOMAIN SHUFFLING**

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

- **ADVENTAGES**

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

- **CONCLUSION**

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

Module 019: ALIGNING, INSERTION & DELETION

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

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

- **ALIGNING INSERTION**

For example we have following two sequences

1: A C T G A C T G

1: A C T G A C T G

2: A C G A C T G

2: A C G A C T G

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

1: A C T G A C T G

2: A C . G A C T

Module 020: ALIGNING MUTATION IN SEQUENCES

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

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

1: A C T G A C T

2: A C G G A C T

1: A C T G A C

2: A C G G A C

- **CONCLUSION**

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

Module 021: INTRODUCTION TO DYNAMIC PROGRAMMING

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

To deal with this situation we modify the dot plot.

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

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

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

Module 022: DYNAMIC PROGRAMMING ESSENTIALS

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

In algorithm we calculate the step involve in sequence compression for example if we if we compare two sequences of length “n” than it would be “n²”

And its order is **O (n²)**

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

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

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

Module 023: DYNAMIC PROGRAMMING METHODOLOGY

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

For example

Match = +a

Mismatch = -b

Gap = -c

Score = #match + #Mismatch + #Gaps

Match Rewards = 10

Mismatch penalty = 2

Gap penalty = 5

C	T	G	T	C	G	-	C	T	G	C
-	T	G	C	-	C	G	-	T	G	-

-5	10	10	-2	-5	-2	-5	-5	10	10	-5
----	----	----	----	----	----	----	----	----	----	----

Total = 11

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

Module 024: NEEDLEMAN WUNSCH ALGORITHM-I

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

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

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

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

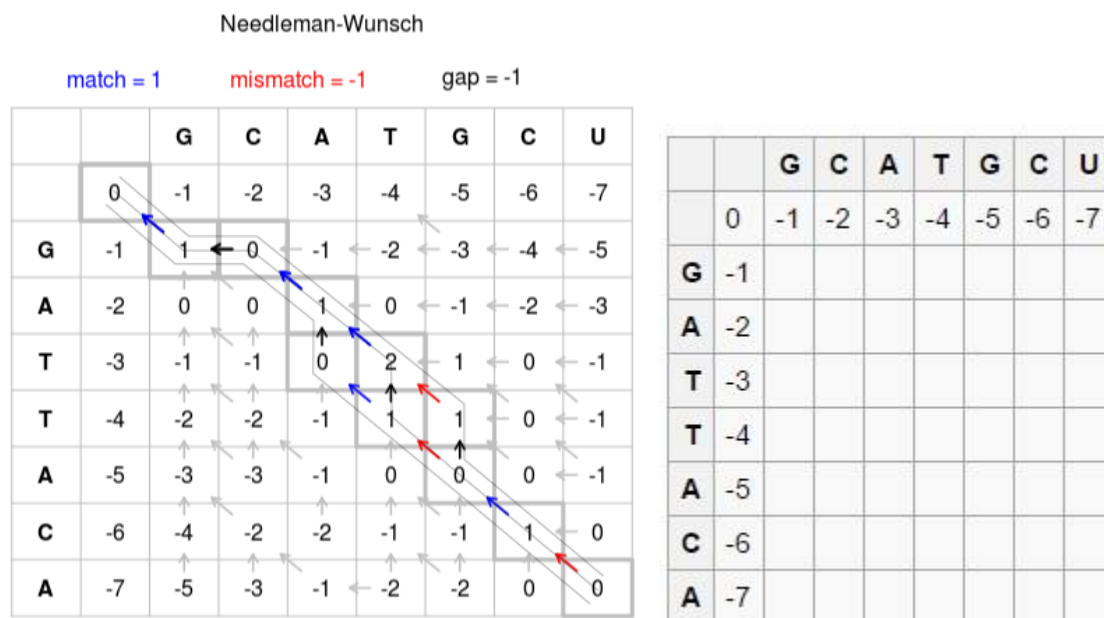


Figure 11 Needleman-Wunsch pairwise sequence alignment

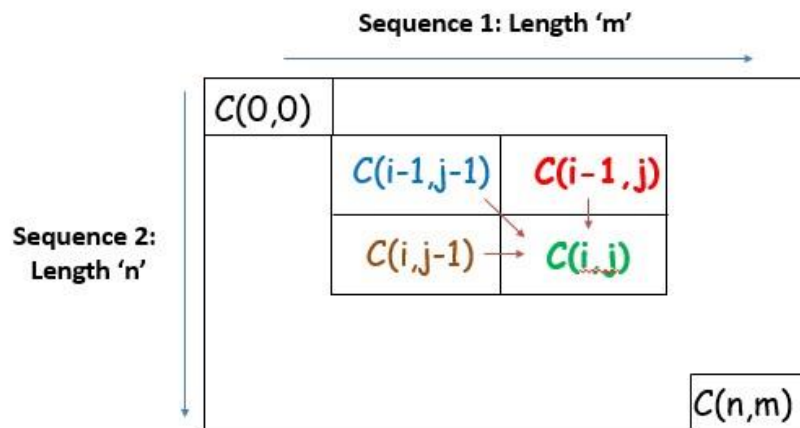


Figure 12 Needleman Wunsch algorithm way of computation of nucleotides

Module 025: NEEDLEMAN WUNSCH ALGORITHM-II

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

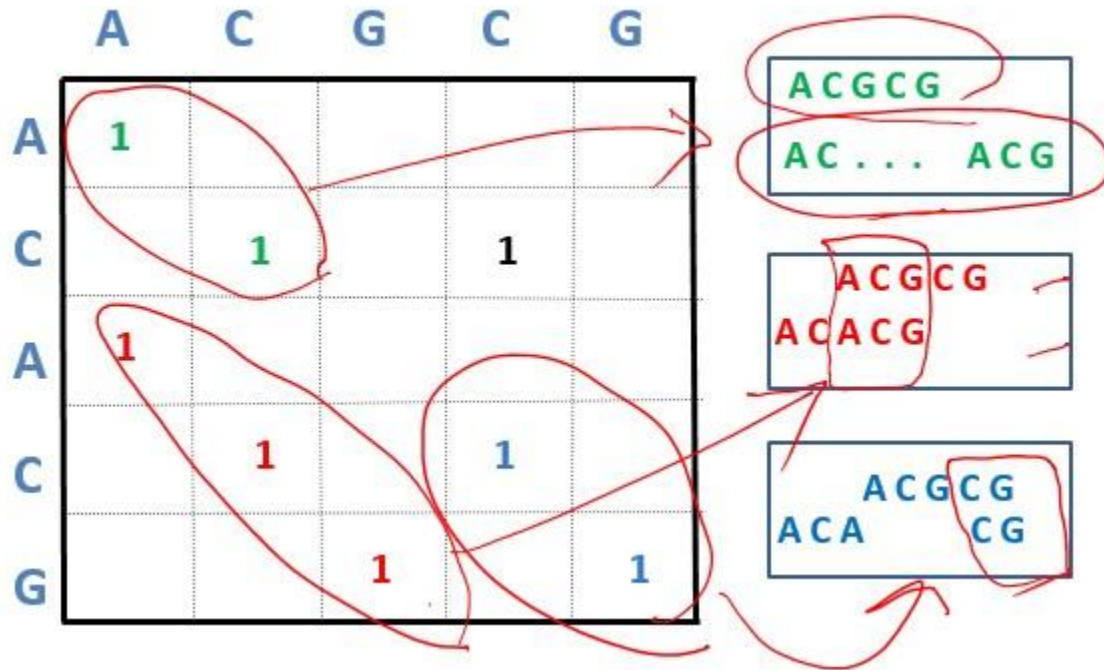


Figure 13 various combinations of sequences through dot plot

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

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

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

		Index: $j = 0 : m$				
		A	C	G	C	
Index: $i = 0 : n$	A	0	$-\gamma$	-2γ	-3γ	-4γ
	C	$-\gamma$	?			
	A	-2γ				
	C	-3γ				

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.

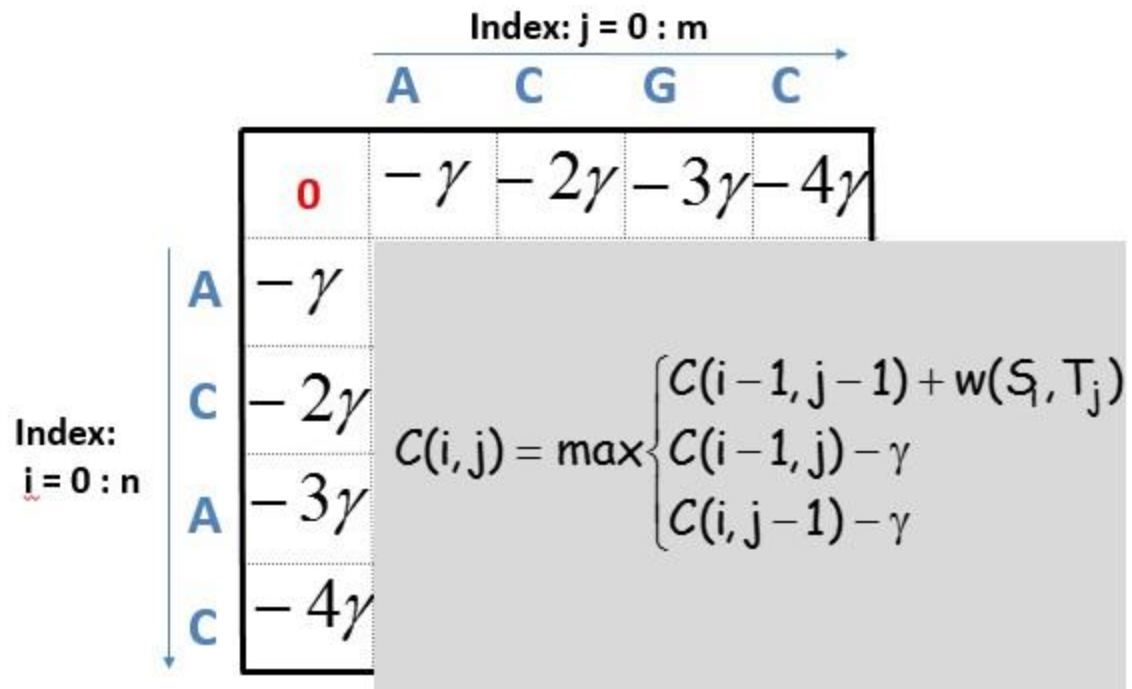


Figure 15 maximum score element is selected from all three sides comparison

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

Module 026: NEEDLEMAN WUNSCH ALGORITHM-III

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

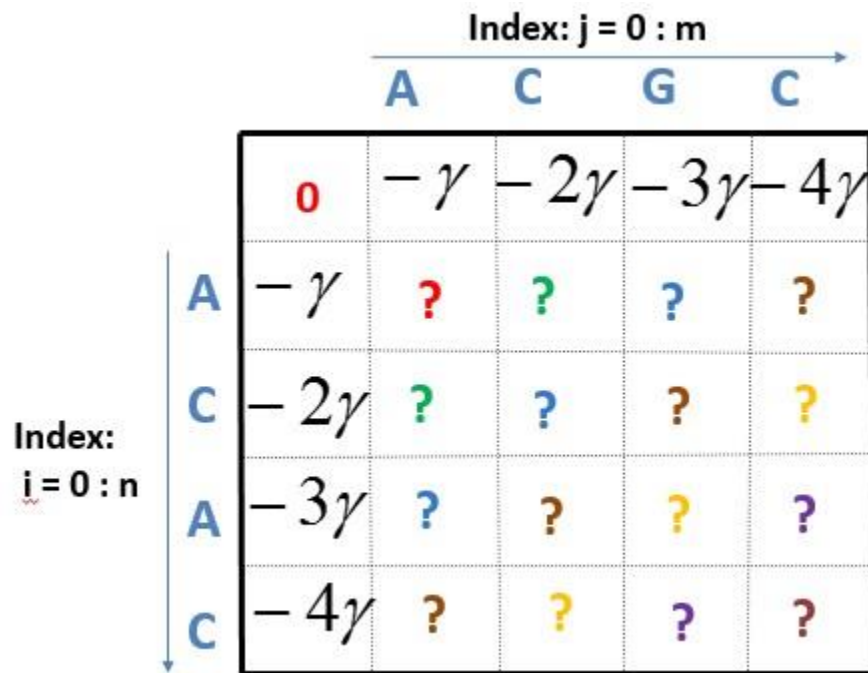


Figure 16 filling order of sequence alignment

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

Module 027: NEEDLEMAN WUNSCH ALGORITHM EXAMPLE

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

For example:

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

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

Figure 17 score computation from all sides

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

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

Figure 18 the best score is computed in diagonal way

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

Module 028: BACKTRACKING ALIGNMENTS

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

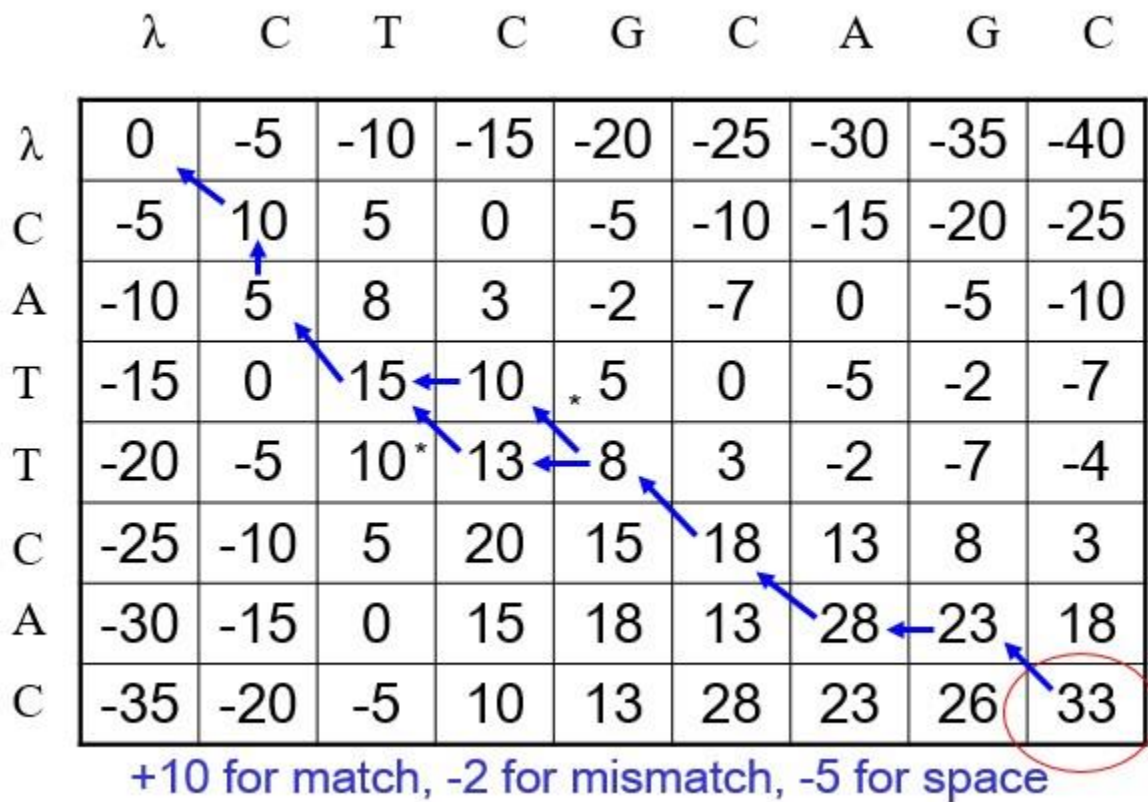


Figure 19 traceback method starts from end maximum score.

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

Module 029: REVISITING LOCAL AND GLOBAL ALIGNMENTS

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

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

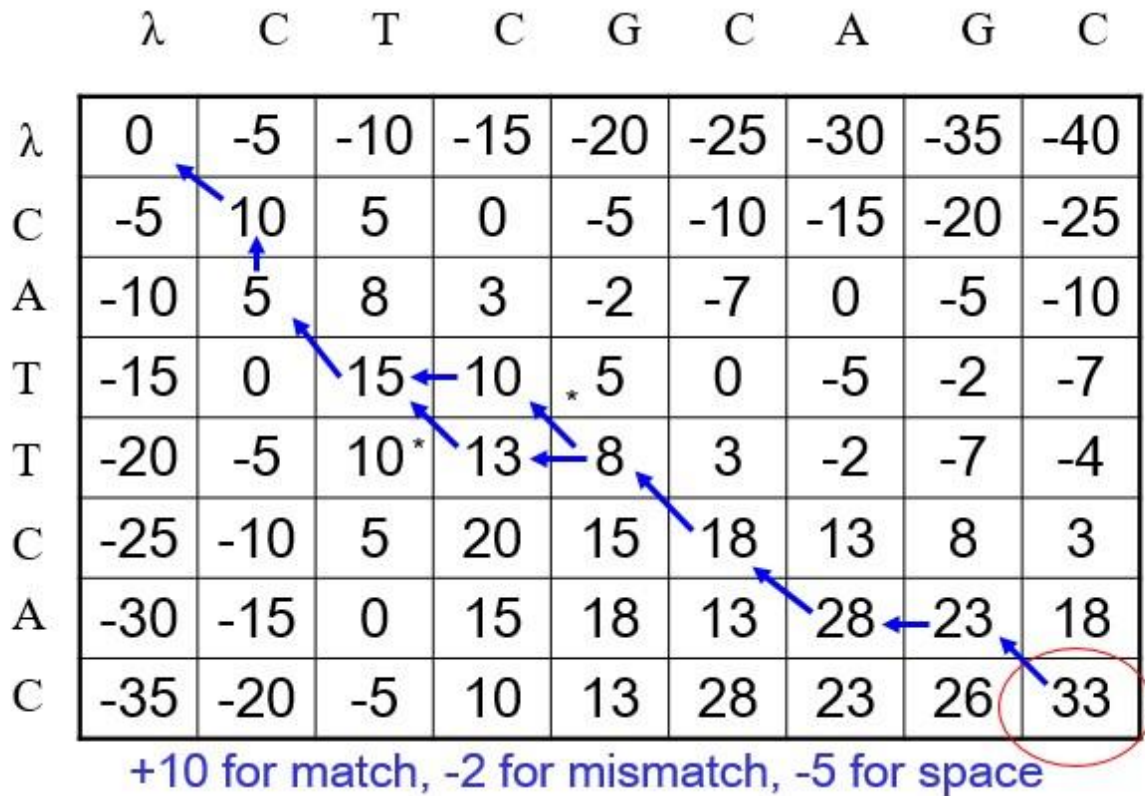


Figure 20 Traceback in Needleman Algorithm

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

Module 030: OVERLAP MATCHES

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

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

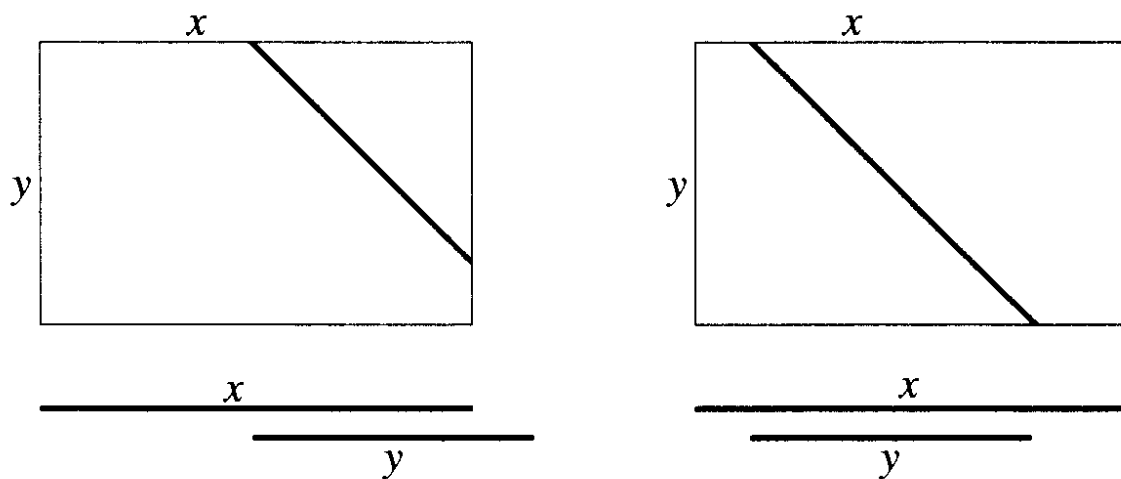


Figure 21 leading and trailing edge mismatches versus global alignment by gap-insertion (stretching) of sequences

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

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

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

Figure 22 Traceback method

Module 031: EXAMPLE

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

In the following example of amino acid alignment we can understand the ways of traceback.

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

Figure 23 Traceback in amino acid sequence alignment

Scoring stage is:

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

Sequences are:

GAWGHEE

PAW-HEA

Module 032: MOVING FROM GLOBAL TO LOCAL ALIGNMENT

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

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

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

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

(Zhang 1998)

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

Module 033: SMITH WATERMAN ALGORITHM

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

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

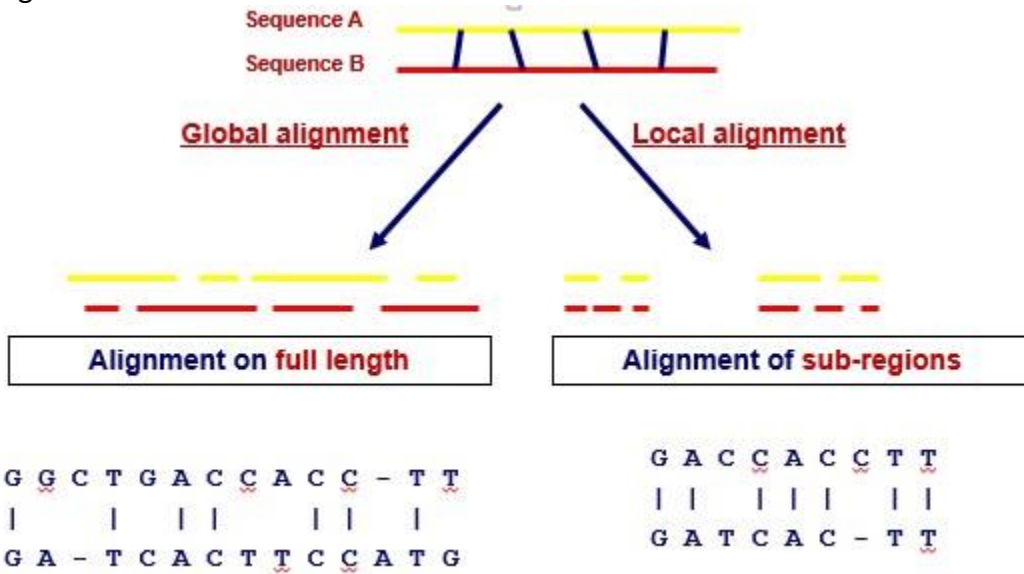


Figure 24 Global and local sequence alignment comparison.

The Smith Freshman algorithm is different from Needle man.

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

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

Module 034: EXAMPLE OF SMITH WATERMAN ALGORITHM

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

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

And in the matrix we place top line of zero and first Column 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 Column are filled with zero in Smith Freshman Algorithm

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

Module 035: REPEATED ALIGNMENTS

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

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

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

Standard Smith Waterman Algorithm

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

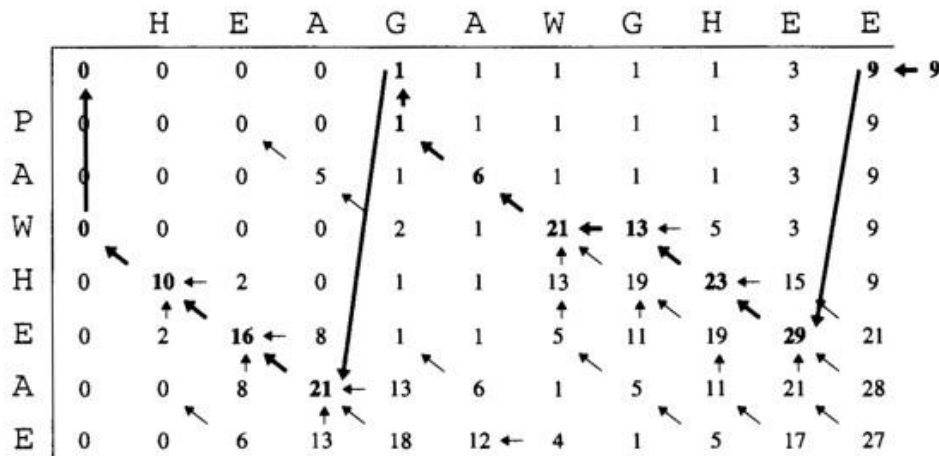
+1 for a match, -1 for a mismatch, -5 for a space

Figure 27 (-5) is threshold score in table

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

Module 036: EXAMPLES OF REPEATED ALIGNMENTS

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



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

HEAGAWGHEE

HEA.AW-HE.

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

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

Module 037: INTRODUCTION TO SCORING ALIGNMENTS

There are two types of alignments;

- Optimal Alignments
- Best Alignment

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

- Appropriately rewarded for matches and mismatches.

• INTRODUCTION

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

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

• CONCLUSION

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

Module 038: MEASURING ALIGNMENTS SCORES

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

For example:

Needleman Wunsch Algorithm

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

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

Figure 29 Needleman Wunsch algorithm match, mismatch scoring

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

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

Module 039: SCORING MATRICES

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

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

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

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

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

Figure 0.12 Ubiquitin Protein where amino acids matching

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

Module 040: DERIVING SCORING MATRICES

Each amino acid have different property.

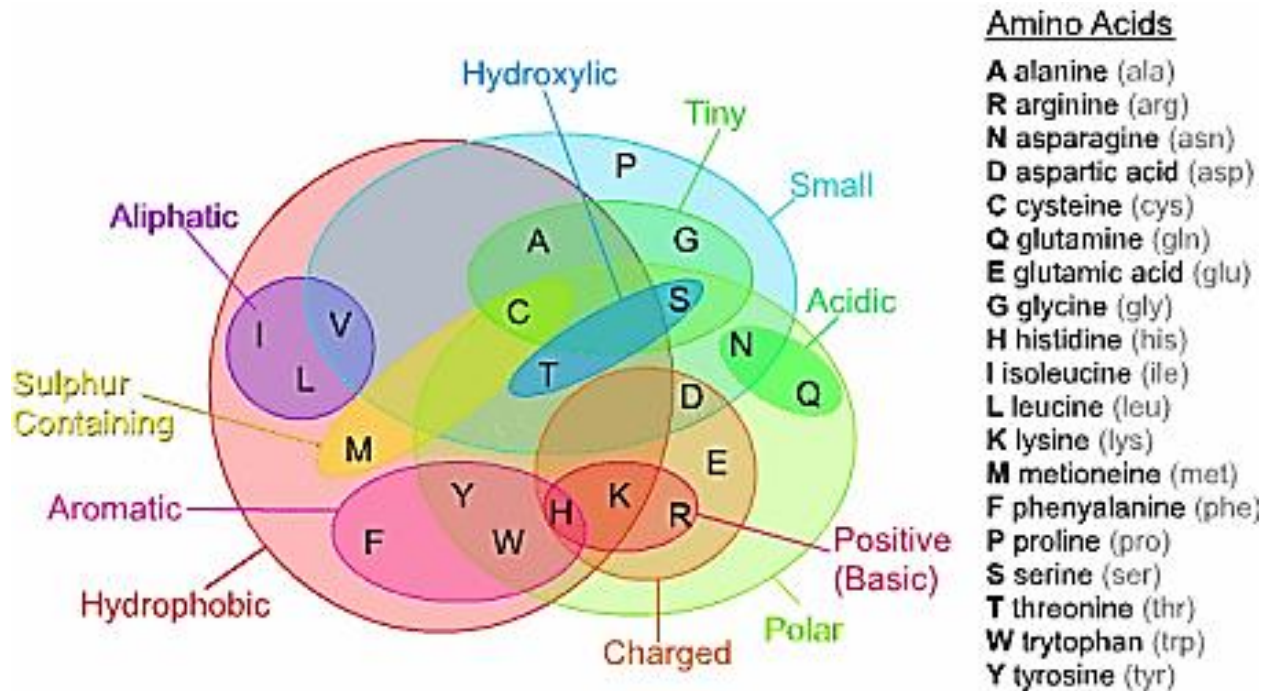


Figure 0.13 properties of amino acids (Image Esquivel et al. (2013))

And each amino acid have different frequency.

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

Amino acid	Frequency	Amino acid	Frequency	Amino acid	Frequency	Amino acid	Frequency
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		<i>aligned pairs</i>	<i>observed frequencies f_{ij}</i>
A	A	C	A		A to A	28/60
A	A	C	C		A to B	8 /60
A	A	B	A		A to C	6/60
A	A	B	C		B to B	6/60
A	A	B	B		B to C	10/60
					C to C	2/60



<i>amino acids</i>	<i>observed frequencies f_i</i>
A	14/24
B	6/24
C	4/24

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

Module 041: PAM MATRICES

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

There are two types of scoring matrices.

- PAM
- BLOSUM

PAM means “Point Accepted Mutation”

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

- **PAM UNIT**

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

- **STEP TO COMPUTE PAM MATRICES**

1. Align the protein sequence which are 1-PAM Unit diverge.
2. Let A_{ij} be the number of times A_i is substituted by A_j .
3. Compute the frequency f_i of amino acid A_i .

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

$PAM 'n' = (PAM1)^n$

Module 042: BLOSUM MATRICES

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

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

```

WWYIRCASILRKIYIYGPGVGSRLRT
WHYVRCASILRHL YHRSPAGVGSITK
WFYTRAASTARHLYLRGGAGVGSMTK
WWYVRAAALLRRVYIDGPGVNSLRT
  
```

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_{ij} of aligned pair A_i to A_j . Hence, f_{ij} 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

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

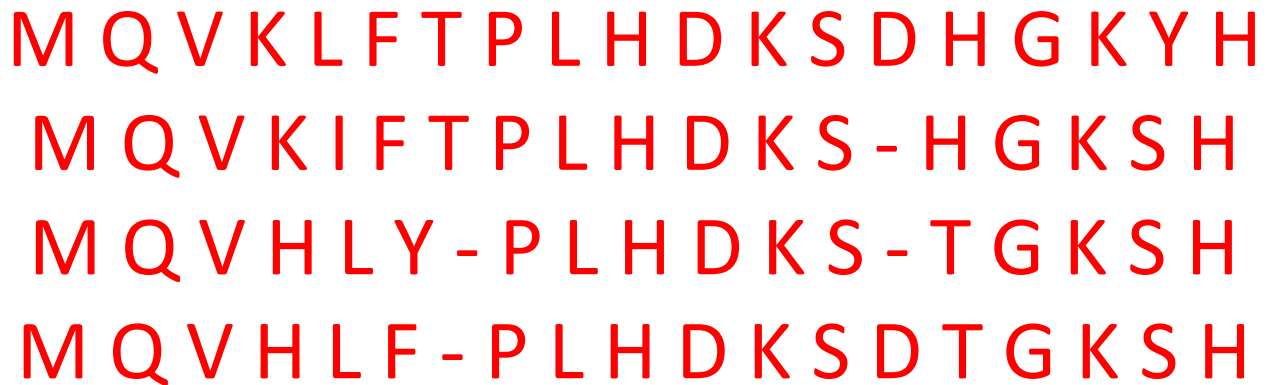
Figure 0.15 formula for computation of BLOSUM MATRICES.

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

Module 043: MULTIPLE SEQUENCE ALIGNMENT

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

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



```
MQVKLFTPLHDKSDHGKYH
MQVKIFTPLHDKS-HGKSH
MQVHLY-PLHDKS-TGKSH
MQVHLF-PLHDKSDTGKSH
```

Figure 0.16 multiple sequence alignment

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

Module 044: MORE ON MULTIPLE SEQUENCE ALIGNMENT

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

APPLICATION OF MSA

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

METHODOLOGY

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

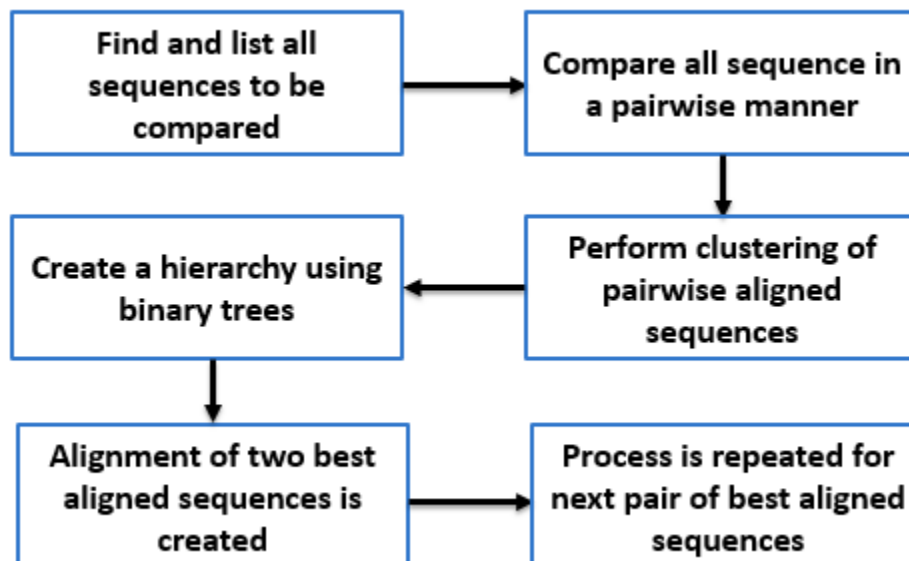


Figure 0.17 Methodology