

Bioinformatics
Prof. Michael Gromiha
Department of Biotechnology
Indian Institute of Science Education and Research

Lecture – 31a

Overview I

In this lecture, I will provide an overview on different aspects of bioinformatics we learnt in this course, specifically on various algorithms or databases as well as the different applications of bioinformatics. So, we started with the fundamentals of bioinformatics right. So, what are various fields of science which contribute to bioinformatics?

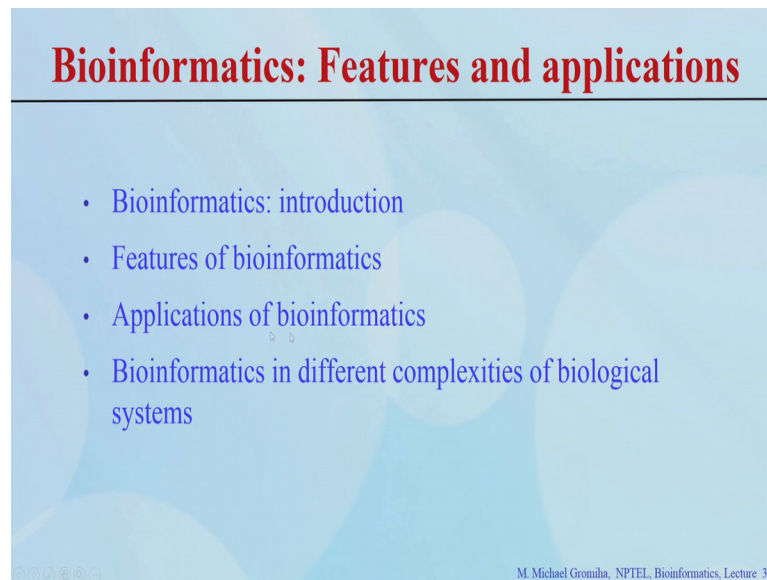
Student: Information.

Yeah, mainly the biological sciences and biological sciences provide wealth of experimental data because data is very important. So, experiments provide lot of data for DNA databases and then sequences, structures, pathways networks, various information we get from the biology. So, that information is very important, information technology plays a major role as well as other disciplines, like physics, chemistry and mathematics, statistics we discussed about all the aspects how the different fields of science contribute to interdisciplinary area. What are the various features of bioinformatics we discussed?

Student: (Refer Time: 01:18).

Development of databases and algorithms, we can derive some hypothesis for example, we have a data. I will try how we generate this data and what are the factors which influence to get that specific data. So, we can give some hypothesis and based on the hypothesis we derive some important features and if you get some properties then we can relate. We can try to develop some models prediction algorithms and they can use it for the validation then also we discussed about the structure based design how the different aspects of this bioinformatics tools and databases are useful to identify some lead compounds for different targets.

(Refer Slide Time: 02:03)



Bioinformatics: Features and applications

- Bioinformatics: introduction
- Features of bioinformatics
- Applications of bioinformatics
- Bioinformatics in different complexities of biological systems

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

Now, we discuss about different applications of bioinformatics and then how the bioinformatics contribute and different complexities of biological systems.

We start with the DNA sequence and then DNA structure and then protein sequence, protein structure, protein interactions and then this interact networks and the pathways then the organism the organ, tissue, up to the organism. So, various aspects we can say in each category we seen the applications of bioinformatics either in the databases or they models or servers rights in different ways, understanding this complexities of this biological systems. Then we discuss each aspect separately mainly if we focused on the protein side as well as some aspects of a DNA right.

(Refer Slide Time: 02:56)

DNA Sequence Analysis

- DNA and RNA
- Complementary strand
- Transcription and translation (software)
- Reading frames
- Sequence based parameters

Handwritten notes in red ink:

- Arrows pointing to 'DNA and RNA' and 'Complementary strand'.
- Sequence: 5' ATAGC 3'
- Software: EMBOSS
- Sequence: 5' GCTAT 3'

M. Michael Groniha, NPTEL, Bioinformatics, Lecture 31

If you discuss about the DNA, what is the difference between DNA and RNA.

Student: Two (Refer Time: 03:01).

Right, one is H and OH, the sugar and the base.

Student: (Refer Time: 03:08).

So, what is the based DNA?

Student: Thymine

Thymine

Student: Thymine

And the RNA.

Student: Uracil

Uracil, you can see the difference between the DNA and RNA. So, when we discuss the complimentary stands we get the this is the, ATAGC, if it is 5 prime - 3 prime, then what is the complementary strand in the 5 prime - 3 prime.

Student: GC.

Right G.

Student: C.

C.

Student: T.

T.

AT right, first you get this strand this in 3 frame 5 frame then we change the inverse right. So, you change the direction when you change this sequence. Then we discussed about the transcription translation then also the degeneracy of this, from the codon to the amino acids and what are the software we discussed on the translation as well as this complementary strand.

Student: EMBOSS.

EMBOSS we discussed about EMBOSS. So, what is EMBOSS?

Student: (Refer Time: 04:09).

European Molecular Biology Open Software Source.

Student: open software source.

Open Software Source right, this contains various packages. So, we can use this package to get the various parameters of this DNA. Then we discussed about this reading frames. What are the different reading frames.

Student: 3 for the (Refer Time: 04:30).

Right 3 for the forward and 3 for the backward, so 6 reading frames. Then what they various sequence based parameters we discussed.

Student: (Refer Time: 04:39).

We can see the flexibility and the rigidity and the base for preferences, compositions various other aspects based on this DNA sequence. Look at this DNA sequences also important to see this parameters to understand the various functions of this DNAs and

this how the interact to this proteins. Then we moved on to the different types of databases. So, when you started data, what is a database?

Student: Data collection of.

It is a collection of data in the computer readable form because we have several information available in the literature and various aspects, various biological data. So, they publish the literature it is very important to collect and put in a proper database right.

(Refer Slide Time: 05:26)

Databases

- Databases
- Characteristics
- Nucleotide sequence databases
- Nucleotide property database
- Literature database

Handwritten notes on the slide:

- DDBT, EMBL, GenBank
- AA, AT, AC, AC
- PURMED
- 1. Contents
- 2. Ontology
- 3. Schema
- 4. Format
- 5. Data retrieval
- 6. Links

Flowchart:

```
graph TD; MIN[MIN] --> Search[Search] --> Display[Display]
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

So, what are the various characteristics of databases we discussed.

Student: Reliability in my data.

Right first we discussed about what are the various, we need the contents then we need the ontology. What is ontology? So, because when we develop database there will be several technical terms which we use in the database. So, we need to give the complete details of this in database then we then what are other characteristics?

Student: Schema.

Schema, what is schema?

Student: Data

And how to put your data what is the major aspect if you have the this is your main information then supplemented with other information for example, what are the direct information related with this main data. For example, if you take the thermo dynamic data we need the protein name and the source and the experimental conditions and all. Then we can give more additional information regarding a sequence and structure and all other things. Then we give the search option, then we gave to give the display how the output your data. So, you have to make the organization organization of your database how to do that. Then what are the other; what are the next one you have the specific format. What is format ?

Student: (Refer Time: 06:54).

Right they have the same order. So, we have this order what are the terms you use in your database. So, it should be same in all the entries then easier to fetch the data as well as for the user should be easier. For example, if you take the protein data bank started with the header compounds finally, it go some to sequence residue information then secondary structure: helix strand then stop with the atoms right. So, if you are interested in the 3D coordinates then you go with atom records right. So, if you see or any PDB id.

So, they follow the same format. So, we have to use some, specific format and then the data retrieval right. How to retrieve the data? So, you should provide proper search options. Then it is we need to think about what are the various options the users like to have or if you use the thermodynamic data they are interested in the free energy data. Likewise what are the information the users prefer accordingly you design your web server sites so that they can easily obtain the data then this some search options then as soon as they search they should get the reliable data it is very important to get this reliable information and the information what they require that is very important.

So, in this case you have to make the interface such a way that users can retrieve the data second aspect is that should be available all the time, that should be fast enough to show the data right. So, otherwise if this not properly maintain or it is not available all the time then the users they do not use it again we have to be very careful. Then you have to prove proper links to the external databases because when you develop a database you get the information from the different resources.

On the other hand there will be several other databases which are related to the your specific database and in this case you have to provide the links to all the relevant databases in this case you said you can get the information if they need more data. For external you take the separate data. So, you get the data from the literature provide the literature information and this case they can expand the knowledge to read the papers. So, like the database. So, you have several characteristic of features and we discussed about the various databases. What database for the nucleotide sequence databases?

Student: Genbank

I will see the DDBJ, EMBL, Genbank right. So, they have the different information similar information.

Student: (Refer Time: 09:34).

Similar information they also follows a similar format also. So, they same time they developed different places, then they form the network. So, in this case they can share the data. So, in this you can deposit data at one place and can be shared by others. In this case you can have the uniform data when all these databases. Then we discussed about different properties like nucleic acid property database which contains data for the different base pairs for example, the dinucleotides for example, AA, AT, AC, AG and so on like the flexibility or the rigidity and the stability melting temperature the stacking energy hydrogen bonds.

So, we give all the information in this database. So, you can use these information available in database to see if you have any sequence what is the average flexibility or average rigidity because sometime with the nuclei they DNA flexible enough to interact with the protein and you can see if you use this data from this nucleotide property database. In the literature database right, so, what is the major literature database for the biology.

Student: Pubmed.

Pubmed right. So, there are various databases like physical abstract, chemical abstracts, biological abstracts are available, but currently for the biologist and the medicine medical field. So, PUBMED is very widely used because it includes data from various

literature for, various journals, they cover various journals and wide variety. So, you can get sufficient information from the pubmed database. In addition you can see the scopus or the science citation index right about science that also you can use as literature databases. So, then we move to the protein side right. So, what is the building block of proteins?

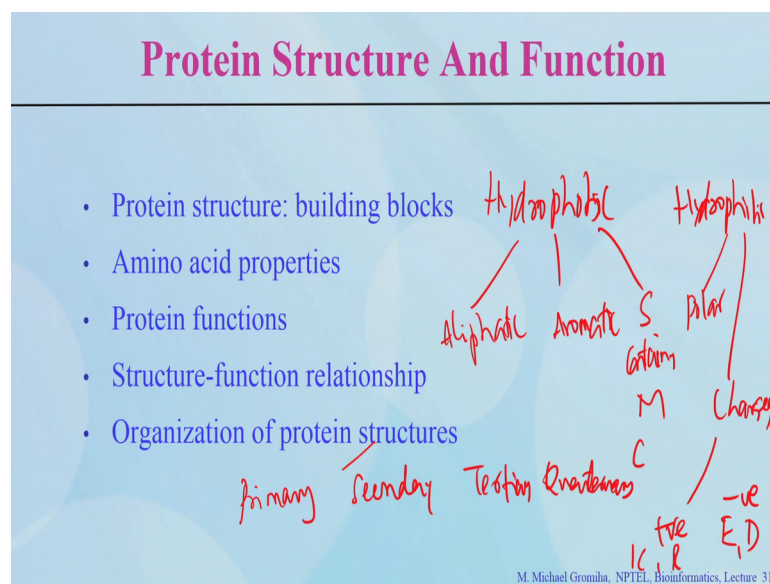
Student: Amino acids.

Amino acids right. So, how many different amino acid residues ?

Student: 20.

20 depending upon the characteristic features, how you classify these amino acids?

(Refer Slide Time: 11:36)



Student: Hydrophilic, hydrophobic.

Hydrophilic hydrophobic they have the hydrophobic and hydrophilic, what are the subclasses.

Student: (Refer Time: 11:45).

If we take the hydrophobic,

Student: (Refer Time: 11:48).

Aliphatic, aromatic and sulphur containing amino acids. What are the sulphur containing amino acids?

Student: Cysteine, methionine.

Methionine and cysteine. So, we discussed cysteine. Cysteine can form this disulphide bonds. This is also important for the hydrophobic interactions. Then the hydrophilic: it is polar and charged. In charged we have positive and negative. What are the negative charged amino acid?

Student: Aspartic acid.

Aspartic acid and glutamic acid; Glutamic acid and aspartic acid- positive charged: lysine and arginine. you can see the different types because these amino acid residues they form specific types of interactions right, so that they can form the compact global shape and maintaining the stability of particular product. So, then we discussed about the protein functions. What are the various functions?

Student: Enzymes.

Enzymes.

Student: cell signaling (Refer Time: 12:50).

And cell signaling.

Student: Antibodies.

Antibodies.

Student: Transport.

The transport proteins.

Student: (Refer Time: 12:56).

Right.

Student: Structural

Structural aspects, blood clotting proteins, antibodies right. So, they can perform various functions. We discussed about various types of proteins: globular protein, membrane protein the fibrous proteins each one of them are responsible for different types of functions right. So, then we want to understand the structure function relationship the function depends mainly on the structures. So, it is necessary to understand about the functions. So, what are the different types of protein structures?

Student: primary.

Like primary.

Student: Secondary.

Secondary.

Student: Tertiary.

Tertiary.

Student: Quaternary

And quaternary. So, what is the primary structure? These are amino acid sequence. So, here they will give amino sequence and the covalent bonds, but we do not know about the other locations of amino acids and all; And secondary structure?

Student: (Refer Time: 13:50).

So, regular arrangement of these amino acid residues this mainly the alpha helix and beta strand, tertiary structure?

Student: 3 range.

It will provide the exactly the coordinates. Where the location of all the atoms in a residue as well as in the complete protein. What is quaternary structure?

Student: subunits.

different subunits. Protein has different subunits they form these quaternary structure right. So, in between that there are two different structural aspects you can see the super

secondary structures plus the domains. Then we see the different types of databases for protein sequence, protein structure and protein function. If we take the protein sequence database. So, what is the initially developed protein sequence database?

Student: PIR.

PIR they developed Protein Information Resource and then they developed this swissprot so this PIR and swissprot they merge together to form this UniProt.

(Refer Slide Time: 14:38)

The slide is titled "Protein Sequence Databases" in a bold, pink font. Below the title is a horizontal line. A bulleted list in blue text contains the following items: "Protein Information Resource", "Uniprot" (underlined), "Features", "Obtaining data from UniProt", and "Applications". To the right of the list, there are handwritten notes in red ink: "+ Swiss prot" next to the first bullet, and "A, V, L" and "C, W" stacked vertically next to the "Features" bullet. At the bottom right of the slide, there is a small footer in blue text that reads "M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31".

- Protein Information Resource + Swiss prot
- Uniprot
- Features A, V, L
C, W
- Obtaining data from UniProt
- Applications

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

UniProt has manually curated sequences plus computer annotated sequences. Currently how many data in the uniprot database? Around 80-85 million sequences and what is the unique features of uniprot?

Student: (Refer Time: 15:00).

Right, less redundancy.

Student: Highly integrated.

Highly integrated.

Student: (Refer Time: 15:08).

And links with the other databases right the integration with other databases. So, this is reason why Uniprot is widely used. Also what are the current information which are available in uniprot database?

Student: Functions.

You can see the functional database, data and you can see the pathways and interactions If we have a protein. So, we get almost all the information from this uniprot database. So, how to obtain the data from uniprot?

Student: Search with.

You can search you can search with the uniprot name or you can search the MESH terms and also you can get the data with specific redundancy. you can what are the various cut-off available in the uniprot?

Student: 100.

90 percent, 100 percent and 50 percent you can search and then you can get these information. We can get the data for any specific protein as well as the group of proteins you can use you can use different MESH terms you can obtain the data. Now, what are the applications of uniprot.

Student: You can get the sequence protein.

Yeah you can say get the sequence for any aspects for whatever the classes or the information structure if it is the function if you want to understand these specific features you can get obtained from this uniprot. So, also uniprot provide the statistics of the different proteins what in the average length of the protein in the for the sequence available in database.

Student: 300.

300, 314-315 protein sequence residues and generally they are only about 100 to 300 residues and some proteins are long length. So, average you can say about 315 residues. Then also they give the statistics for the different amino acid residues which residue is highly occurring amino acid residues, alanine, valine, leucine these are having about 10

percent, then few residues are rarely occurring amino acids. What are rarely occurring amino acids? Cysteine.

Cysteine.

Student: tryptophan

Tryptophan. These are rarely occurring amino acids in the uniprot database. So, then we have this sequences you do not want to compare and how far two sequences are similar? how they are distant from the sequences?

So, you try to understand this sequence similarity we develop algorithms for the align aligning sequences.

(Refer Slide Time: 17:24)

Sequence Alignment

- Pairwise alignment
- Gap penalties
- Scoring matrices
- Nucleic acids and proteins
- Development of a PAM matrix

Handwritten notes:

A_i

$A \rightarrow V$

$A \rightarrow K$

$A \rightarrow V$

$A \rightarrow K$

$A A A C A$

$A A A C T - -$

Original 1

Gap 2

$[A \rightarrow V]$

$[A \rightarrow K]$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

So, what is a pairwise alignment?

Student: Alignment of two sequences.

Align two sequences right. So, what is various ways to align two sequences. For example, it is a sequence.

Student: It will (Refer Time: 17:40).

Right, you can first you try to use the without gaps then you can align with gaps with gap penalties then we see the alignment based on some scores how to give scores, matching score. What is a matching score?

Student: If it matches.

If it matches with these the alignment for if you put like this, this is this matching. So, this is a match score. So, what is then mismatch score, what is mismatch score?

Student: (Refer Time: 18:10).

Yeah for for example if you take these, is not matching. So, you have to again mismatch score. Then what is gap penalty? If there is a gap then we have to give penalty for this gap. Because compared with this mutations gaps are rarely occurring. In this case you have to give more penalty to the gaps then the mismatch score. What are all different types of gaps we discussed. Which penalties?

Student: Origination (Refer Time: 18:34).

Origination penalty.

Student: Extension.

And the gap penalty. Gap penalty means number of how many number of gaps, this is how many types gap originates this is the how many number of gaps. For example, what is origination penalty here?

Student: 1.

It is 1, because gap originates only ones; what is a gap penalty.

Student: 2.

There is 2 because 1 to 2 gaps like ways that. Then we try to discuss about the scoring matrices. What is scoring matrix?

Student: Mutation.

This will tell you what is a mutation rates actually we can derive the scoring matrix in various aspects. For example, if you take the nucleic acids you can see the purines and the pyrimidines, we can do transversions and transitions based on that we give penalties. If you discuss about the case of amino acids you can do with physiochemical properties and the genetic code that how many variations then finally, we can do it with the actual rates the actual mutation rate. For example, if you take the set of homology sequences say about 80 percent or 90 percent sequences in this case there will be several mutations and see what is the most proper mutations we using this proper mutations you can derive a matrix from one amino acid to another amino acid. So, if you remember you can see there are some cases they are acceptable some cases it is not acceptable.

Mainly if you mutate same residues then no mutations that is highly preferred ones even in that case a depending upon the amino acids for example, compared with tryptophan or the cysteine with the alanine. tryptophan is more conserved than alanine, likewise if you have a similar type of amino acids for example, alanine to valine right. So, this is preferable or not is preferable for example, if alanine to lysine it is not preferable either they physicochemical properties or scoring matrices. Scoring matrix you can develop from the mutation rates of amino acid residues how far each residue is mutated to different types of residues we can make that. So, this is how we made the scoring matrices. What are different matrices we discussed?

Student: PAM and BLOSUM.

PAM matrix and the BLOSUM matrix; what is PAM matrix?

Student: Point accepted mutation.

Point accepted mutation matrix right. So, they how to derive the PAM matrix, what the various aspects we need to consider to derive the PAM matrix?

Student: Frequency.

So, first we see the frequency of the amino acid

Student: (Refer Time: 21:15).

Length of the sequences and how this residue is mutated and which residues mutated to which residue for example, if it is A_{ij} right. So, how many times alanine is mutated to valine, how many times alanine is mutated to all the other residues 20 residues, how many alanine present in the sequence, what is the length of the sequence? So, all these aspects we need to consider when developing this PAM matrix right. Then what is BLOSUM matrix?

Student: (Refer Time: 21:45).

(Refer Slide Time: 21:41)

The slide is titled "Sequence Alignment" in a pink font. Below the title is a bulleted list in blue text:

- BLOSUM matrix
- Dynamic programming
- Pathways for alignment
- Global alignment
- Local alignment

Handwritten in red ink are the names "Needleman" and "Smith-Waterman". A red arrow points from "Needleman" to "Global alignment". A red arrow points from "Smith-Waterman" to "Local alignment". To the right of these names, there is a red bracketed list: "max { match/mismatch, insertion, deletion }". Below this list is a red plus sign and a red circle.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

This blocks substituted matrix right. So, what is the difference between PAM and BLOSUM?

Student: homologous sequences.

Yeah BLOSUM they takes the almost all the alignments to see the mutation rates BLOSUM will take the conserved regions. So, they can say the to avoid the gaps and then see the conserved regions how they can make this alignment. We discussed about dynamic programming, then what is dynamic programming?

Student: divide the problem into

Yeah, divide the problem into small peices finally, solve the problems and then finally, combine together and you solve the whole system right. So, that is called dynamic

programming. So, what are the different aspects would we use dynamic programming for this sequence alignment and two different alignments? global alignment.

Student: Local.

And the local alignment. Because we can have two sequences you can use various ways to align the sequences right, where there is a different pathways right. So, we use two different alignments. So, what is global alignment?

Student: All sequence.

Whole sequence, you can align the whole sequence. So, what is local alignment?

Student: (Refer Time: 22:49)

Right, only the small part of the sequence you can align. How, what is the algorithm used for global alignments?

Student: (Refer Time: 22:58)

Wunsch Needleman alignment; Wunsch Needleman algorithm - What are the conditions used to fill the matrix in Wunsch Needleman algorithm.

Student: Match.

Maximum of.

Student: Match.

This is the match or mismatch, match we have score mismatch we have score.

Student: Then gap

Then gap that is insertion or deletion.

Student: Deletion.

Right. So, we have the rank we have the score for each case we have match score, we have mismatch score, we have insertion gap penalty insertion deletions. So, based on

these we can give the maximum value provided from all these 3 values right. So, what is local alignment, which algorithm is used for local alignments?

Student: Smith-Waterman.

Smith-Waterman right. So, what is the condition for the Smith-Waterman algorithm

Student: Match, mismatch and Zero.

This is plus then a 0. So, in this case there is no negative values in this case you can restrict for the small gaps, any small matches fine. So, how to align these sequences we have two sequences, first we see the various different options if you have local alignment see there 4 values then fill the value numbers and from the back see the highest score you go back to trace back then accordingly you can put the gaps and then you can make this alignments right.

(Refer Slide Time: 24:34)

Sequence Alignment

- BLAST
- Features
- Alignment scores
- Multiple sequence alignment
- Online resources

Handwritten alignment example:

T	A	I	T	L	A	C	A	V
A	T	S	L	A	K			

Handwritten annotations:

- 3/6
- 4/6
- 6/8
- 75%

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

Then what are the various programs we use for sequence alignments? BLAST is widely used tool. What is expansion for blast?

Student: Basic local alignment

Basic Local Alignment

Student: Search tools.

Search tool right. So, what are the various features of the blast?

Student: You can do

Right, you can search for any sequences you can get the all the sequences which are matching with the particular sequence and align the sequences right. So, you can, what are the various features we get. What are the various scores we get from blast?

Student: (Refer Time: 25:06)

Right we can see the p-values we can see the e-values and the similarity.

Student: Max score.

Right maximum score, the identity.

Student: sequence coverage.

Right the identity, these are the score you can get from blast and coverage. From that you can see whether the sequence alignment is perfect or not. Now, for example this is the sequence. So, in this case what is the sequence identity? So, how many identical sequences?

Student: 3 3 position.

If you take this one then we can see the coverage of 6 for 8 right. So, this is the match, here is the match, here is the match. So, 3 right. 3 per 6 you can see and the similarity this is similar T and S are similar. So, you can take this you can get 4 by 6 because coverage is only 6 you can see this is the average is only 6 by 8, 6 by 8 means how many percentage; 75 percentage, yeah 75 percentage and then what is the multiple sequence alignment.

Student: Multiple sequences.

Align more than two sequences right. So, again this case there are various ways to get the multiple sequence alignment right. So, you can first see the similar sequences then go with this dissimilar sequences then put it together and you can see which residues which come in the same position you can get the information for a multiple sequence alignment. What are the online resources available for multiple sequence alignment?

Student: Clustal omega, clustal (Refer Time: 27:00).

Clustal omega.

Student: mafft.

Then mafft, muscle.

Student: Muscle.

Right. So, you can use various software for the aligning the sequences. this multiple sequence alignment.

(Refer Slide Time: 27:11)

Conservation Score

- Conservation
- Frequency of amino acids
- Conservation score
- Online resources

Handwritten notes in red:

- $n(i)$
- N
- Entropy
- variant
- Sum of pairs
- $\frac{x - \bar{x}}{\sigma}$
- AL2CO
- CONSERV

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

Then we discussed about the conservation. What is conservation?

Student: Occurrence of residue in .

Yeah, residue is occupying the same position. For example if you have the 100 sequences, about 100 sequences take position number 5 it is alanine, how many sequences they have alanine in the position number 5. So, this will give you the information regarding the same amino acids presents in the same position right. So, how to calculate the conservation score? There are two sub procedure first what I have to do.

Student: (Refer Time: 27:42).

First we have to get the frequency of amino acids and then.

Student: Score.

We need to convert it into score. What are the various ways you get the frequency of amino acids?

Student: Weighted frequency.

Weighted frequency.

Student: Unweighted frequency.

Unweighted frequency and the random independent counts. What are the random unweighted frequency, how to get these frequency.

Student: Count of the.

n_i of i by N .

Student: Right

Right, for each position then how many times each amino acid represent and then see how many amino acids at particular position. Then we get the frequency then we converted this into score how to convert into score what the various a ways to convert the frequency to score.

Student: Entropy based entropy based.

Right the entropy based.

Student: Variance based.

Variance based.

And sum of pairs you can see the different ways to get this right. So, what is the program available to get the conservation?

Student: AL2CO.

AL2CO. How it works?

Student: We can give multiple sequence.

Yeah, we can give the multiple sequence well this is the alignment it is the multiple sequence alignment then they use any of this algorithms. So, for we can choose where we need entropy based measure or the variant based measure and also the frequency also we can choose based on that we will calculate the score it will normalize a score. What is formula to normalize the score: $x - \bar{x} / \sigma$. So, you can get kind of Z-score you can normalize the values then what are other online resources to get the conservation?

Student: Consurf.

Consurf right. So, you can here if you give the PDB id automatically it will get the sequence and calculate the overall sequences and get these conservation and for each positions. This will give a pictorial representation say the highly conserved as well as the less conserved, variable regions you can see. It can also provide the information regarding each position, what are the other residues which are present in particular position if any residue is the variable. So, then we did the phylogenetic trees. How to construct the phylogenetic tree? What is the information we get?

Student: Distance between the

How the sequence related which two sequence are closely related which ones are distant related right, So what are the various ways to construct the tree?

Student: UPGMA method.

UPGMA method this is the very popular common method. So, what is the basic information required to construct the UPGMA method tree?

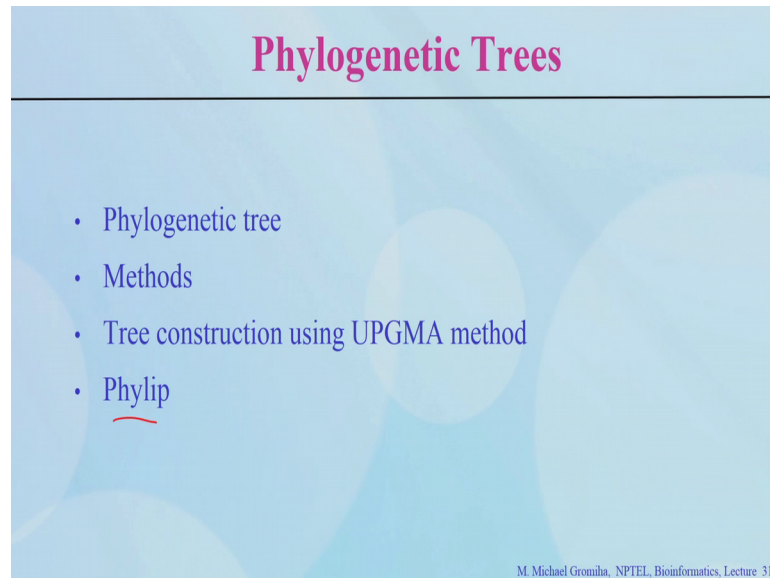
Student: Distance.

Distance you see the mismatches depending upon the mismatches you can say. If two sequence are less mismatches that means.

Student: Then they are.

They are closely related right. So, we can construct this phylogenetic tree. Then we discussed about the phylip right.

(Refer Slide Time: 30:28)



The slide has a light blue background with a darker blue header bar. The title 'Phylogenetic Trees' is written in a bold, dark red font in the header. Below the header, there is a bulleted list of four items in a dark blue font. The last item, 'Phylip', is underlined with a red line. In the bottom right corner, there is small text in a dark blue font.

Phylogenetic Trees

- Phylogenetic tree
- Methods
- Tree construction using UPGMA method
- Phylip

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 31

So, this widely used a program. So, we can use a phylip to construct the phylogenetic trees. Then see we can do this from amino acid sequences, we can get the alignment, we can get the multiple sequence alignment and the conservation, closed related sequences all the information you get from this sequence.