

Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture – 06a
Pairwise Alignment

In this lecture, we will discuss on pairwise alignment. So, in the earlier classes, we dealt with different types of sequences right for the macrobiological macromolecules what are the different databases we use to obtain the sequences? For example, for the DNA sequences.

Student: EMBL.

EMBL, Genbank and DDBJ right DNA data bank of Japan for the case of protein sequences.

Student: PIR and UniProt

PIR and UniProt right. So, if you have the sequences from different organisms how for they are similar? How far they are different? So, if the sequences whether they resemble each other and for any particular function, in this case, we need to compare the sequences and you have to align properly and you can then you can see the similarities and differences. For this case, if there are various techniques to compare the sequences, and we will discuss the details in this lecture.

(Refer Slide Time: 01:09)

Refresh

- Protein primary structure
- Protein Information Resource (PIR)
- UniProt
- Features of UniProt
- Obtaining data from UniProt

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

So, let us refresh ourselves regarding the contents we discussed in the previous class. So, what did we discuss in the last class? Protein primary structures. So, what is the primary structure?

Student: Linear sequence of the protein.

Linear sequence of the protein right. So, here the main chain is the same and side chain is different right. So, if you see you can see NH₂ Cα C N Cα COOH right. So, here you have the hydrogen and the R group. So, the main chain is a same, repetition right there is an elimination of water molecule, and here you have the side chain R₁ R₂.

So, there are 20 different types of amino acid residues right you can form any chains right based on this R₁ and the R₂. Then we discussed that nature selects particular specific combination amino acid residues to form a functional protein. Then we discussed where we shall get the information. So, there are several proteins which contain the sequences currently more than 70 million sequences are known right where shall we get the information. So, initially, we started to collect manually and publish a book called the Atlas of protein sequences right. So, then they develop this into the protein information resource called PIR in Georgetown (Refer Time: 02:27) University.

In the meantime the (Refer Time: 02:28) Munich research institute of Bioinformatics right they also tried to collect the sequences right for example, the Swiss-Prot and both

the places they develop several tools to analyze sequences. Later on they formed the consortium (Refer Time: 02:42) right they made the UniProt that is the universally accepted protein sequence database. Now, what are various features of UniProt, what are the major aspects of UniProt? High level of annotation.

Student: Minimal redundancy.

Minimal redundancy.

Student: (Refer Time: 03:00) high integration

And largely high integration with other databases right. So, it has various features UniProt. So, they classify into two or three (Refer Time: 03:07) different groups. In the first two parts, we can see the general information regarding the sequence, regarding the structure and the function and general information regarding a particular protein and second part we give the sequence information and structure information, and interaction information along with the links toward the databases right and in the last part they mentioned about the enzymes and pathways and so on.

They have a high level of integration as well as the links with all other databases in the literature. How to obtain the data for UniProt for example, if you want to obtain the data for transcription factors, how to get the data?

Student: (Refer Time: 03:44) Go to UniProt.

Go to UniProt.

Student: In the search option.

Search option use 'transcription factors'

Student: Transcription factor.

Then you will get a list of sequences if you want to remove redundancy, you can also remove the redundancy, but there are only a few options available in the UniProt what are the various options available in the UniProt.

Student: 90 percent.

90 percent.

Student: 50 percent

50 percent and 100 percent right you can all the sequences write the 90 percent and 50 percent. So, if you want 50 percent you will select the 50 percent and you can download the data right then you can do the different analysis. Then if you are interested in any specific protein then you give their particular protein. So, you get the information regarding their particular protein, for example, last class we discussed about the hemoglobin.

So, we got the sequence we got the function, we got the post-translational (Refer Time: 04:26) modification sites and the binding sites all the information you can get from the UniProt. Now we discuss about the alignment, what do you think about alignment?

Student: (Refer Time: 04:39).

Yeah. So, you can see how the evolutionarily related right, they evolved for various period of time.

(Refer Slide Time: 04:50)

Pairwise alignment

1: A I T S A V
2: A L S S A V

□ **Alignment** between two or more protein or nucleic acid sequences represents an explicit hypothesis regarding the **evolutionary history** of those sequences. As a direct result, it facilitates many recent advances in understanding the information content and function of genetic sequences

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

So, if we have one sequence and we have another sequence right. You can see the one sequence here and we have another sequence, how they are related to each other, how long it took to get different types of sequences. If you take the sequence 1 and sequence

2 you can see some differences. Here there is no difference, here there is no difference and you can see a difference here, and then you can see the relationship between two different sequences to understand how they are similar, what are the different functions whether they are affected by the functions and so on.

(Refer Slide Time: 05:24)

Pairwise alignment

```
>sp|P68871|HBB_HUMAN Hemoglobin subunit beta OS=Homo sapiens
VHLTPEEKSAVTALWGKVNVDENVGGEALGRLLVYPWTQRRFFESFGDLSTPDAM
GNPKVKAHGKKVLGAFSDGLAHDNLKGTFTLSELHCDKLHVDPENFRLLGNVLV
CVLAHHFG KEFTPPVQAAYQKVVAGVANALAHKYH

>sp|P02112|HBB_CHICK Hemoglobin subunit beta OS=Gallus gallus
VHWTAEKQLITGLWGKVNVAECGAEALARLLIVYPWTQRRFFASFGNLSPTAILGN
PMVRAHGKKVLTSFGDAVKNLNLIKNTFSQSELHCDKLHVDPENFRLLGDILIVLA
AHFS KDFTPECQAAWQKLVRVVAHALAHKYH
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

So, here I get two examples; the first one is the human hemoglobin and the second one is chicken hemoglobin. So, if you look into this two sequences, can we see any difference or similarities between the 2 sequences?

Student: (Refer Time: 05:36).

Here this is similar and you can see some places right WGKVVN right. So, here you can see right. So, when you look at your eyes you can see some cases, but we require an algorithm right to quantify how far these two sequences are similar to each other which part of this sequences are similar which part of the sequences are different right. So, here again, you can see KYH is the same, but again here is the difference.

(Refer Slide Time: 06:09)

| Sequence comparison | |
|---|--|
| Different types of pairwise comparisons | |
| Method name | Situation |
| Dot plot | General exploration of your sequence Discovering repeats Finding long insertions and deletions Extracting portions of sequences to make a multiple alignment |
| Local alignments | Comparing sequences with partial homology Making high-quality alignments Making residue-per-residue analysis |
| Global alignments | Comparing two sequences over their entire length Identifying long insertions and deletions Checking the quality of your data Identifying every mutation in your sequences |

So, in this case, there are various methods available to compare these 2 sequences like this called the pairwise comparison. The simplest one is dot plot right this will compare these two sequences and if the sequences are the same, then we put a dot right I will explain in a couple of minutes right. And then from this plot, we can see whether the same residue is located at the same place or there is any shift like this is one residue shift in the first sequence or any shift in the second sequence that there we will see.

So, this plot will give some information and regard you can see some repeats, if you can any region which is the same in between the two sequences and you can see any long insertions or the deletions for example, if there is shift in the first sequence or the shift in the second sequence, and also you can see some portioned sequence which can be aligned together for the different sequences this is the first one, then the second one you can do local alignments.

Here you can see instead of the whole sequences, you can do it for some small portions; whether any small portions we can have the alignment between these two sequences. Then we do a global alignment, the global alignment will give you the whole alignment for the complete set of the sequence. For example, if you see here you can align for the whole sequence or you can align only some specific portions to the sequence.

How to align the sequences, how to score the sequences, that we will discuss in a couple of classes.

(Refer Slide Time: 07:36)

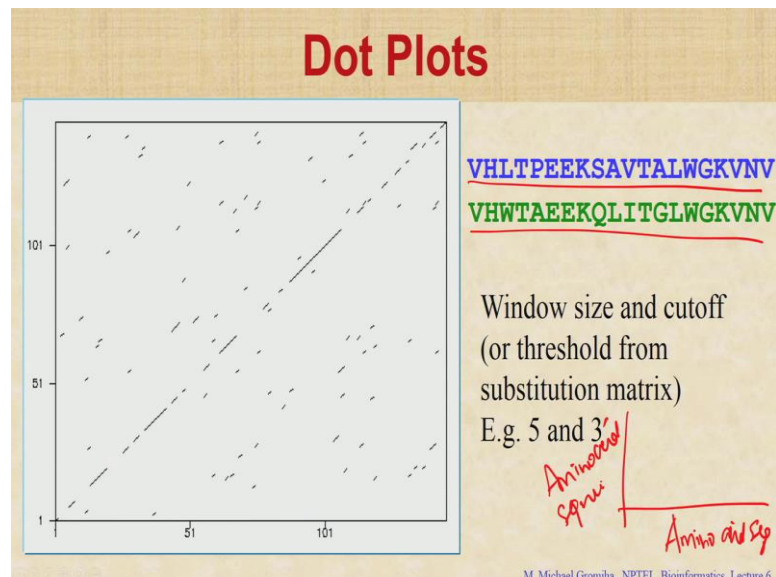
Dot Plots

- One of the simplest methods for evaluating similarity between two sequences is to visualize regions of similarity using dot plots.
- To construct a simple dot plot, the first sequence to be compared is assigned to the horizontal axis of a plot space and the second is then assigned to the vertical axis.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

So, what is the dot plot what is the information we can get from dot plot? It is one of the simplest methods because we have the sequence A and sequence B. So, you compare these two sequences and make a dot if it is they are similar to each other, and you can get the similarity plot between the 2 sequences.

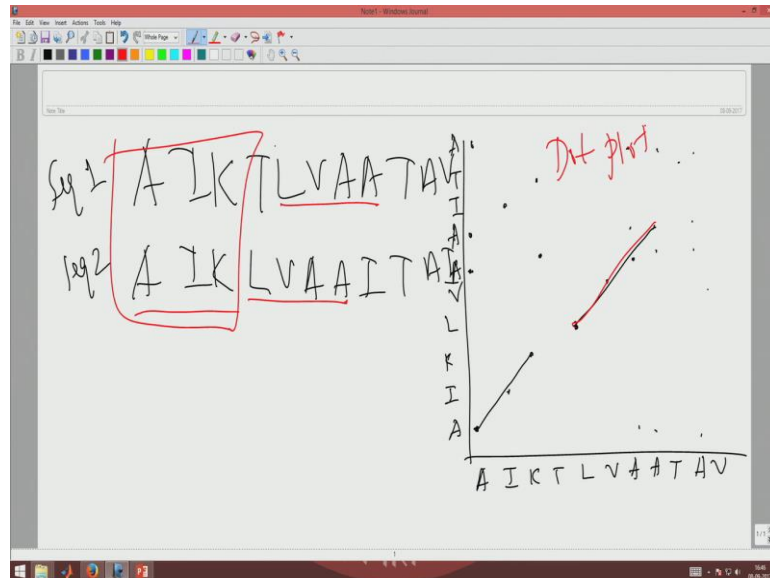
(Refer Slide Time: 07:55)



How to do it? So, I will give you the one sequence where the second sequence right. So, take the compare the sequences take sequence one and sequence two right here this is the

x-axis, and here is the y-axis, this amino acid sequence, here you put the amino acid sequence.

(Refer Slide Time: 08:23)



So, if you have two sequences I will write it clear. So, for example, I put this sequence here right. So, you put another sequence here. So, this is sequence 1 this is sequence 2. So, you can make a plot right. So, do it here AIKTLVAATAV right go the y-axis. So, here is a second sequence AIKLVAAITA right.

So, in this A here. So, where the 'A's present in the y-axis, here you have 'A'. So, then where here you have a here you have A, here you have A. So, (Refer Time: 09:27) Next is 'I' where are the 'I's here and here one 'I' right then 'K'. So, 'K' is here that is all, then 'T', here one 'T' here right.

Student: (Refer Time: 09:46).

Here, here there is another one, then L, one L is here, no other L right then V where is V? V is here, then A, A again here, there is one A here, another A here another A here right then this A this A here, here, here, here, then T here this A here, here, here, here, then V, one V is here.

Student: (Refer Time: 10:40).

Only one right. So, if you see this one. So, if you see these are the cases you can see the continuous dots. So, if to put to the line here this is continuous and here this is one shift and then you can see the continuous line here right. So, this will tell you if it this is the diagonal then one can infer from this diagonal wants yeah. So, these 3 are diagonal AIK is same. So, here you can see this is same.

Student: (Refer Time: 11:07).

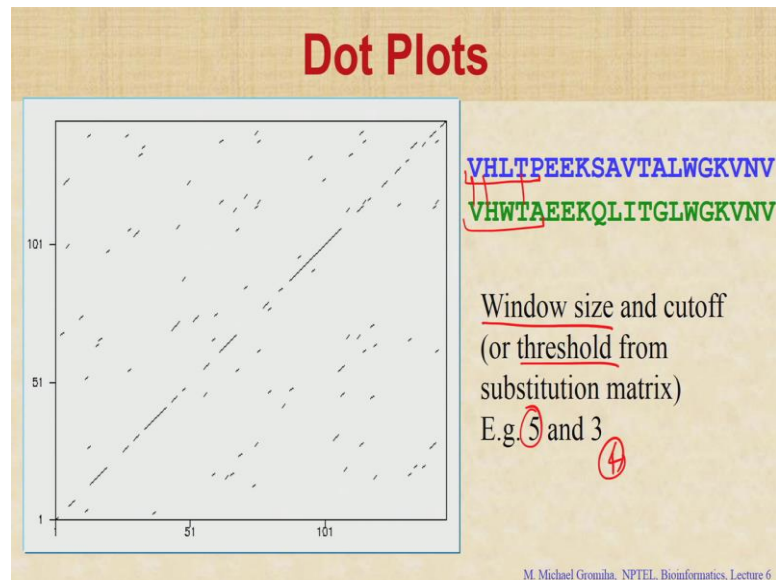
Right the position the same now second one if you see look at here, here this is a shift. So, you can see a shift between sequence one and the sequence two. So, you can see you can still see here LVAA here, you can see either same one here with the shift of one residue. So, if you have the sequences it is little difficult to see where are the residues which are exactly matched and where they have the shift, but we make a dot plot like this is the dot plot easily you can look at the residues where they have matches and where they have the shift either in the first sequence on the second sequence. So, here there is a shift in the first sequence because there is shifted here right. So, I can say down.

So, likewise if you the next another sequence you can have a shift the above the diagonal right you can from this one you can see where you have the shift in the first sequence or the second sequence. So, what is the application of this dot plots?

Student: We can find alignment and also we can see if there is any (Refer Time: 12:08) shift from the alignment.

Right, you can easily see if you make a plot, immediately you can see that whether there is an exact match either at the diagonal or at the any of the up the diagonal or the down the diagonal. So, here this is how to construct the dot plots, first dots are placed in the space of each position, where the sequence elements are identical and then it will give you the identity between the two sequences right and the diagonals of this dot plot. So, this here we exactly looked at the same one then we can also used a different window size.

(Refer Slide Time: 12:33)



You can use a different window size and you can use some threshold; for example, if you take a window size of 5. So, you take this 5 compare this 5 and among the 5 how many them are matching. So, then you put a threshold, for example, 3. So, your 3 matching 1 2 3 then you can say there is a match then go to the next window, this 5 then whether put the threshold 3 which is matches are not if we match put here dot. Then you can get different types of plot if there is a one sequence difference, then do you consider there is specific a difference because they make the window average right either you can 3 or 4 you can put right. So, you can get a plot right this is a smooth plot with respect to the two sequences sequence one and sequence 2.

If there are single mutation right here we do not find the difference, but we make a real dot plot with the same sequence, then you can see the difference you only a mutation you can see a difference right there is a difference. So, this is a protein data. So, here x-axis is the amino acid sequence, y-axis the amino acid sequence right here we kind of put a 150 amino acid residues right. So, you can see some regions which are similar right this is this one right we have the sequence. So, if we see there are some regions which are same first few residue are same, then again some residues are same.

So, we can see the regions where you have the long stretch of residues that are around 100 residues. In the sequence, if you have the 100 the residues you can see the difference. So, there are many residues they are closely related to each other. So, now,

we can for any proteins we can make a plot, and you can analyze where are the regions where they are similar to each other between any two sequences. So, then what is simple alignment? This simple alignment as a discussed earlier just have that sequence 1 and you have sequence 2 right just how far they are aligned with each other; that means, what is a how about the match between the characters in each sequence.

(Refer Slide Time: 14:31)

Simple alignment

- The three kinds of changes between sequences are
 - (i) a mutation that replaces one character with another
 - (ii) an insertion that adds one or more positions
 - (iii) a deletion that deletes one or more positions
- Insertions and deletions occur less frequently than mutations. In these cases, gaps in alignments are commonly added.

AITVSA
AITA--A

AITVSA
AITAA

AST → AIST A → V AIT → AIST

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

So, I will show if you have different sequences for the different sequence for example, you put two sequences and you put another sequence here for example if you have this one AITVSA and AITAA. So, you can make these two sequences aligned with each other with different aspects. One that is a replacement of characters for example, if you see here. So, this here valine is changed to alanine and this type of changes; it is called a mutation or substitution.

So, if you change or replace one amino acid by another amino acid and that is called mutation. There are two other aspects one is called the insertions here we add more data more sequences right for example if there is AIT, if you add any of the amino acid in between from anywhere, for example, you add S then this will become AIST this called the addition. So, first one is the mutation, mutation refers to the change of amino acids right for example if there is A and this is changed to V or is a V or change to A this mutation.

An insertion that adds one residue right in any sequence, then there is another called deletion in this case some residues are deleted in the sequence, for example, if there is a sequence AIST and this I is deleted then we get the sequence AST right here this I is deleted. So, these are the 3 kinds of changes in the sequences, either you have mutation or you have the insertion or you have the deletion. So, there if you look into this sequences the insertion-deletions they occur less frequently than mutations. If you see different sets of sequences from the same organisms you frequently see the mutations.

Changing one amino acid by another amino acid right so, but less possibility of insertion-deletions, but the insertion-deletions we can represent by gaps right, but this is not frequent. So, this will influence with a high penalty, when you make a sequence alignment ok.

(Refer Slide Time: 16:53)

Example

Seq 1 AATCTATA and AAGATA Seq 2

- Dot plots are useful for visual inspection of the regions of similarity between sequences.
- A numeric system for evaluating sequence similarity has obvious advantages for objective determination of optimal alignments.
- Scoring alignment without gap: credit for aligned residues and penalty for mismatches

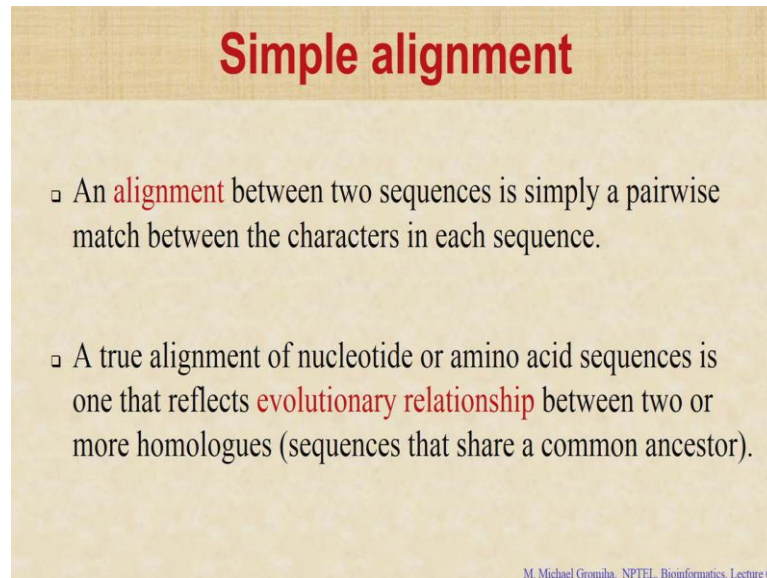
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

Now, you show two examples first this is the sequence 1 here sequence 2. So, how far these two sequences are related with each other? So, first, as we discussed we can have dot plot easily we can see. So, you can get a construct a dot plot and you can see where you can have the same amino acids or any insertion-deletions. Then we need a numeric system to evaluate how far these two sequences align with each other.

They align well or this is a good alignment or a bad alignment. So, we need this score in this case we need to develop some of these matrices, to score whether sequence one is related with the sequence two or not. So, in this case, we need a scoring measure. So, we

need 3 different types of measures, we need right because we discussed about 3 different types of changes what are 3 different types of changes.

(Refer Slide Time: 17:45)



Simple alignment

- An **alignment** between two sequences is simply a pairwise match between the characters in each sequence.
- A true alignment of nucleotide or amino acid sequences is one that reflects **evolutionary relationship** between two or more homologues (sequences that share a common ancestor).

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 6

Student: Mutation.

Mutation

Student: Insertion.

Insertion and deletion right. So, insertion and deletions we represent as a gap. So, we give the gap penalty for the mutations right we need to give the penalty depending upon the types of mutations, then some case you will use the same residue there is no change. So, that will be rewarded. So because that maintains the same residue so the two sequences are similar to each other.