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

Lecture – 13a
Construction of Non-redundant Datasets I

In this lecture we will discuss about the construction of non redundant datasets. Why do we need non redundant datasets? What is the meaning of non redundancy, how this datasets influence the prediction models or any large scale data analysis. So, in the last lecture, do you remember, what did we discuss in the last lecture?

Student: Hydrophobicity profiles.

Mainly about hydrophobicity profiles right. So, what is hydrophobicity profile?

Student: Sequence versus.

It is a plot connecting, the amino acid sequence.

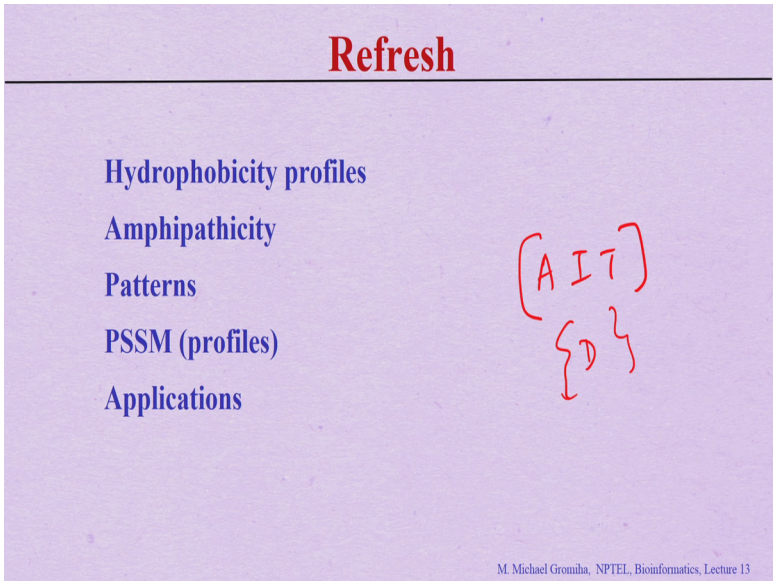
Student: hydrophobicity.

Versus hydrophobicity values. So, we have experimental hydrophobicity values for the 20 different amino acid residues right for each residues you assign the values and construct the plot, which information we will get if you construct hydrophobicity plot?

Student: So, like secondary structures.

Correct, you can see some sort of patterns or motifs right. So, you can see any specific patterns for example, in alpha helices right or beta strands for example, alternating hydrophobicity, transmembrane segments dominated with the hydrophobic residues.

(Refer Slide Time: 01:32)



And we can see some profiles and we can see the patterns, and relate these patterns to any specific structure or function right.

Then we discussed about amphipathicity like; what is amphipathicity?

Student: hydrophilicity.

It is a periodicity of this polar and nonpolar residues in any amino acid sequence. You can see some specific periodicity in the case of alpha helices or beta strands and so on. Then we discussed about patterns, where patterns are defined with some symbols, for example, square bracket right. What is the meaning of square bracket?

Student: Inclusive.

Right you can include any amino acids which are given within this square bracket right for example, if we have square bracket and AIT. So, any residues within this groups is allowed. If it is curly brackets,

Student: not allowed

So, the residues which are inside this bracket for example, if it is D, the D is not allowed at that particular position.

So, likewise you can make any motifs or you can see any patterns, you can also search the database right for using the PIR or any search right. You can see the whole UniProt sequences, whether you can identify any specific patterns, because the patterns and motifs are important for several functions. Then we discussed about the profiles where position specific scoring matrices or position weight matrix right. How to construct the position weight matrix or scoring matrix?

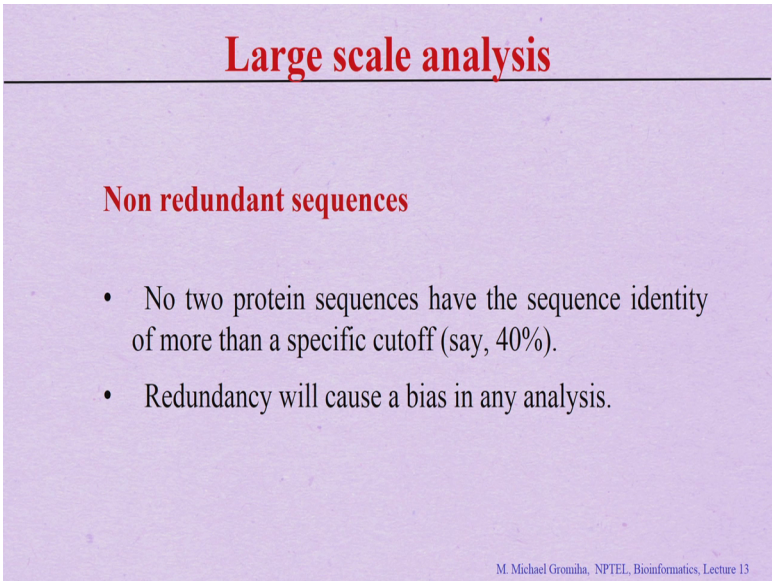
Student: Sir we take the frequency.

You take the frequency of residues, first we need the multiple sequence alignment. From the multiple sequence alignment we get the frequency, where we convert the frequency into this matrix, alignment matrix right, by normalizing with the probability of residues right for example, in the case of nucleotides or for example, the amino acids. Because a nucleic acids right the probability is 0.25, the amino acid sequence 0.05. So, we can use the pseudo counts and finally normalize right you will get the matrix.

So, we discussed various applications for example, the prediction algorithms and the binding sites right different types of functional important sites and so on. So, if you do any analysis in bioinformatics, it is very important to construct dataset, because dataset plays an important role in the analysis as well as in the prediction. If you have very good data sets right, then you can relate the reliability of any prediction methods is very high right.

So, because the results are biased with a dataset right. In this case it is very important to construct a dataset which are not redundant with each other. In this case what is the meaning of redundancy, what is the meaning of non redundancy?

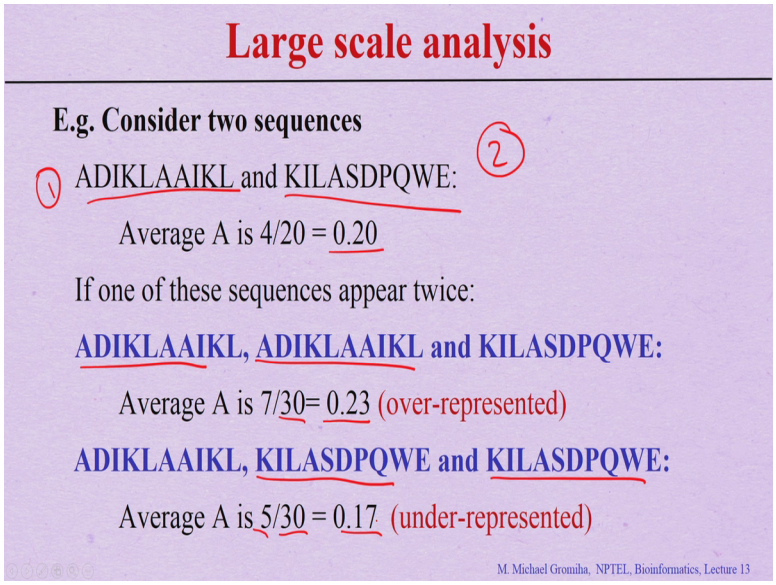
(Refer Slide Time: 04:13)



For example if the two protein sequences which are similar to each other, you can take any specific cutoff for example, 70 percent or 80 percent. When its 80 percent in the sequence A with the sequence B right, 80 percent of the sequences are the same or similar. See if you add up these sequences this will introduce a bias of this sequence; because they influenced with the data which we obtained right, from these sequences.

So, in this case we need to avoid bias, we need to construct the sequences, non redundant sequences at any specific cut off depending upon the problem you choose. If you have 100,000 sequences right, we have more number of data right then we can decrease the cut off, we can use 20 percent or 30 percent right, so that you will get sufficient number of data for the analysis. If the initial data set is small then we need to set that cutoff accordingly so that you will get sufficient number of data for analysis. Because for the redundancy this will cause a bias, so it is important to construct a data set of completely non redundant.

(Refer Slide Time: 05:28)



For example anything with 40 percent or so. So, for example, here I have two sequences right, this is one sequence, sequence 1 and here you have another sequence, sequence 2 right. If you take these two sequences, what is the amino acid composition for alanine? Total 10 residues.

Student: Four.

Total 10 residues. How many alanines?

Student: Two

Right 1, 2, 3, 4. So, 4 by 20 this equal to 0.20 right. In some cases for example, if one of the sequences appear twice; that means, we introduce a bias in this analysis for example, these sequence appears twice. Now totally how many residues? 30 residues right. Out of 30 residues how many alanines?

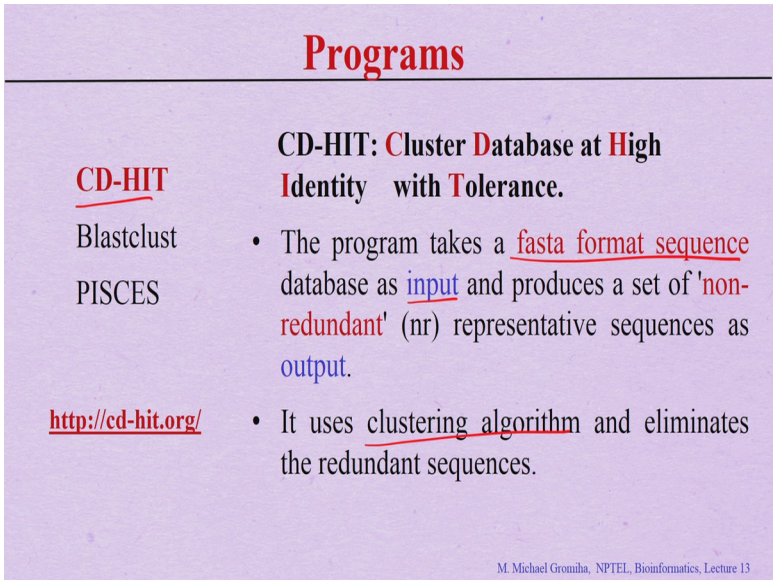
Student: 7.

It is 7. So, if you take the composition now at 0.23, this is over represented. If you have this second sequence twice, then what is average alanine? Totally 30 residues.

Student: 5.

5 by 30 equal to 0.17 right. Likewise if you include the same or similar sequences in your dataset, and we extract information, you can see a bias right that you have introduced because of these same sequence or similar sequences right you are including in your data set. Since it is very important to construct datasets for any analysis, which are non redundant with each other. So, how to get this non redundancy, how to construct non redundant sequences? So, in bioinformatics there are various approaches right, I will explain the one or two important approaches, and certain software available to construct non redundant data sets.

(Refer Slide Time: 07:18)



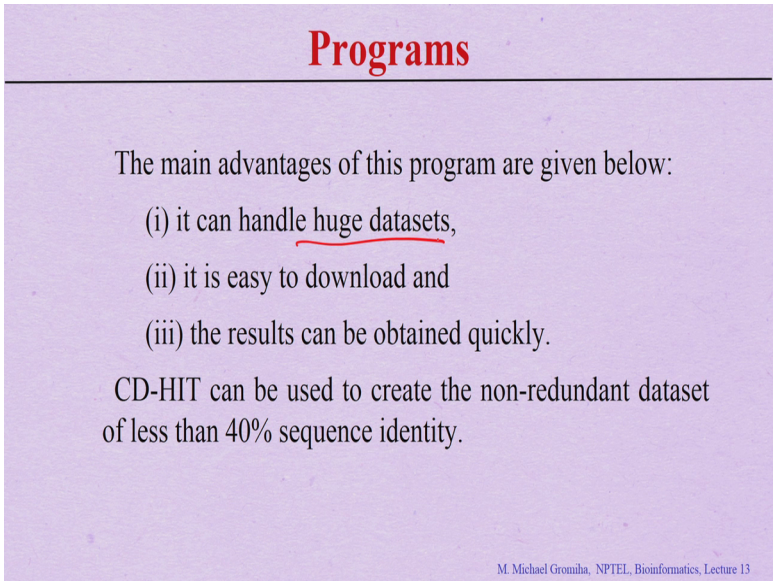
One of the most popular software that is called CD HIT. CD HIT refers to the cluster database at the high identity with tolerance.

So, what this program does? So, our aim is to get the non redundant sequences. For example, if you have 1000 sequences or 10,000 sequences. You need to extract the non redundant data from this initial data set right. So, here you have to give these sequences and good, because in this case we need to remove redundancy from the sequence. So, you take the fasta format right, fasta format we discussed earlier, it starts with the greater than symbol right. So, it takes all the sequences as input and it gives a set of sequences which are non redundant, that depends upon the user's choice, we can specify the cut-off, we can specify the word size right. I will explain now right and this CD HIT will treat all

your sequences as input, and depending upon your convenience or your needs right it will give you a non redundant dataset.

So, how it gives a sequence, that it uses a clustering algorithm, to eliminate the redundant sequences; it takes your sequences, and makes some clusters, and based on the clustering it will take the non redundant from picking of the sequences from each clusters. What is the use of this program, why this CD hit is very popular? Because this program handles huge data sets right.

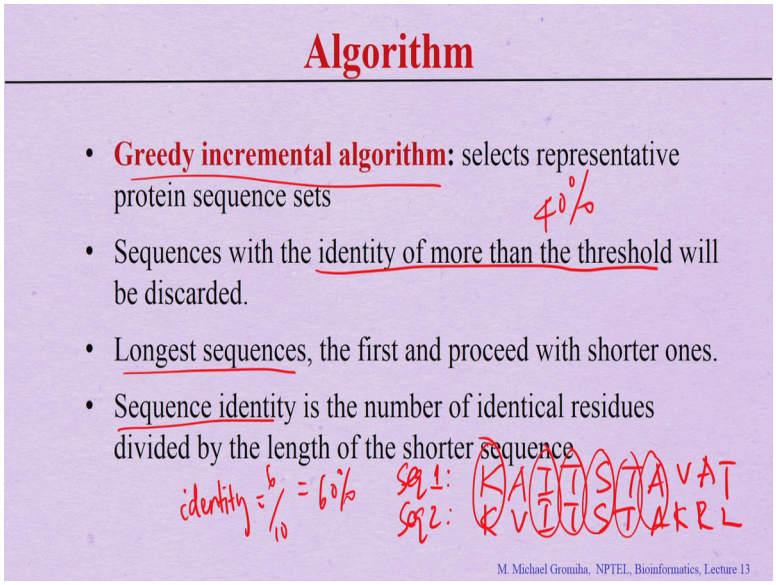
(Refer Slide Time: 08:48)



Even if we have thousands of data, it can handle the huge data sets and give you the results very quickly. Then it is easy to download, and this CD HIT, we can get the results very quickly.

So, you can use a CD HIT to create non redundant sequences at different sequence identities. For example, 40 percent, 50 percent, 60 percent so on. The disadvantage of CD HIT, it has some limitations to restrict the sequence redundancy, either up to 30 percent or 40 percent, depending upon the program you use online, or you get the downloadable version. So, that is only disadvantage, other than that, if you want to get a specific sequence cut off up to 30 percent right we can use CD HIT and it is easy to use to get the results very quickly.

(Refer Slide Time: 09:42)



Let us see how the CD HIT works. So, it uses the greedy incremental algorithm to select the protein sequence sets. So, it takes some clustering technique, right try to assign each sequence, and making into particular clusters and then check any of these features and using the features you try again and again and again right, this is called a greedy algorithm, unless the deviation is very less, even though it is not satisfied. So, look for the other better solutions again and again right and to find the most probable solution for getting your non redundancy sequences.

So, how it does, it takes a sequences, and we can specify the identity for example, 40 percent and if this specific threshold, if it is more than the threshold, for example, 40 percent that will be discarded. So, then it takes the longer sequences first, and then proceeds with the shortest ones and what is the sequence identity? Right that we discussed earlier. Sequence identity is the number of identical residues divided by the length of the shorter sequence. For example, if you have two sequences; this is sequence 1, then sequence 2.

So, what is the sequence identity?

Student: 6.

Right this is same, 2, 3, 4, 5, 6; 6 by 10 this equal to 60 percent. If a two sequences which are more than the particular threshold for example 40 percent right will we

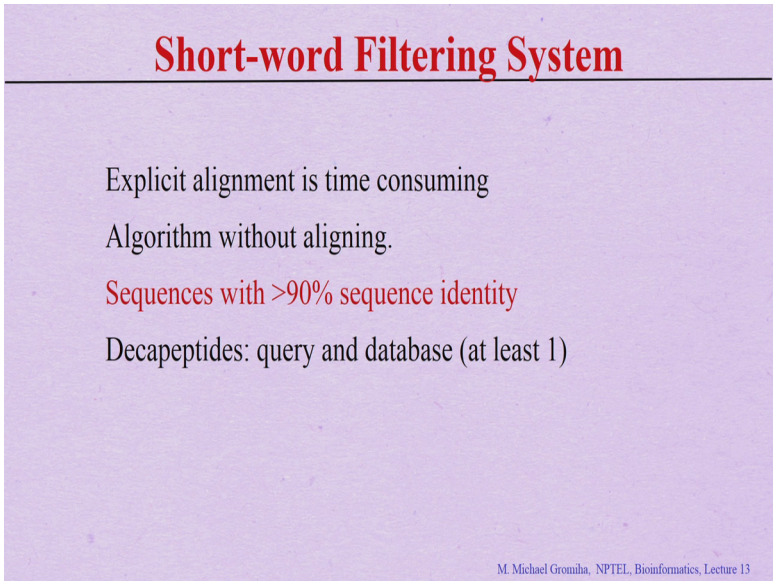
consider both or will we consider only 1? Only 1, right because it is more than 40 percent. So, do not keep both. So, they keep 1 and discard other one right. This is how it works.

So, if we take the explicit algorithm alignment. So, which algorithm we use for sequence alignment?

Student: BLAST.

BLAST right. So, BLAST is widely used, if you have several sequences, works on thousands of sequences, if you want to make the sequence alignment, complete sequence alignment, it is very time consuming, right it takes lot of time to make the alignment and find this has identity or not. So, this implements an algorithm without aligning the complete sequences.

(Refer Slide Time: 12:19)



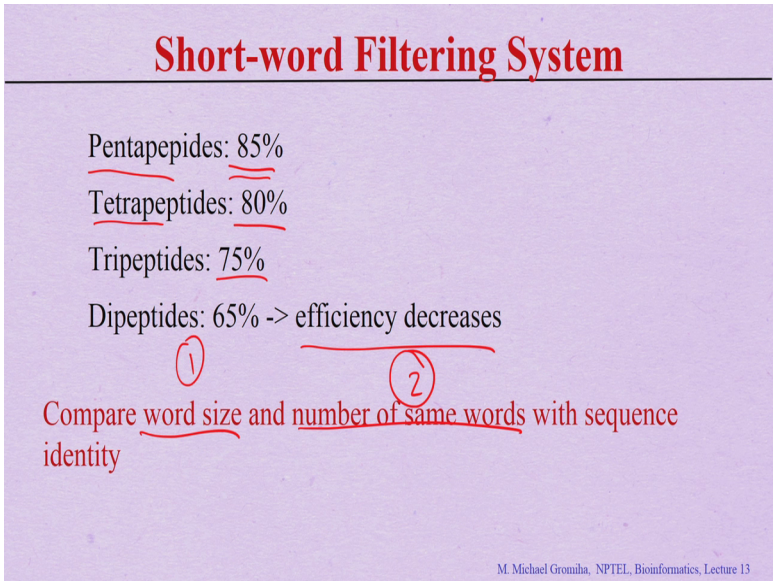
So, it takes a specific set or specific subset, and then sees whether you can find any subset of residues which are same between two sequences. For example, if two sequences A and B, they are considered to be 90 percent sequence identity, there is at least one peptide right at least of 8 residues like decapeptides or 10 residues, which are available both in query and database. If you look into these two sequences which are identical or similar or the high identity, then you can see long stretch of residues which

are similar to each other for example, 100 residues and identity is 90 percent. How many residues are same 90 residues are same?

Student: 90.

90 residues are same. In this case if you look in the sequence, at least some segments, if you see the large stretch of residues which are the same. We start that assumption, they try to use only decapeptides also the different lengths, and see whether any segments right which are same in both that sequences. They do not have to do a complete alignment, look for the stretch of residues, when they are seeing, then they assume that they are 90 percent identity. See did you see because of the reason, why because the highly redundant, there should be some stretches which are the same among different sequences.

(Refer Slide Time: 13:45)



Then with a different sequence identities for example 85 percent or 80 percent or 75 percent, they try to reduce the length for example, 4 residues or 5 residues like pentapepides right or tetrapeptides or tripeptides, and they also set more number of segments.

For example, if you take pentapeptides, there will be they treat that may be for a example 5 or 6 pentapeptides minimum 5 pentapeptides, the same then you can say that these two

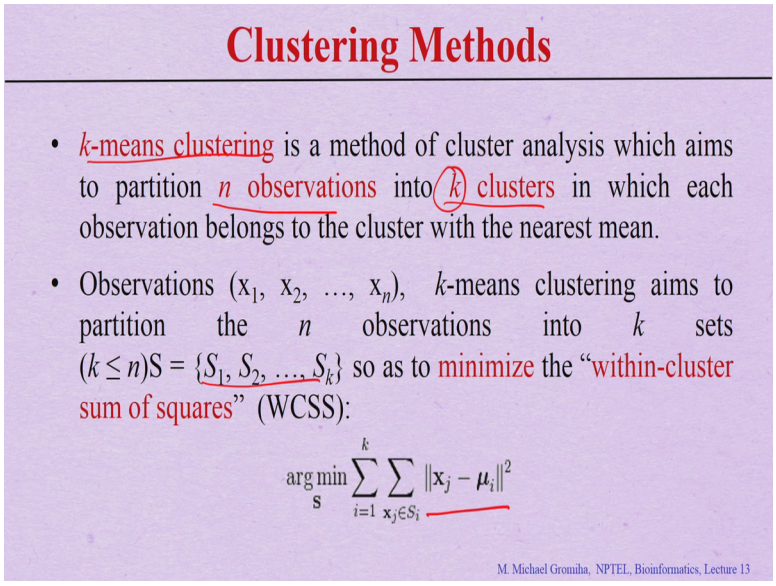
sequences which have the identity of about 85 percent. And the second aspect is, when they reduce the word size then the efficiency decreases right. why?

Student: Because this.

Because if we have a dipeptides or tripeptides, there is possibility of several pepetides, but they are not similar, even then you can see several tripeptides or dipeptides they are similar to each other, the possibility is very high. Even if you restrict the number of times right you can see the dipeptides or tripeptides. So, they started with the longer segments and reduce the segments and then increases number of segments to identify the residues which are similar or redundant with each other.

So, there is the word size and the another one is a number of same words, try to identify the sequence identity that two conditions, one is word size and the second one, the number of same words, how many times the same words appear different at different locations.

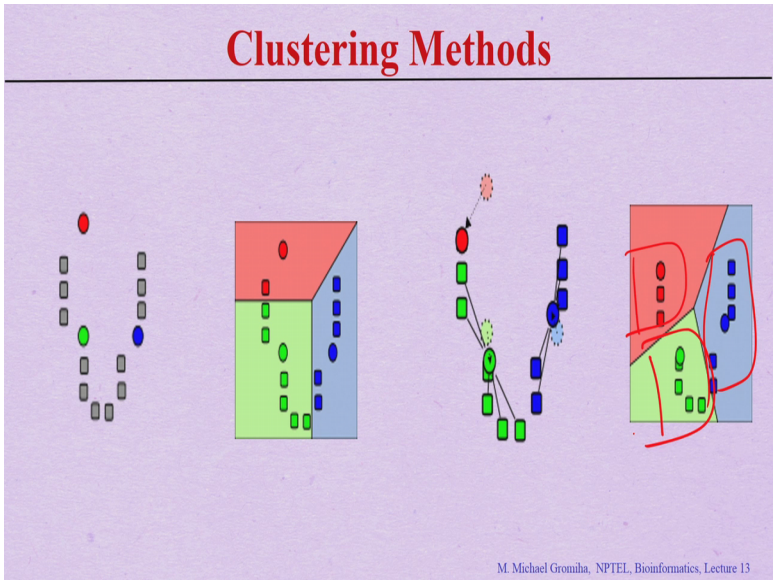
(Refer Slide Time: 15:20)



So, how the CD HIT works right. It uses the clustering techniques one of the popular techniques there is called the k-means clustering; how the k means clustering works? It is a method of cluster analysis, which aims to partition n observations right for example, if you have n sequence right, n observations into limited number of clusters.

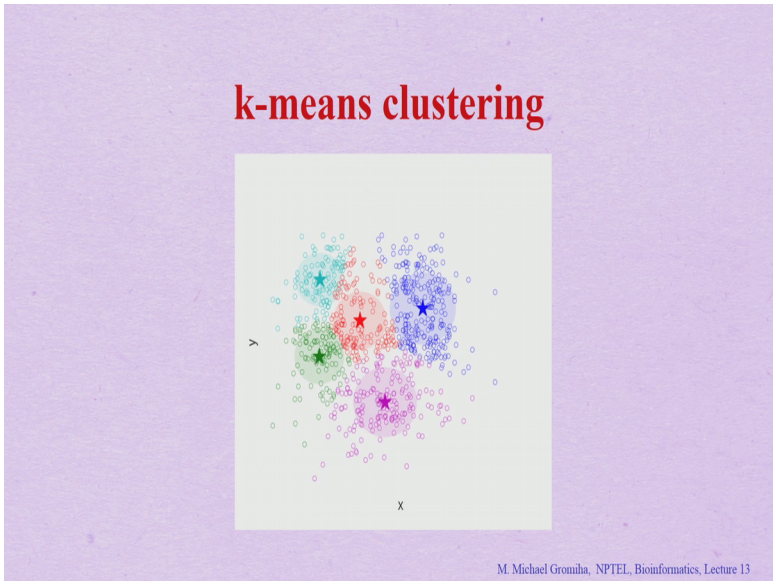
For example here k clusters, say 5 clusters or 10 clusters. In which each observation belongs to the cluster with the nearest mean right because there should be within this nearest mean. For example, if you have n observation like (x_1, x_2, …x_n) right the k means clustering right tries to put these n observations into k sets like different sets. So, that it minimize the within-cluster sum of squares right. If you see the within the clusters right you can see the mean and any new sequence right; the distance between these two features right this should make it as minimum. So, take any clusters, all the representative sequences within the clusters right there should be the minimum, fine.

(Refer Slide Time: 16:21)



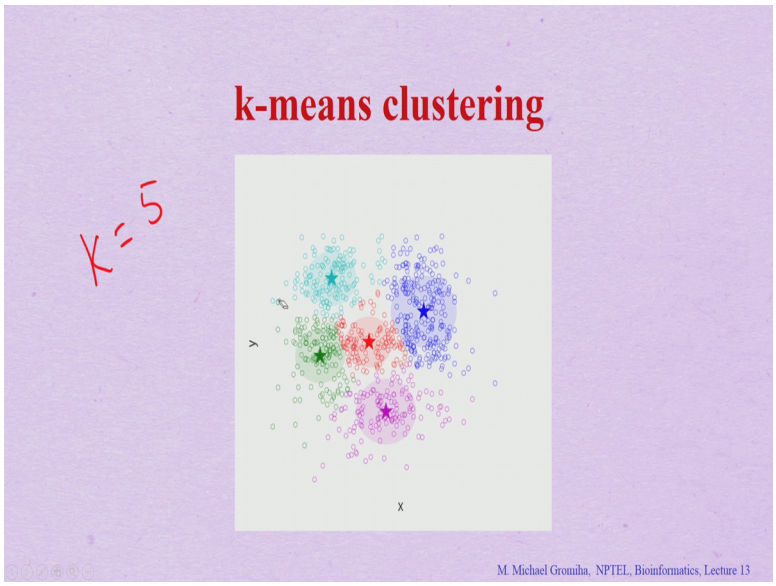
So, here I show one example, here they like to have three different clusters and for any sequence they try to find out the most probable cluster, and then once it is found with one cluster then they reorganize change the values, then again they assign the other residues so that for each cluster the value should be minimum. So, finally, you can see here this one cluster, here this another cluster and put the third cluster.

(Refer Slide Time: 16:50)



So, I can show one animation. So, here we know n data, n observations. So, we need to decide how many clusters you want to form. In this figure, we had right how many clusters are desired? 5 right.

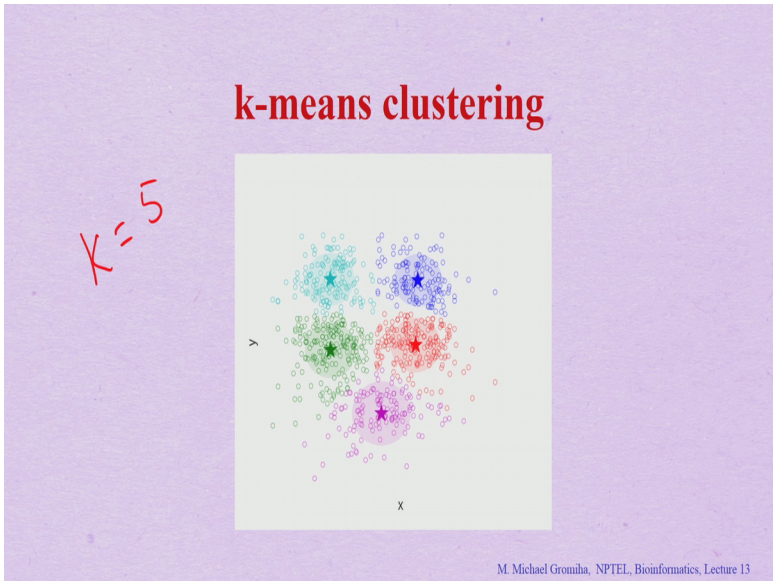
(Refer Slide Time: 17:07)



Here k equal to 5 right. 5 clusters. So, first randomly they put into 5 clusters, and try to see the average values for these 5 clusters and try to rearrange themselves. So, that you can see for any specific clusters, the mean value that it should be the minimum within that particular cluster.

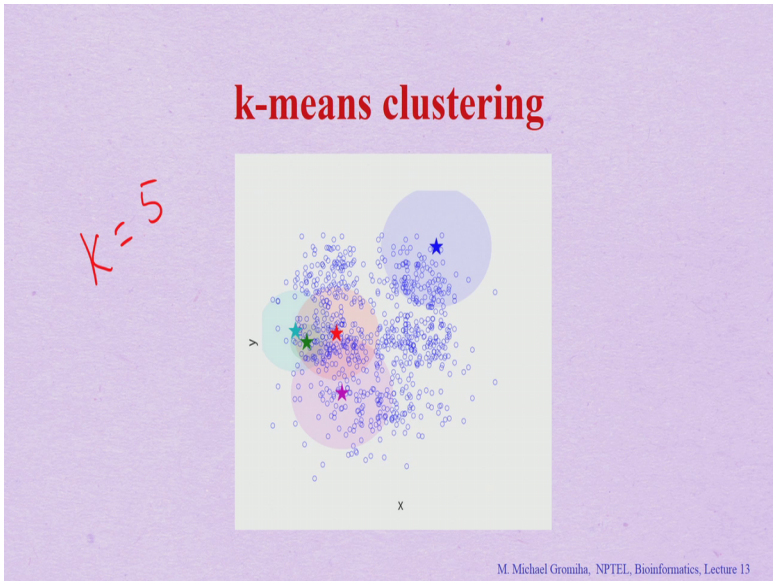
So, you can see cluster here, and you can see the cluster here right. So, they will see 5 different clusters right we can see the different iterations right, how these points move, you can see the 5 stars right move everything together, and when converge they can see the clusters are very a restricted right, the deviation is very restricted. So, they can form in particular clusters.

(Refer Slide Time: 17:53)



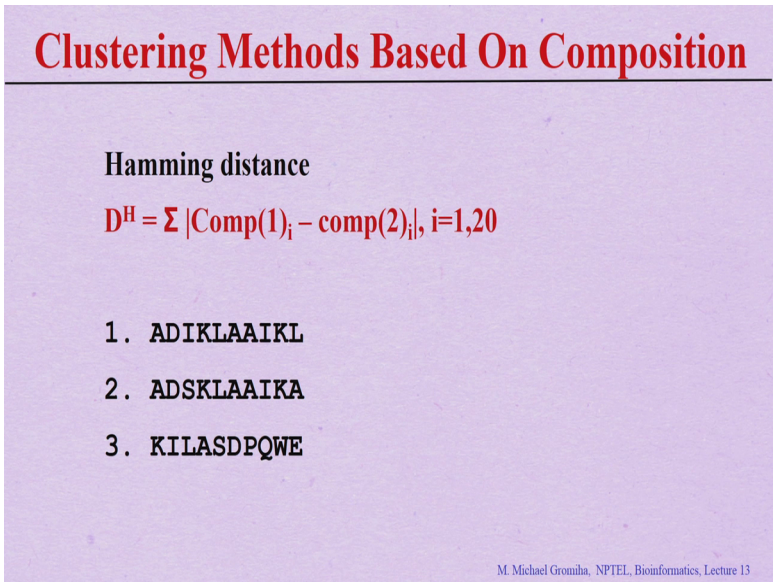
Then we take the representatives from each clusters right. So, that they can form any non redundant data, any sequences or any data.

(Refer Slide Time: 17:57)



So, what are the various features, how to get these sequences in a particular cluster. There are several ways, there are several features you can a think of to cluster these sequences.

(Refer Slide Time: 18:14)



So, one is a hamming distance, based on the composition for example, if you deal with sequences, we discussed various features; what are various features we discussed for the amino acid sequences?

Student: Occurrence.

Occurrence.

Student: Composition.

Composition, pair preference, molecular weight, hydrophobicity. So, any features right you can calculate for example, if you take the property say hydrophobicity. If you have 100 sequences, you can calculate the hydrophobicity for all 100 sequences and make some threshold like for example, it is between 10 and 12 kcal are 12 and 14 kcal, 14 to 16 kcal right within that clusters we can put together. So, how many sequences do you require? So, you can put these different clusters right and then see you what is average values. Then if I add or remove anything, how these value changes and if it is standardized then you can get the representatives from each cluster.