

Lecture – 13b
Non-redundant Datasets II

(Refer Slide Time: 00:17)

Clustering Methods Based On Composition

Hamming distance $\text{Comp}(i) = \frac{n(i)}{N}$

$D^H = \sum |Comp(1)_i - comp(2)_i|, i=1,20$

A C D E F G H I K L M N P Q R S T V W Y

1. ADIKLAAIKL → 0.3 - 0.1 - - - - 0.2 0.2 - - - - - - - - - - - - - -
2. ADSKLAAAIKA → 0.4 - 0.1 - - - - 0.1 0.2 0.1 - - - - - 0.1 - - -
3. KILASDPQWE → 0.1 - 0.1 0.1 - - - 0.1 0.1 0.1 - - 0.1 0.1 + 1.01 - 0.1 -

$D^H(1-2) = \frac{0.4}{0.8} = 0.5$ $D^H(2-3) = \frac{0.04}{1.01 \times 7} = 0.0059$ $D^H(1-3) = \frac{0.09}{1.01 \times 5} = 0.0189$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 13

So for example, there one example is Hamming distance, so this is based on the amino acid composition; how to calculate the composition?

Student: Number of amino acids

$n(i)$ divided by N ; so N is a total number of residues in a sequence and this $n(i)$ is residue of particular type; any residue, alanine or whatever, 20 residues. Now I get 3 sequences; so how many clusters you can put in these three sequences? And which two sequences are close to each other? So, just we can use this formula, Hamming distance; this is equal to composition 1 of i , minus composition 2 of i ; take the absolute difference, i varies from 1 to 20.

So, what is the different composition for a sequence 1? How many residues? 1, 2, 3, 4, 5, 6, 7, 8, 9, 10; for simplicity I put 10, but usually what is average length of protein sequences?

Student: 100.

Right you can see large crowd around 200 to 300 or 100 to 300; average length is 315 residues depending upon the sequences, deposited in the UniProt database. Also I showed the graph connecting the number of residues versus the frequency. You must know 100 to 300 residues; so, for example if we take; for simplicity I use small peptides what is the composition for A in the first sequence?

Student: 0.3.

0.3; C, 0; D 0.1; E?

Student: 0.

F?

Student: 0

0; G?

Student: 0

0. H 0; I?

Student: 0.2.

0.2; K?

Student: 0.2.

K is 0.2.

Student: L is also 0.2

L? 0.2; M no; N?

Student: 0.

No.

Student: All other 0.

All are 0; 0.2, 0.4, 0.6, 0.91; then for the second one?

Student: 0.4

0.4.

Student: 0.

0.1.

Student: 0

E, F, G, H, I is 0.1; K?

Student: 0.2

0.2; L?

Student: 0.1.

0.1; 4, 5, 6, 8, 9 there is one more, S; 0.1. So, third sequence; so, what is the composition for alanine?

Student: 0.1

0.1; no C; D 0.1; E 0.1; F no, G; H, I 0.1; K, 0.1; L 0.1 M, N, P 0.1; Q 0.1; no R, no S.

Student: 0.1.

W 0.1; 1, 2, 3, 4, 5, 6, 7, 8, 9; one is missing S 0.1. So, now we calculate Hamming distance between 1 and 2; what is the value Hamming distance between 1 and 2?

Student: 0.1? 0.2, 0.3.

Point?

Student: 0.3.

0.3.

Student: Hm.

This is 0.1, 0, 0.2, 0.3, 0.4; now D, H between 1 and 2 and 3; 0.8.

Student: 0.8.

Very high; so between the 1 and 3; 1.0; so, we have the different sequences; something is wrong right. So, three different sequences; so we can calculate the Hamming distance between any pairs. Now, if you see this one; which two sequences are close to each other?

Student: 1 and 2.

1 and 2 are close to each other; if you see this one; 1 and 2, they are close to each other; we get the value as 0.4. So, we can see the different sequences; then if you have any specific threshold for example, if we have threshold of 0.5; then 1 and 2 are close to each other; either 1 or 2. Then we can take 3; you can make into different clusters and from each cluster, we can take a representative data right and finally, you get the non redundant set of sequences; this one is Hamming distance.

(Refer Slide Time: 05:18)

Clustering Methods Based On Composition

Euclidean distance

$$D^E = \{\sum [Comp(1)_i - comp(2)_i]^2\}^{1/2}$$

1. ADIKLAAIKL $D^E(1,2) = \sqrt{0.01 + 0.01 + 0.01 + 0.01} = \sqrt{0.04}$

2. ADSKLAIAKA $D^E(1,2) = 0.2$

3. KILASDPQWE $D^E(2,3) = \sqrt{0.14} = 0.3..$
 $D^E(1,3) = \sqrt{0.11} = 0.3..$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 13

So, another is similar to Hamming distance; we can also get the Euclidean distance. The difference between Hamming distance and Euclidean distance is; what is the difference?

Student: Sum of squares.

Is sum of squares, so you can take composition of 1, minus composition 2; the whole squared, because we do not care about the sign. Then you have to take the root; so what is the value between 1 and 2?

Student: 0.4

That you get the difference right; 0.1, 0.2, 0.3, 0.4; then you have to take the root.

Student: 0.1.

0.1 squares; that is equal to 0.04; 0.1 square equal to 0.01.

Student: Yes.

Plus 0.01 plus 0.01 plus 0.01 and take the root; this is equal to 0.04.

Student: Point.

0.2; this D^E of 1, 2; how about D^E of 2, 3? 2 and 3 is 0.09; 0.01; here it is 0, here it is 0.01; 1, 2 3 times. So; that means, 0.01 into 5; that is equal to 0.09 and 0.05, this is equal to 0.14; so, root of 0.14; this is equal to 0.4? 0.3?

Student: 0.3 something

0.3 something; so, likewise D^E of 1,3; what is 1,3?

Student: 0.4... 0.04

0.04; is 0; this is.

Student: 0.01.

1.

Student: 2

2, 3, 4, 5, 6, 7 times; 0.01 into 7; so, 0.11; so 0.11 is equal to 0.3 something. So, then this is the closest one? Here also you can find distance you can see this is the closest among these three sequences. So, you can make any pairs and you can make any different clusters, you can do any cutoff; I can do that. This here we used the amino acid

composition as a feature; likewise we can use any feature to assign the clusters and you can obtain the non redundant sequences.

So, how to install these CD HIT? This is freely available and this program is very easy to install and will work with this Linux operating systems.

(Refer Slide Time: 08:37)

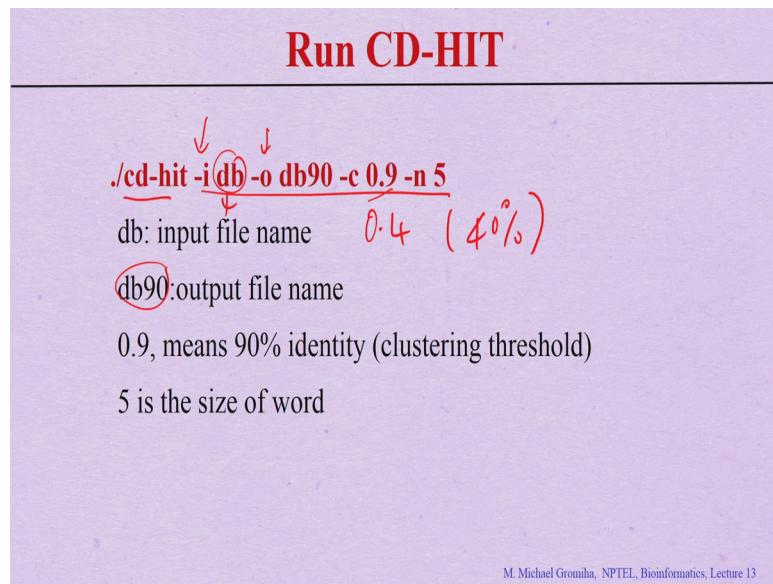
CD-HIT Installation

- **Unpack** the file with
“**`tar xvf cd-hit-v4.5.4-2011-03-07.tgz --gunzip`**”
- **Change directory** by “**`cd cd-hit-v4.5.4-2011-03-07`**”
- **Compile** the programs by “**`make`**”
- **Run** the program

M. Michael Grombilka, NPTEL, Bioinformatics, Lecture 13

So, first you download the files and then unpack using the tar and then you have to go to your directory. And you can install the program by using make, right and you can run the program. So, first it requires to download the program, you need to unpack and then to compile and you got to run the program. This is the steps we use when compelling the program right; in the UNIX systems.

(Refer Slide Time: 09:02)



So, how to run CD HIT? Because this is the program that you put dot slash CD HIT and this is the command we have to give. So, what is for i? That is input; we give this input file name; the db gives the input file name and what is o?

Student: Output.

This output; this is the output file name; db 90 is the output file name and here we give the c; 0.9 and n minus 5; what is 0.9?

Student: (Refer Time: 09:31)

This identity; so, which identity; you require whereas, 90 percent or 80 percent or 40 percent; if you need 40 percent; what you have to give?

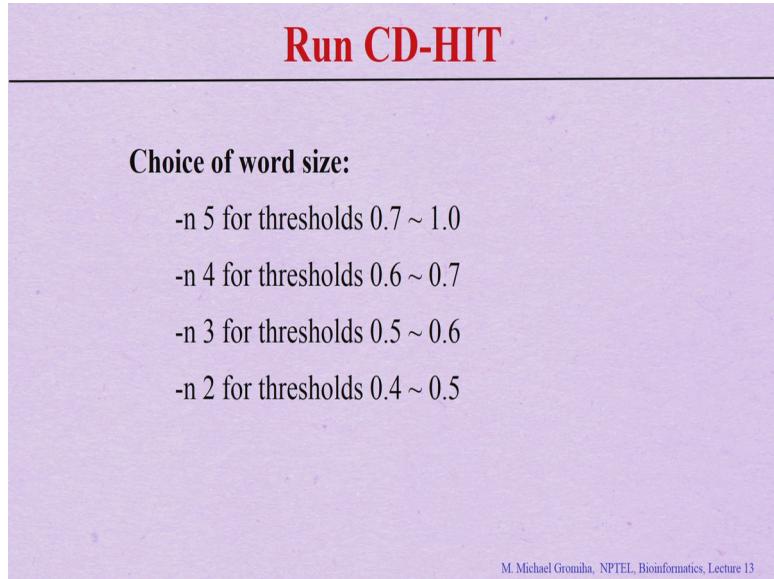
Student: 0.4

0.4; for example this is 40 percent or n equal to 5; what is n equal to 5?

Student: Word size.

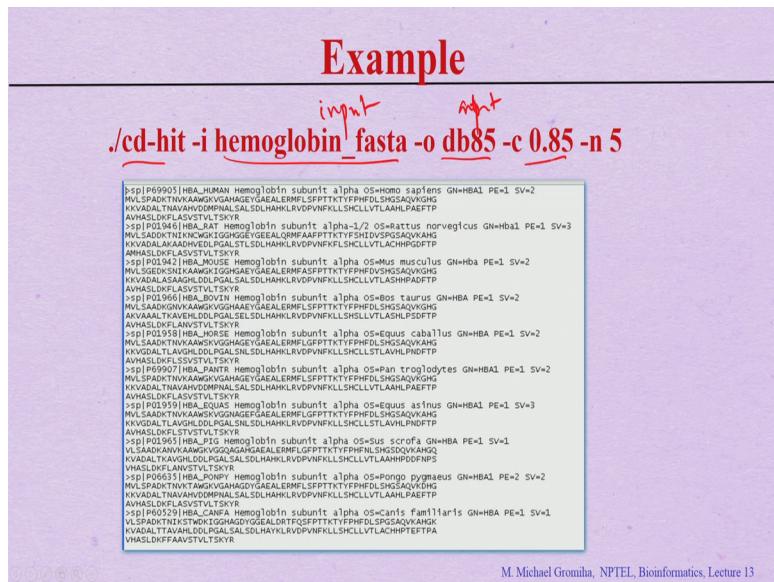
Word size; so in this case we look for the pentapeptides; how many times each pentapeptides occurs within those sequences? So, depending upon your sequence identity and the word size, so the CD HIT will run and then give you the sequences.

(Refer Slide Time: 10:04)



For example n; this is a choice, they use the default say n equal to 5 for the thresholds 0.7 to 1; n equal to 4; they use 6 to 7 and 3 for 5 to 6; this is normal threshold we use for running the CD HIT.

(Refer Slide Time: 10:18)



So, now we give the sequences; so, we have all the sequence in fasta format and we run CD HIT; this is the input file, and here this is the output. So, with 85 percent sequence identity and using the word size n; so, you can use any word size right here n equal to 5. So, then you see you will get the final data.

(Refer Slide Time: 10:40)

```
[gromiha@INSIGHT1 cd-hit-v4.5.4-2011-03-07]$ more db85  
>sp|P69905|HBA_HUMAN Hemoglobin subunit alpha OS=Homo sapiens GN=HBA1 PE=1 SV=2  
MVLSPADTKTKVAAAGHGTVGAHAEGEYGEAELERFLSFFPTIKTYFHFPLSRLGCSAQVKGHG  
KVKVADALTNAVAHVVDQNPWALSAISDLHAEKLRLVDPVNFKLLSHCLLVTLAALHLPAEFTP  
AVHASLDKFLASVSTVLTSKR  
>sp|P01946|HBA_RAT Hemoglobin subunit alpha-1/2 OS=Rattus norvegicus GN=HBA1 PE=1 SV=3  
MVLSDADTKTKVAAAGHGTVGAHAEGEYGEAELERFLSFFPTIKTYFHFPLSRLGCSAQVKGHG  
KVKVADALAKAADHVEDLPGALSTLSDLHAEKLRLVDPVNFKLHSCLLVTLAACHHPPGDFTP  
AMHASLDKFLASVSTVLTSKR  
>sp|P01965|HBA_PIG Hemoglobin subunit alpha OS=Sus scrofa GN=HBA PE=1 SV=1  
VLSAAADKVKAAUAGKVGQAGAAGAEALERMFICGFTPTIKTYFHFPLSRLGCSAQVKGHG  
KVADALTTAVAHLDLPGALSAISDLHAEKLRLVDPVNFKLHSCLLVTLAAHHPDDYNPS  
VHASLDKFLASVSTVLTSKR  
>sp|P60529|HBA_CANFA Canine hemoglobin subunit alpha OS=Canis familiaris GN=HBA PE=1 SV=1  
VLSADPKTMIEKSTWKIGCHAGDGYGGCEALDRTFQSFPTIKTYFHFPLSRLGCSAQVKGHG  
KVADALTTAVAHLDLPGALSAISDLHAEKLRLVDPVNFKLHSCLLVTLAACHHPTEFPA  
VHASLDKFLASVSTVLTSKR  
[gromiha@INSIGHT1 cd-hit-v4.5.4-2011-03-07]$  
Connected to 10.93.219.140  
SSH2 - aes12
```

Now, you will get the data then you can use this data for any analysis. And now you have two sets of data; one is the redundant sequence, one is non redundant sequence. If you compare your composition or your any properties; you can see the difference between the non redundant ones as well as redundant ones.

Also it depends upon; how much your data is reduced. If you can see much reduction in your data sets, then you can see much deviation from the properties; what you calculate from the whole data set or from the reduced data set. There is not much redundancy, that already your data set is non redundant; then you cannot see much difference. Then we will get similar data; when you analyze with a whole set of data or with the reduced set of data.

(Refer Slide Time: 11:27)

Blastclust

- **Blastclust** is a program within the standalone BLAST package used to cluster either protein or nucleotide sequences.
- The program begins with pairwise matches and places a sequence in a cluster if the sequence matches at least one sequence already in the cluster.
- In the case of proteins, the blastp algorithm is used to compute the pairwise matches.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 13

So, there is another program; this is also widely used program one of the applications of Blastclust is you can get the data with any redundancy. Even if you want to have 20 percent identity or 25 percent identity; you can use with this Blastclust. This is the standalone program which is available within BLAST and you can use this to cluster either protein sequence or nucleotide sequence. You can use it for both protein as well as the DNA sequences to get the clusters.

So, how it works? It must begin with the pairwise match and then places the sequence in a cluster. If the sequence matches at least one with the other, if two sequences have high identity; then if you are interested to threshold at 25 percent; which is more than 25 percent, you put one in the clusters. In the case of the proteins, they used a blastp for the alignment; for the case of nucleic acids, which program they use?

Student: blastn

blastn. So, you take the blastn to align the sequences.

(Refer Slide Time: 12:31)

Blastclust

The general command to create a set of non-redundant set of protein sequences is

*blastclust -i infile -o outfile -p T -L .9 -b T -S 95, 7A AITLVIKS
ITVAIKS*

where infile and outfile are input and output files, respectively.

T stands for protein;

The coverage of the length and sequence identity cutoff are 90% (-L .9) and 95% (-S 95), respectively.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 13

Then you get, this is a command; which you can use to extract the data. You can first get the program using the standalone program, you can get non redundant sequences using this sequence. So here it's the same; i is the input file, so infile is the input file and outfile is the output file; then we get some numbers -p T; -L .9; -b T, -S 95.

So, what this T stands for? Protein, for protein say put T. For example, they have the L; L equal to 0.9 and what is a meaning of L equal to 0.9?

Student: coverage

It is a coverage; so, its coverage is about, that is 90 percent; what is the meaning of coverage?

Student: Sequence coverage.

Sequence coverage you know for example, if we have one sequence; this is AITLVIKS; now you can see aligned only some part of the sequences; for example, aligned only here, in this case what is the query coverage?

Student: 80? 70 percent.

Coverage equal to?

Student: 70 percent.

70 percent; so, this is how much coverage you can accept. And then you can see this 95; what is S 95?

Student: Identity.

Identity; you can say this about 95 percent sequence identity. So, here you can restrict with a sequence identity plus the coverage; what is the advantages of using coverage?

Student: how much you have.

For example, if you have 1 proteins; 100 residues, another proteins 200 residues. Only some part; only 10 residues are aligned and 10 residues are same. If you do not use your query coverage, the aligned 10 sequences and 10 sequences are same. So, what will be the identity?

Student: 100.

100 percentage, but you see the query coverage out of 100; with aligned only 10. So, coverage only 10 percent and that means; all the 90 are totally different. In this case, they are different from each other; so to include this information; they use the query coverage in this program. Right now, we discussed about two different algorithms or the two different programs.

Student: CD HIT.

CD HIT.

Student: Blastclust.

(Refer Slide Time: 14:49)

PISCES

- PISCES is a protein sequence culling server to produce subsets of non-redundant sequences using Protein Data Bank entries or Uniprot sequences in FASTA format.
- Sequence identities for PDB sequences are determined by the combination of Combinatorial Extension **structural alignment** and **PSI-BLAST alignment**.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 13

And the blastclust; so, there is another program that's called PISCES. This is a protein sequence culling server; to get the subsets of non redundant sequences. Here we have different options; either you can use the protein sequences already available in this server; mainly from the PDB. Or you can give your own sequence; for example, if you only give your own sequence, you can give the UniProt sequences in fasta format. If you want to compare with the existing ones; then you can use the existing PDB data and you can compare and you can get any non redundant sequences.

So, how they get this non redundancy? If you have the PDB sequence; they use a structural alignment and the PSI-BLAST alignment. So what is PSI-BLAST alignment? It is a kind of multiple sequence alignment, there is a position; specific iterated BLAST alignment. So, they have the multiple sequence alignment by putting same sequences or similar sequences at the first. So, they have this sequence alignment as well as a structured based sequence alignment; they are combined together, if you have the PDB entries.

(Refer Slide Time: 15:54)

PISCES

- Non-PDB sequences are culled with **sequence identities from PSI-BLAST**. PISCES does not search the non-redundant sequence database, but rather use the user's input sequences as the database.
- This server will usually be used to cull a related set of sequences, for instance those from a PSI-BLAST search.
- It takes the amino acid sequence in FASTA format and sends the list of non-redundant protein sequences by e-mail.

→ <http://dunbrack.fccc.edu/pisces/>

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 13

If they do not have this PDB entries; for example, non PDB; then they get this sequence identity only from PSI-BLAST; it is because structures are not known. So, they have these PSI-BLAST sequences, they do the homologues sequences and this will give you the values and the non redundant sequences using only the sequence information.

So, this you can use this to cull any protein sequences; I did various ranges of sequence identities. So, what is the input sequence for this PISCES server? It takes amino acid sequence in fasta format as the input and then give the non redundant sequences. So, this is the web server; so where you can submit your query sequences and you will get the output either by email or you can get download.

If there is less number of data; you can get it immediately, but the problem with this file server is it can handle only less number of sequences. There is some limitations about the query sequence, when in the case of CD HIT or theBlastclust; you can get any number of sequences; it can handle huge amount of sequences. So, that is a difference between the CD HIT as well as with the PISCES servers.

(Refer Slide Time: 17:03)

The screenshot shows the PISCES homepage with a purple header containing the word "PISCES". Below the header is a black sidebar with white text. The sidebar lists several bullet points about PISCES's capabilities, such as its ability to provide meaningful results at low sequence identity (15-30%) and its use of PSI-BLAST. It also mentions the creation of culling lists from the PDB's Data Uniformity Site and the weekly update of PDB sequences. At the bottom of the sidebar, there is a red arrow pointing right next to the text "Access the server to create your own lists" and "Download precompiled CulledPDB lists and standalone PISCES and databases".

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 13

So, this is the online tool; so if you want to get the sequences which are already available in the server; you can get it. Or if you want to create your own data set; so, for example if you have 1000 sequences or you want to get the non redundant sequences; in this case you can get this, access the server; to create our own list.

(Refer Slide Time: 17:26)

The screenshot shows the PISCES interface for selecting a task. At the top is a logo with the word "PISCES" inside a stylized fish shape. Below the logo is the question "What do you want to do?". A list of options follows, each preceded by a radio button. The fourth option, "use the fragments of sequences from the Sbjct. line in the BLAST output which you will upload", has a red circle around it, indicating it is selected. At the bottom of the list are two buttons: "Submit" and "Reset". A red arrow points to the "Submit" button.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 13

If you click on that; this will ask for the different options, which one you want. So, you can if you have your own sequence; you can keep the last one; here you can see your own file of sequences then if you submit.

(Refer Slide Time: 17:39)

The screenshot shows the PISCES web interface. At the top, it says "PISCES" and "Please browse or paste FASTA format sequences or BLAST/PSI-BLAST output file". Below this is a text area containing a sequence of amino acids: "I-479237|Gonbank|Outer membrane integral membrane protein|outD protein MLLLSGSVLLMASSLAWSAEFSASFKGTDIQEFINTVSK NLNKTVIIDPSVSGTTITVRSY DMNNEEQYQQFFLSVLDVYGETVIPMDNNVLKIIIRSKDA". There is also a "Browse..." button for uploading files. Below the sequence area, there is a section titled "Set sequence identity threshold:" with three input fields: "Maximum percentage identity: 25" (circled in red), "Minimum chain length: 40", and "Maximum chain length: 10000". A red arrow points to the "Submit" button. At the bottom right, it says "M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 13".

Then you can get this next page; here you have to paste or type your fasta sequence in this text box. You have the sequences; you can give the text box here or you can upload file, they have the upload file option; you can browse from your computer and you can upload the file which contains amino acid sequences in which format?

Student: Fasta.

Fasta format because it accepts fasta format of any UniProt sequences. Then you can also give your options; what is the maximum percent identity you want? 25 percent or 30 percent or 40 percent or right; so, you can give the threshold as per your convenience; depending upon the problem, depending upon initial data sets and your requirement; you can set the threshold.

Normally we get 25 percent as a threshold; then they asking about the chain length, whether we need to consider all the sequences or discard some short sequence and large sequence. So, here we can give 40 or 50 because several proteins are more than 50 residues; only few exceptions. So, in this case you can give the minimum length; why do you needing the minimum length? This can avoid considering the peptides; there are many short peptides. In this case you can avoid these peptides from these not redundant data set and maximum chain length that also you can specify; for your calculation purpose; then it takes all these information as input.

(Refer Slide Time: 19:08)

Sequence id	length
7467903 Genbank Outer membrane integral membrane	456
QFVQARSGMPMNVGLQINRNSNTKYNKVNGLGTFPRFTATSGLITI	783
EQLPFLPQWVLSAALAAKAAKAAKAAKAAKAAKAAKAAKAAKAAKAAKAA	785
ENLTDVY	295
LAAYADPGRKAGKTFKVEIANDTNGEGLPLP1SLDEFTKAAAHANFDI	452
HEARHHRVYTTSALWMKGGLLKVQSVLLLSGGTNNWQEVVQINNRR	652
YTTPLWVNTT1T1YYEGGQYTRIRISANWNTNSVQLDQSAIKXIAVGVS	889
HEPCTPQKTRVJAVANQVQAIAQSYNGVQEVTSVSTLLOVDAEQKLY	201
EMKTCVHNNSVNLASVQKLLTGELTAKKLKVYFVYEEBPNLKK	260
KMF1G	292
> 11559475 Genbank Outer membrane integral membrane	5640161 Genbank Outer
MTHNPFM'W'MERIAATTLLVAGTQVAAVQVTLNWNADIDQVKAIGAT	5640161 Genbank Outer
GKTTLVQ	P13949 SwissProt Outer
WPEAKNSKQGPFPTTQDRAFQASRSQVTVVQVQVQVQVQVQVQVQVQ	P10170 SwissProt Outer
IPRNNTVATAFANNTIVTDQAGNVRVIA1C1SISQDAGAQVQVQVPLRNA	292
NAIDLAAQALQMDLGFGIGHEDUTLWVNTDOPPTNLLRLRANRNLRAA	7208425 Genbank Outer
AKEFLVWQQLDAPAEVFGNDVYLRLRNLVLAQNTLRLMLGKNDGSGS3A	560
BHDNQVQ	15598604 Genbank Outer
GOOLGTQVQ	891
GOOLGTQVQ	13470835 Genbank Outer
VIGQDQ	794
VIGQDQ	1197903 Genbank Outer
VIGQDQ	855
VIGQDQ	12620518 Genbank Outer
VIGQDQ	200
VIGQDQ	P31600 SwissProt Outer
VIGQDQ	999
VIGQDQ	15596468 Genbank Outer
VIGQDQ	616
VIGQDQ	P15727 SwissProt Outer
VIGQDQ	482
VIGQDQ	15229 Genbank Outer
VIGQDQ	778
VIGQDQ	P16466 SwissProt Outer
VIGQDQ	1577
SDDNNJRDQ	3228547 Genbank Outer
QDQ	700
QDQ	P06970 SwissProt Outer
> 14653 Genbank Outer membrane integral membrane	812
MRHNSVVKPFRFHUTTEIATYCLLALAPAAJNTVQSNLTDQGLP7IAGO	P22420 SwissProt Outer
SHDOWLQFKEKSLDQLESLKQEEDISWPAFSLKLMQASLVLVKFGQKL7IELL	505
VQELEALSLFALMEFEFWNTSAQAFPSRPMFPPFAAGSGQWMPFNEI	565
VQELEALSLFALMEFEFWNTSAQAFPSRPMFPPFAAGSGQWMPFNEI	P35077 SwissProt Outer
VQELEALSLFALMEFEFWNTSAQAFPSRPMFPPFAAGSGQWMPFNEI	584
VQELEALSLFALMEFEFWNTSAQAFPSRPMFPPFAAGSGQWMPFNEI	12721580 Genbank Outer
VQELEALSLFALMEFEFWNTSAQAFPSRPMFPPFAAGSGQWMPFNEI	444
VQELEALSLFALMEFEFWNTSAQAFPSRPMFPPFAAGSGQWMPFNEI	654
VQELEALSLFALMEFEFWNTSAQAFPSRPMFPPFAAGSGQWMPFNEI	1197903 SwissProt Outer
VQELEALSLFALMEFEFWNTSAQAFPSRPMFPPFAAGSGQWMPFNEI	350
GUTTPHTP952QVAFFPQKQATITLQ10NEVSQVNTVYMA3GLASRG	P06111 SwissProt Outer
MTTFAGANLFLSABMILITRTIRLNQAASEASVATLASQAEVNLFEQSQ	257
DVGLARNAFPWV1KQPFPTLFLV1QVPAFLSVVLLGLV1VSTDPRLMV	P24126 SwissProt Outer
TLCGQFQVQ	530
EGSSPQVQ	P29041 SwissProt Outer
F1L50G113LQGQSTTQ3V1LQDLP7QZDTRTSEBV1ILMTFK	759
1HEGOTDQQGPFPTTQHATDAPLQGQPFPTV	5759281 Genbank Outer
FTFLN1T1NEFPLGQTIT1NFNTFS85GAVSNPLPFG72F1V5V60E	462
QGLL1T1VPEPLGSLFSTYQ1FKEKQAJDQAQSLNPLQILTOPTLVQGE	
AAQVQLTETSV1IAYSVDTVQGQQTAVR1T1VLENOVGLLTNV1VDR10N	
QFTTLRNV1V3A9TQGTVQFQGSGANGAINE1L1NKRRELTSGVVRRLRQQT	
F1L50G113LQGQSTTQ3V1LQDLP7QZDTRTSEBV1ILMTFK	
1HEGOTDQQGPFPTTQHATDAPLQGQPFPTV	

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 13

And if you submit then this will give you the sequence, here this is the complete sequence and this is the list for the ID's. You can see the id here and here we give the complete sequences and here we can give sequence ids.

So, you can use a PISCES to get the non redundant sequences, but the only disadvantage of PISCES is; it can handle only less number of sequences then it uses a clustering technique, then they will finally, they give you the final set of non redundant sequences. So, till now what did we discuss in the class? today's lecture?

Student: Redundancy.

First we discussed about the redundancy, so what is redundancy?

Student: Over representation of sequences

Right we can see if there are same sequences; if one or more same sequences which are present in any data set. If we construct in a specific data set which consists of 100 sequences; are there any sequences or any 2 or 3 sequences which share the common information with the high sequence identity. So, if we share high sequence identity then these sequences are called redundant sequences; then how to get the non redundant sequences.

Student: (Refer Time: 20:29)

For example, if you have two sequences; different sequence identity and a third one this has the different sequence identity. If you want to have the sequences with more than 50 percent then what we have to do?

Student: (Refer Time: 20:43)

When the first two it is very high; so you can keep one and discard another one. Because if you have the same sequences that will reduce a bias in the final calculations as well as for any prediction algorithms. We checked two examples like what will happen if the same sequences are present two times or three times, depending upon the features; some of them are over represented and some of them are underrepresented, should not avoid the bias; we need to construct non redundant sequences.

So, what are the various program we discussed to construct the non redundant sequences?

Student: CD HIT

CD HIT.

Student: Blastclust

Blastclust.

Student: PISCES

And the PISCES; so, what is the advantage of using CD HIT?

Student: It can handle large dataset.

It can have the large data sets.

Student: very fast

It is very fast; so, the disadvantages of this CD HIT.

Student: upto 30 percent

It can handle up to 30 to 40 percent sequence identity because it has two versions; one is online and one is the standalone version. So, you can handle only up to 30 percent or 40

percent sequence identity based on the program you use. So, what principle it used for clustering the sequences?

Student: Hamming distance

K means clustering; to make these clusters, how this K means clustering works? Try to consider all the sequences and we can do some number of clusters. So, this K clusters and then try to put in the sequences, similar sequences within that particular cluster and they rearrange all the values again and again. So, that the minimum redundancy is maintained within that only particular cluster.

This would take any cluster, you can see there is completely close to each other then they take the representatives from different clusters; in this case you can get the minimum redundancy among your sequences. So, what are various features we use for clustering techniques? We discussed two aspects; both are based on?

Student: distance.

This amino acid composition; so, one is the Hamming distance and the Euclidean distance. So, if you have any pair of sequences; you can calculate the values and make any threshold, if the pairs are more than that threshold then you can discard one. So, in getting the non redundant sequences; some advantages, some disadvantages. For example, if we have 100 sequences; made into 10 clusters, first cluster has 20, second cluster has 5 and another cluster has 1. Then how to get the representative data; if it is 1; it is easy, you can then pick it up; if that is 10 how to you get the representative one?

Student: the common one

Generally, we will take randomly pick up one; you can get this one. This is the reason; you have to cluster again and again and you can see what will happen if you pick up these different sequences from each cluster. You can see almost they are in similar features; so you can get similar data, even then you can do this test to confirm that your data are significant or not. So, then the second one is a Blastclust; what is the advantages of Blastclust?

Student: It can also handle large data

It can also handle large data sets and you can go with the less sequence redundancy; even 20 percent, 25 percent you can do. And it also considers the coverage; query coverage, so with that you can use the Blastclust for clustering the sequences. The third one we discussed about PISCES; here it is easy to use, because if you give all the sequences on length; you will get the non redundant sequences.

But only disadvantages is; it can handle limited number of sequences. In this case, you have to do this again and again; take your sequence and get non redundancy and then make another set and get the non redundant sequences and cluster together and put again, you have to run again to get the non redundant sequences.

So, we discussed about the construction of non redundant sets and various features of the primary sequence. There are various parameters we calculated and we used for different applications. Now in the next class, we will discuss about the next level; what is next level of protein structure? Secondary structures; what are different secondary structures?

Student: alpha helices

Alpha helices beta strands coil and all; so, we will discuss about what are the different secondary structures? What are the probable allowed positions of these secondary structures in the construction of Ramachandran plot? How to predict these secondary structures from any given amino acid sequence? What is the reliability of these prediction techniques and so on; that we will discuss in the next following couple of classes.

Thank you for your kind attention.