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

Lecture - 9b
Conservation score II

There are 3 different ways to get the frequencies of amino acids; what are 3 different ways to get the frequencies?

Student: Unweighted.

Unweighted frequency.

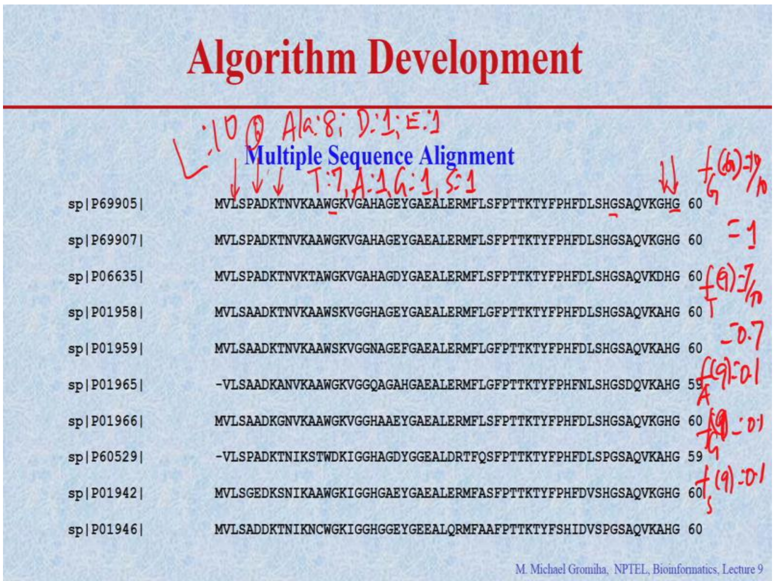
Student: Weighted.

Weighted frequency and?

Student: Independent.

Independent counts. So, fine now we have the frequencies, we need to convert these frequencies into score.

(Refer Slide Time: 00:38)



Algorithm Development

Multiple Sequence Alignment

```
sp|P69905|     MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp|P69907|     MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp|P06635|     MVLSPADKTNVKTAWGKVGAHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKDHG 60
sp|P01958|     MVLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG 60
sp|P01959|     MVLSAADKTNVKAAWSKVGGNAGEFGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG 60
sp|P01965|     -VLSAADKANVKAAWGKVGGQAGAHGAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAHG 59
sp|P01966|     MVLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp|P60529|     -VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHG 59
sp|P01942|     MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG 60
sp|P01946|     MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVSPGSAQVKAHG 60
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

So, for example, if we have this sequence this position; so what is the frequency of G at position number 60?

Student: (Refer Time: 00:45).

10 by 10, right. So, this equal to 1; we take this position; position number 9 and so if you see there are what is the frequency of these amino acids residues; how many times T occurs.

Student: (Refer Time: 00:59).

T occurs 7 times.

Student: Alanine.

Alaline one time.

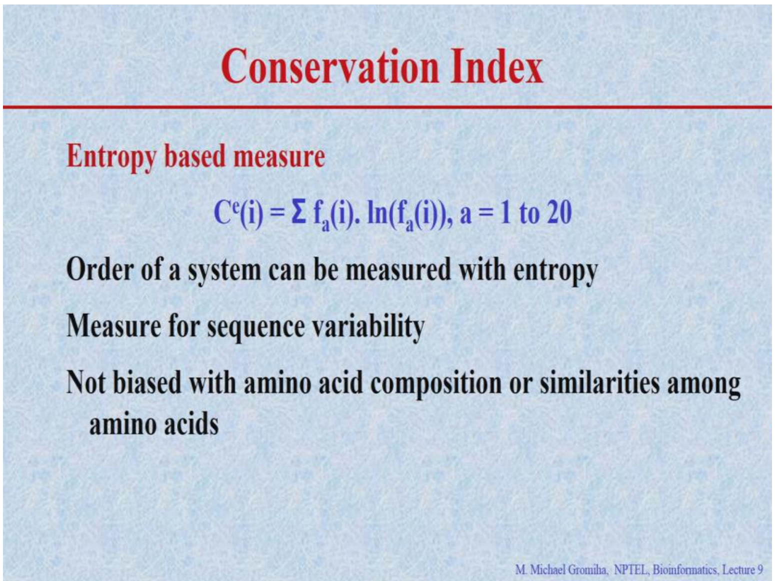
Student: Glycine one time.

Glycine one time.

Student: Serine.

Serine one time, right; so now we convert this into frequencies, right the frequency of threonine at position 9 this equal to 7 / 10 = 0.7. So, frequency of alanine at position number 9 equal to 0.1 right and the frequency of glycine at portion number 1 9; this equal to 0.1 as well as frequency of serine at portion number 9 is equal to 0.1, right to get the numbers.

(Refer Slide Time: 01:46)



Conservation Index

Entropy based measure

$$C^e(i) = \Sigma\, f_a(i).\ \ln(f_a(i)),\ a = 1 \text{ to } 20$$

Order of a system can be measured with entropy

Measure for sequence variability

Not biased with amino acid composition or similarities among amino acids

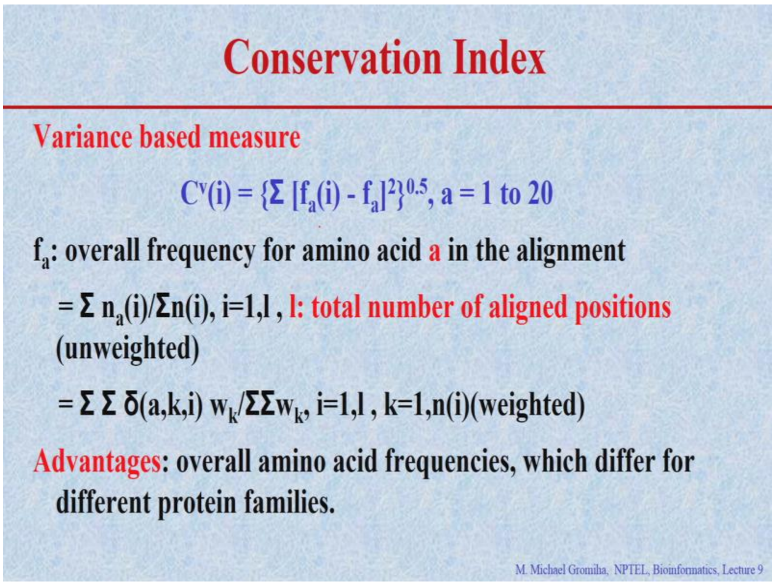
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

So, now we convert these frequencies into scores. So, for first we do the entropy based score. So, here if you see this frequency right at any position I of any specific amino acids if it a specific position is occupied by a single amino acid right you will get this $f_a(i)$ = 1 right if it is occupied by different amino acid residues $f_a(i)$ will be different depending upon the number of times each amino acid residue right occurs at particular position if they randomly distributed then the frequency of 0.05.

So, will get 0.05; other frequency, then the frequency we multiplied with the logarithmic of this frequency right to get the conservation score based on entropy based method here this is not biased with the amino acid composition or similarities among amino acids, because we do not give any bias for this. Similarity of amino acids if it is 7 2 2 for the 3 different amino acids present at the position, it does not matter if it is 7 is threonine or 7 is alanine or 7 is tryptophan. If you do not give any weight for any of the amino acids and there is another method that is called variance based method.

So, here this will consider the frequency of amino acids a same amino acid at the different positions. For example, here if you see the sequence same amino acids located different positions. For example: if you see glycine here and if your glycine here and glycine here and some cases it is highly conserved and some cases with the variable residues and look at the other residues how far they are variable. So, they compare how many positions the glycine occurs how many positions they have similar residues and what is the proportion of these residues they take into consideration when you calculate a score right.

(Refer Slide Time: 03:29)



Conservation Index

Variance based measure

$$C^v(i) = \{\Sigma\ [f_a(i) - f_a]^2\}^{0.5},\ a = 1 \text{ to } 20$$

f_a: overall frequency for amino acid a in the alignment

$= \Sigma\ n_a(i)/\Sigma n(i)$, i=1,l , l: total number of aligned positions (unweighted)

$= \Sigma\ \Sigma\ \delta(a,k,i)\ w_k/\Sigma\Sigma w_k$, i=1,l , k=1,n(i)(weighted)

Advantages: overall amino acid frequencies, which differ for different protein families.

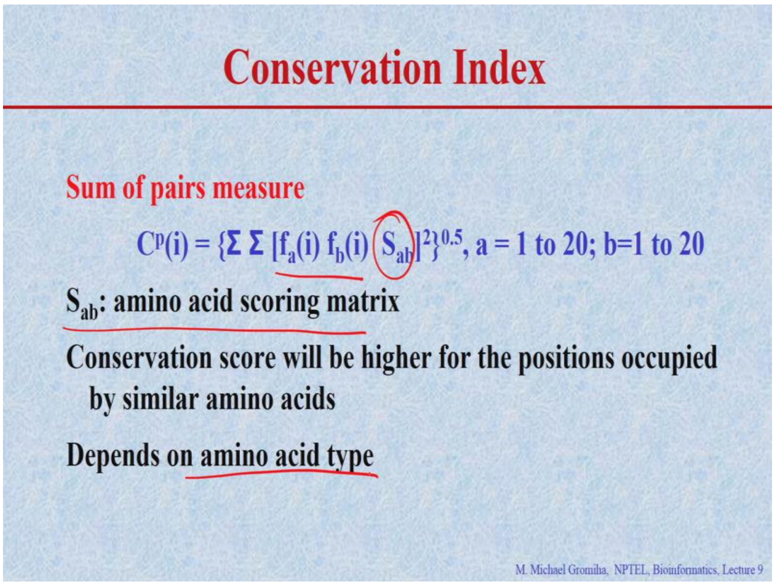
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

Now, the equation is $f_a(i)$ this is the frequency of amino acid a right in the alignment and they take the overall frequency there is f_a overall frequency of this amino acid a in the alignment.

So, and then compare this with any particular positions and then use this equation $f_a(i)$ - f_a right take the square and finally, summation overall the pairs and then take the square root then we will get this conservation score here in the frequency either we use weighted ones or they use the unweighted ones; if use unweighted ones then they get this equation. So, $n_a(i)$ divided by sigma n of i, where i = 1 to l or this is the number of aligned positions this for unweighted, if you want to go the weighted ones then we give the weightage delta a,k,i depending upon the weightage you can give this equal to 1 or this equal to 0. So, what is the advantage of having weightage?

So, here the overall amino acid frequencies which is different from different families; so in order to if you want to implement if you want to include this information then we need to give weightage for the different a protein families then you can see which one is highly conserved compared with the other ones then there is another measure this is called a sum of pairs method right here.

(Refer Slide Time: 04:44)



Conservation Index

Sum of pairs measure

$C^p(i) = \{\Sigma \Sigma [\underline{f_a(i) f_b(i)} S_{ab}]^2\}^{0.5}$, a = 1 to 20; b=1 to 20

S_{ab}: amino acid scoring matrix

Conservation score will be higher for the positions occupied by similar amino acids

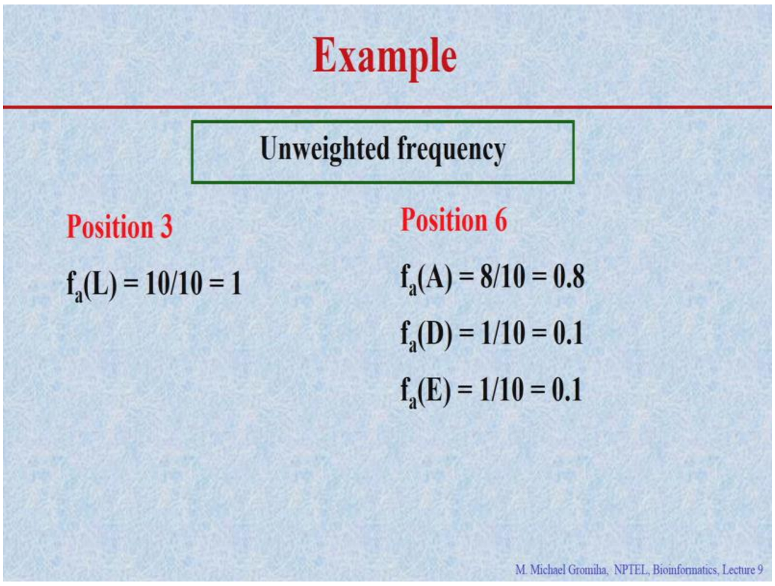
Depends on amino acid type

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

If you have this sequences you are aligned $f_a(i)$ and $f_b(i)$, then we give this scoring matrix and depending upon the aligned residues. So, which scoring matrix, we use we discussed about 2 matrixes either PAM matrix or BLOSUM matrix. So, we use these numbers a,b right where a = 1 to 20 and b also varies from 1 to 20.

So, we use any scoring matrix and then we give the weightage based on these align positions and you can calculate the score. So, this also depends upon the amino acid type if they considered position not the same residues in similar residues are present then we get high score right. So, the value will be high if the positions are occupied by similar amino acids if it is completely different amino acids like lysine and alanine then you get less score compared with the similar amino acids like lysine and arginine; so this way you get the sum of pairs method, ok.

(Refer Slide Time: 05:40)



Example

Unweighted frequency

Position 3

$f_a(L) = 10/10 = 1$

Position 6

$f_a(A) = 8/10 = 0.8$

$f_a(D) = 1/10 = 0.1$

$f_a(E) = 1/10 = 0.1$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

Now, we will see how to obtain the values. So, this is the one example I give 2 different positions, position number 3 if you take position number 3 it is occupied by leucine, I will show the sequence right if you take the position number 3 this position what is the frequency of residues at position number 3.

Student: 10 by 10.

Completely this occupied by.

Student: Leucine.

Leucine right leucine 10 times, if you take the position number 6.

Student: Alanine.

Alanine.

Student: 8 times.

Alanine 8 times.

Student: D 1 time.

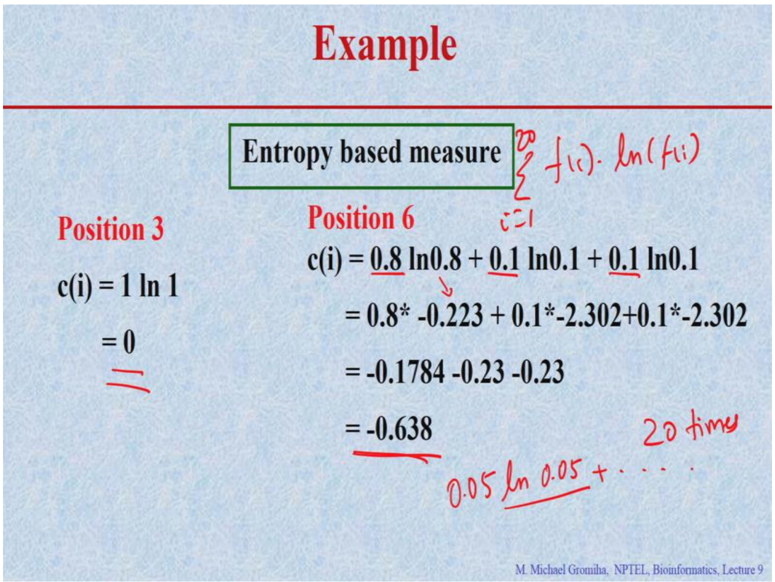
D 1 time.

Student: E 1 time.

E 1 time.

So, now you get the frequency. So, f (a) = 0.8 f (d) = 0.1 and f (e) = 0.1.

(Refer Slide Time: 06:42)



Example

Entropy based measure $\sum_{i=1}^{20} f(i) \cdot \ln(f(i))$

Position 3

$c(i) = 1 \ln 1$

$= \underline{\underline{0}}$

Position 6

$c(i) = \underline{0.8} \ln 0.8 + \underline{0.1} \ln 0.1 + \underline{0.1} \ln 0.1$

$= 0.8^* -0.223 + 0.1^*-2.302+0.1^*-2.302$

$= -0.1784 -0.23 -0.23$

$= \underline{-0.638}$

20 times

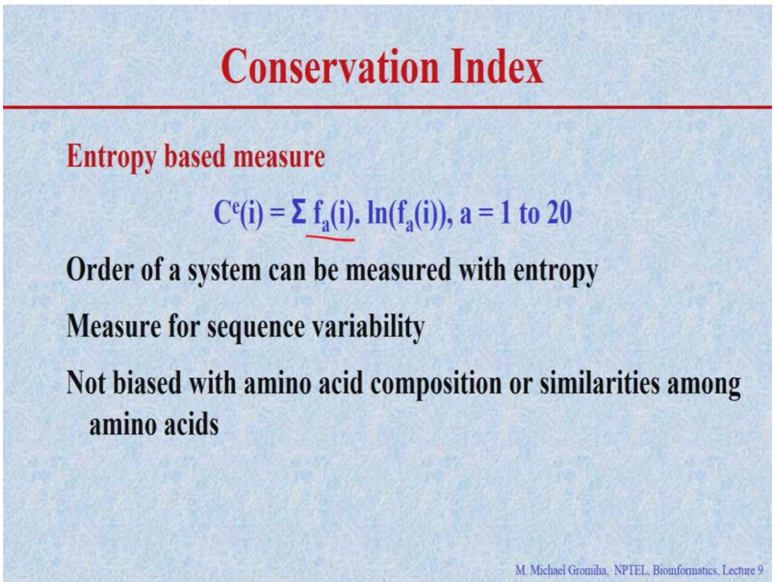
$\underline{0.05 \ln 0.05} + \ldots$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

So, if we take position number 3 a occupied only by leucine and the preference is 10 by 10 this equal to 1. Now we convert these frequencies into score. So, if we take the entropy based method; so c (i) is score 1 * ln(1) = 0 if you take the position number 6 right the equation is f (i) * ln f (i) right, i varies from i = 1 to 20.

So, here we have only 3 amino acids right because 3 residues 1 is the 0.81 is 0.1 and the 0.1. So, another conservation score we calculate ok.

(Refer Slide Time: 07:12)



Conservation Index

Entropy based measure

$$C^e(i) = \Sigma\, \underline{f_a(i)}.\ \ln(f_a(i)),\ a = 1 \text{ to } 20$$

Order of a system can be measured with entropy

Measure for sequence variability

Not biased with amino acid composition or similarities among amino acids

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

This is the equation or you can this is the equation right 0.8; ln (0.8); 0.1 ln(0.1) and 0.1 ln(0.1), right, this is equal to 0.8 into this is the logarithmic values we substitute the numbers and finally, we get the conservation score of 0.638 if 20 different amino acids are distributed randomly then what will be the distribution what will be the score.

Student: 0.05.

0.05 right 0.05 for 1 amino acid 0.05 in to logarithmic of.

Student: 0.05.

0.05; we will get how many after how many times.

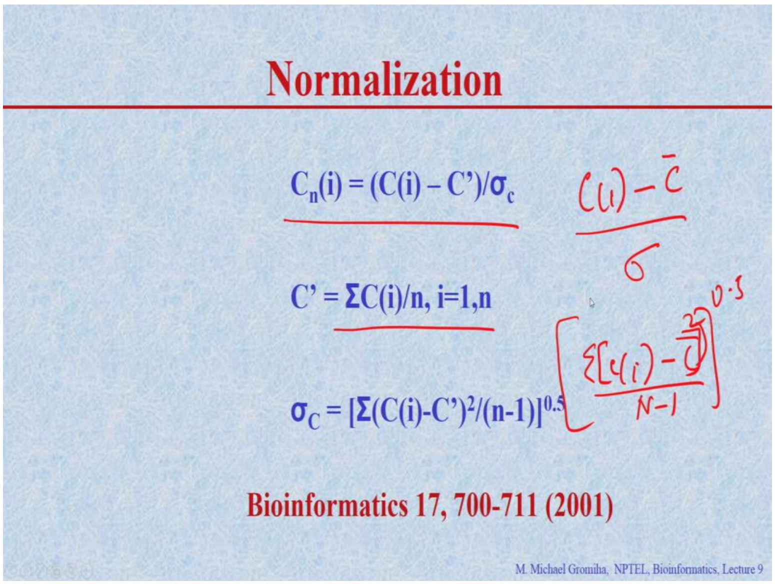
Student: 20 times.

20 times we get so multiplied by 20. So, what is logarithmic of 0.05? You get the numbers and multiplied by 0.05 and get the divide multiplied by 20, then you get the numbers. If you compare these 2 values; for example this is 0 and this 0.638 which one is conserved?

Student: 0.

Is conserved right, because if you have all the positions occupied by the same residue then you highly conserved so you can get the 0, but this if you see negative values. If it is highly variable, then it is going less because you will get more negative values. Now you can normalize these numbers because here we get the numbers maximum of 0 and minimum numbers.

(Refer Slide Time: 08:32)



Normalization

$$\underline{C_n(i) = (C(i) - C')/\sigma_c} \qquad \frac{C(i) - \bar{C}}{\sigma}$$

$$C' = \underline{\Sigma C(i)/n, i=1,n}$$

$$\sigma_C = [\Sigma(C(i)-C')^2/(n-1)]^{0.5} \qquad \left[\frac{\Sigma[C(i) - \bar{C}]^2}{N-1}\right]^{0.5}$$

Bioinformatics 17, 700-711 (2001)

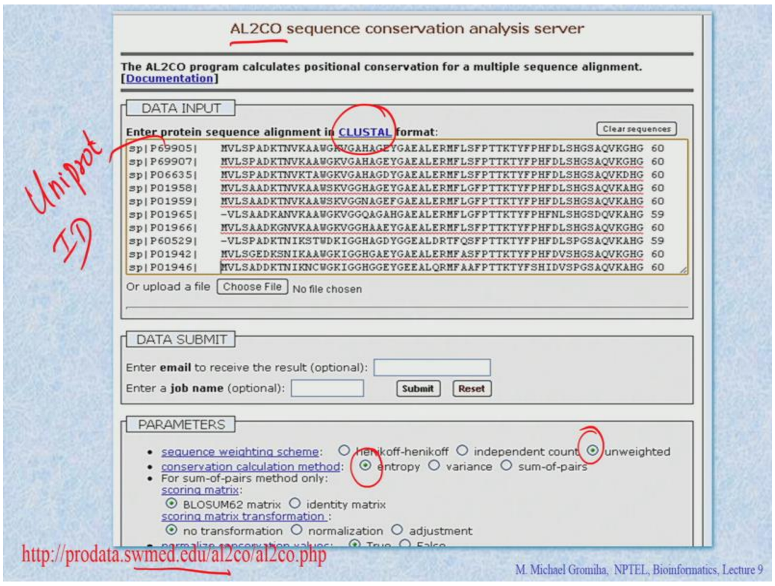
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

So, you can normalize you can normalize using the equation right see normalization of i for new positions this is equal to C (i) right conservation of particular residue right C (i) minus mean of C that is C’ this equal to summation C(i) / n the i = 1 to n number of sequences aligned in this position right and sigma what is sigma standard deviation how to get the standard deviation this equal to C (i) - mean of the C right.

So, then we divided by the square right and finally, get the 0.5 this is not square root right square this C is confuse this is C. So, C (i) - C average you have to square it right divide by n - 1 and take the square root right we will get this deviation. So, will give the C (i) - C average divide by sigma, then you will get the normalization then in this case, you will get the numbers right normalize with this respect to 1. So, you get the negative value or the positive values right with the respect to this one.

So, now this is the method you can calculate the conservation score.

(Refer Slide Time: 09:59)



AL2CO sequence conservation analysis server

The AL2CO program calculates positional conservation for a multiple sequence alignment. [Documentation]

DATA INPUT

Enter protein sequence alignment in CLUSTAL format:

```
sp|P69905|    MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp|P69907|    MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp|P06635|    MVLSPADKTNVKTAWGKVGAHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKDHG 60
sp|P01958|    MVLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG 60
sp|P01959|    MVLSAADKTNVKAAWSKVGGNAGEFGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG 60
sp|P01965|    -VLSAADKANVKAAWGKVGGQAGAHGAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAHG 59
sp|P01966|    MVLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG 60
sp|P60529|    -VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHG 59
sp|P01942|    MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG 60
sp|P01946|    MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVSPGSAQVKAHG 60
```

Or upload a file Choose File No file chosen

DATA SUBMIT

Enter **email** to receive the result (optional):

Enter a **job name** (optional):

Submit Reset

PARAMETERS

- sequence weighting scheme: ○ henikoff-henikoff ○ independent count ● unweighted
- conservation calculation method: ● entropy ○ variance ○ sum-of-pairs
- For sum-of-pairs method only:
  scoring matrix:
  ● BLOSUM62 matrix ○ identity matrix
  scoring matrix transformation :
  ● no transformation ○ normalization ○ adjustment
- normalize conservation values: ● True ○ False

http://prodata.swmed.edu/al2co/al2co.php

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

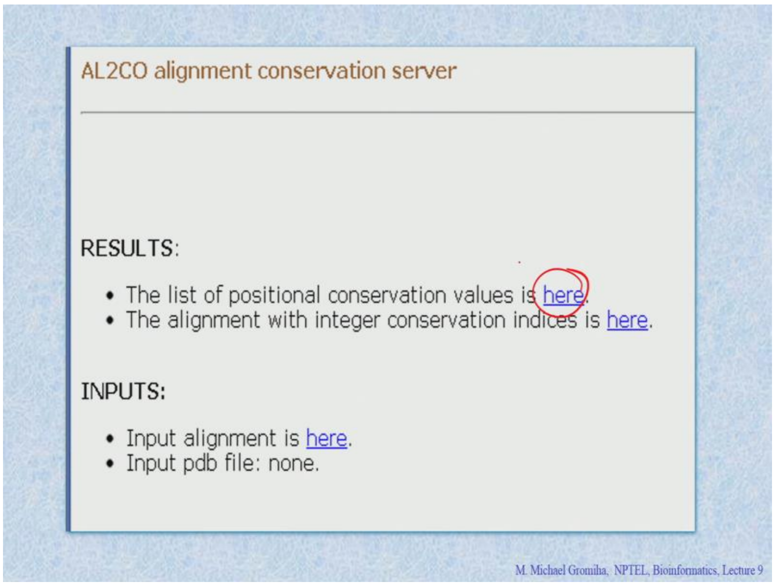
And we have several algorithms available to get the score based on all these method methods for example one of the publicly available method is AL2CO that is alignment to conservation these accepts the multiple sequence alignment in CLUSTAL format; what is CLUSTAL format.

Student: identify and then sequence.

(Refer Time: 10:13) identify and the sequence if you see program for multiple sequence alignment if we give these single sequences and if align using CLUSTAL, then you will get the output this is the identifier this is the UniProt code UniProt id and here you give sequence. So, we give this one and there are various ways you can select this is what we discuss we discussed about the different frequencies either we use independent count or we use unweighted or give weighted and the conservation score we have different ways you can use entropy based method and the variance based method and the sum of pairs method. We discuss one example based on unweighted frequency and entropy based method. So, just for the verification I put the unweighted frequency and the entropic based method.

So, here this is the website you can get the AL2CO method to get the conservation. So, if we click we get the submit button.

(Refer Slide Time: 11:12)



AL2CO alignment conservation server

RESULTS:

- The list of positional conservation values is here.
- The alignment with integer conservation indices is here.

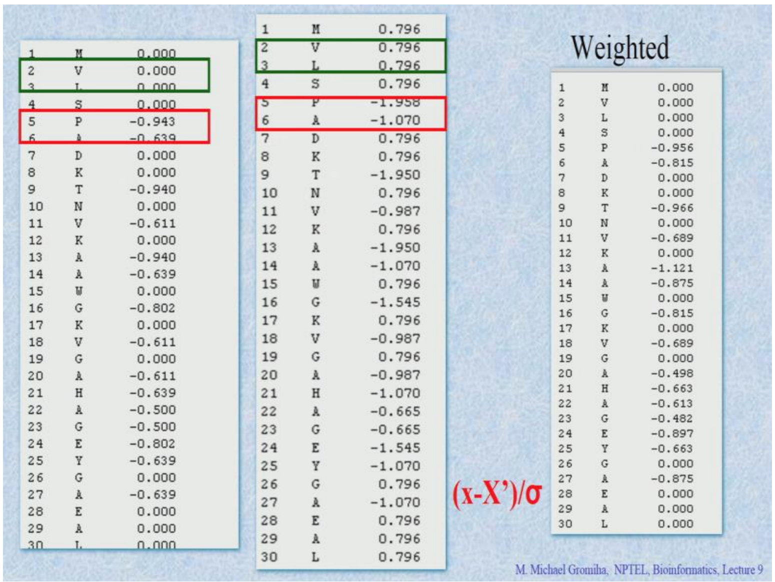
INPUTS:

- Input alignment is here.
- Input pdb file: none.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

You will get the data right which was asking for the positional conservation or the alignment with the individual conservation right will get the values or we can classify into different groups into different numbers. So, we want to check your input this is input alignment is here if you click on here you will get the input alignment or you can get the positional conservation here.

(Refer Slide Time: 11:29)



```
1      M      0.000
2      V      0.000
3      L      0.000
4      S      0.000
5      P     -0.943
6      A     -0.639
7      D      0.000
8      K      0.000
9      T     -0.940
10     N      0.000
11     V     -0.611
12     K      0.000
13     A     -0.940
14     A     -0.639
15     W      0.000
16     G     -0.802
17     K      0.000
18     V     -0.611
19     G      0.000
20     A     -0.611
21     H     -0.639
22     A     -0.500
23     G     -0.500
24     E     -0.802
25     Y     -0.639
26     G      0.000
27     A     -0.639
28     E      0.000
29     A      0.000
30     L      0.000
```

```
1      M      0.796
2      V      0.796
3      L      0.796
4      S      0.796
5      P     -1.958
6      A     -1.070
7      D      0.796
8      K      0.796
9      T     -1.950
10     N      0.796
11     V     -0.987
12     K      0.796
13     A     -1.950
14     A     -1.070
15     W      0.796
16     G     -1.545
17     K      0.796
18     V     -0.987
19     G      0.796
20     A     -0.987
21     H     -1.070
22     A     -0.665
23     G     -0.665
24     E     -1.545
25     Y     -1.070
26     G      0.796
27     A     -1.070
28     E      0.796
29     A      0.796
30     L      0.796
```

(x-X')/σ

Weighted

```
1      M      0.000
2      V      0.000
3      L      0.000
4      S      0.000
5      P     -0.956
6      A     -0.815
7      D      0.000
8      K      0.000
9      T     -0.966
10     N      0.000
11     V     -0.689
12     K      0.000
13     A     -1.121
14     A     -0.875
15     W      0.000
16     G     -0.815
17     K      0.000
18     V     -0.689
19     G      0.000
20     A     -0.498
21     H     -0.663
22     A     -0.613
23     G     -0.482
24     E     -0.897
25     Y     -0.663
26     G      0.000
27     A     -0.875
28     E      0.000
29     A      0.000
30     L      0.000
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

If you go with a positional conservation this is at the different amino acid residues we discussed about few positions.

So, remember with positions we discussed.

Student: 3.

3. So, 3 is the value of leucine.

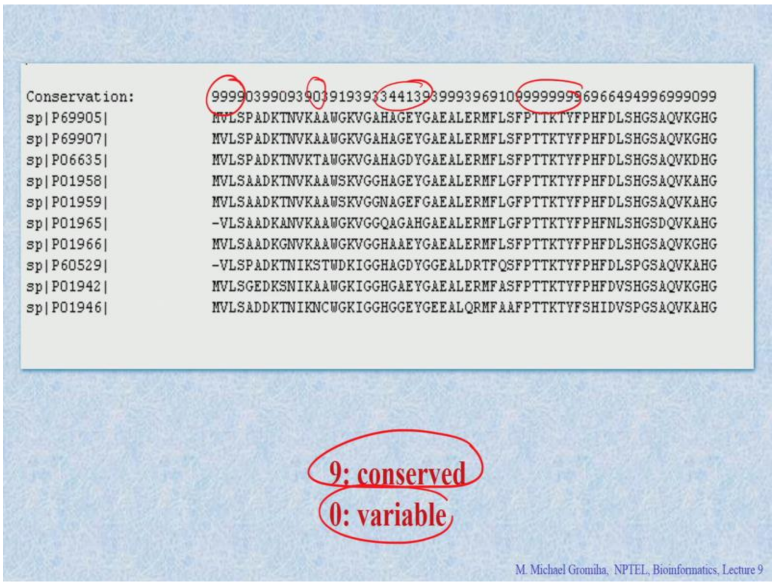
Student: (Refer Time: 11:42).

Right we get the 0, right, here this is the value we get is equal to 0. So, it matched with the number. So, you can see 3 this equal to 0, then we check the position number 6 that is alanine right if the position number 6 we checked the values, if we get -0.638 here also we see in this program we get -0.639, then we checked position number 9. So, this is equal to -0.940 because there is more variable than the position number 6.

So, just I show the sequence if you see position number one 3 is highly conserved 60 is conserved and position number 6 that is 8 plus 1 plus 1 and position number 9 7 plus 1 plus 1 plus 1. So, it is more variable. So, if you look into the answers. So, you can see the distribution of these residues based on conservation right you can see this is 0 highest one and 6 is variable and 9 is again is more variable right. So, you can see the numbers.

So, you can get the numbers into normalized one as I discussed earlier. So, this is normalized one you can use this equation to normalize the score and get the normalized values. So, this is the data which we obtained with the weighted matrix if we use the weighted matrix. If you see here the weighted method right we will get these numbers, but if you compare these 2 if you have this similar set of sequences because in this case we use 10 sequences they are from the similar sequences right similar family of sequences this is why if you see there is a good correlation between this number and this number 0s 0s 0.9; 0.9, 0.6, but you have 0.8. You can see the correlation between the unweighted and the weighted frequencies if you use completely different sequences then you can see a difference between the score obtained with the weighted frequencies as well as the unweighted frequencies.

(Refer Slide Time: 13:33)



```
Conservation:       999903990939039193933441393999396910999999996966494996999099
sp|P69905|          MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
sp|P69907|          MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
sp|P06635|          MVLSPADKTNVKTAWGKVGAHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKDHG
sp|P01958|          MVLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
sp|P01959|          MVLSAADKTNVKAAWSKVGGNAGEFGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHG
sp|P01965|          -VLSAADKANVKAAWGKVGGQAGAHGAEALERMFLGFPTTKTYFPHFNLSHGSDQVKAHG
sp|P01966|          MVLSAADKGNVKAAWGKVGGHAAEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHG
sp|P60529|          -VLSPADKTNIKSTWDKIGGHAGDYGGEALDRTFQSFPTTKTYFPHFDLSPGSAQVKAHG
sp|P01942|          MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHG
sp|P01946|          MVLSADDKTNIKNCWGKIGGHGGEYGEEALQRMFAAFPTTKTYFSHIDVSPGSAQVKAHG
```

9: conserved

0: variable

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

So, now the next aspect is we get some numbers right. So, here we get several numbers 0.00 -0.94 -0.63 and so on how to convert into a numerical scale, because here we get sequence for each sequence we need to assign some numbers. So, I want to do it between 0 and 9 if you see this highly conserved is 9 right 9 is highly conserved and to see the variable it is 0, then they automatically normalized and put the numbers for the each sequence this is the a query sequence. So, I take this as the query one and for this one they put the numbers here these are the places where it is highly conserved and these are the region it is very very highly variable right you can see numbers and this numbers you can use as an input for several predictive algorithms that we will discuss in the later classes this is one method.

Then there are several methods which can calculate the conservation score right because why we are concerned about conservation score this will give you the structural important positions this will give you the functional important positions and these positions. They try to maintain in different sequences means if you alter these residues it will have adverse effects.

So, this is a reason why the conservation score is one of the features right for the predictions second aspect to calculate the conservation score we do not need any structure information we need only the sequence information if you gather the information regarding homologous sequences you can calculate the conservation score from the sequence right getting sequence information that is easy because you have more number of sequences then structures right how many sequences in the UniProt database now.

Student: 79.

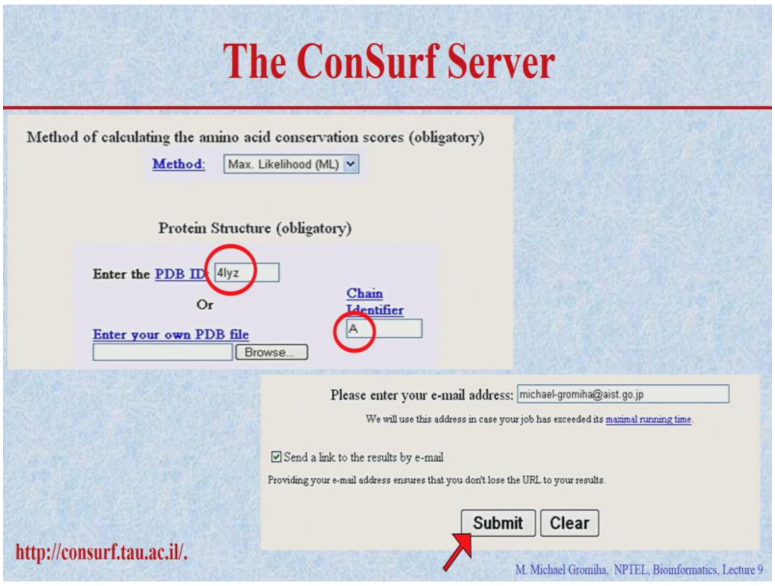
So, 78 million sequences. So, you have more number of sequences. So, given if you have homologous sequences you get sufficient number of sequences for the conservation approximately how many sequence are required to calculate the conservation score.

Student: More than 2.

Right if you get 3 sequences you can get this conservation, but is it reliable or not no right. So, if you get the reliable values. So, at least we need to have more than 100 sequences if you have more number of sequences we can see the variability if you have less number of sequences. Now we discuss with 10 sequences many positions here we get the same residues here if this is the sequence I gave only 10 sequences and if you see G is same at 60 and same at this position right and you see H is same in the position of 59 its mainly because less number of sequences if you go through 100 sequences the probability of glycine in all the 100 sequences is very less compared to the probability of glycine in the 10 sequences, right.

So, if you go with the more number of sequences even if you take 100 sequences and all the 100 sequences if glycine occupies at 6 positions, then the data is significant you can say that glycine is very important at position number 60; so in this case to make sure right for the reliability of your results. So, you should have at least 100 sequences for calculating the conservation score.

(Refer Slide Time: 16:38)



The ConSurf Server

Method of calculating the amino acid conservation scores (obligatory)

Method: Max. Likelihood (ML)

Protein Structure (obligatory)

Enter the PDB ID: 4lyz

Or

Enter your own PDB file

Chain Identifier: A

Please enter your e-mail address: michael-gromiha@aist.go.jp

We will use this address in case your job has exceeded its maximal running time.

☑ Send a link to the results by e-mail

Providing your e-mail address ensures that you don't lose the URL to your results.

http://consurf.tau.ac.il/.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

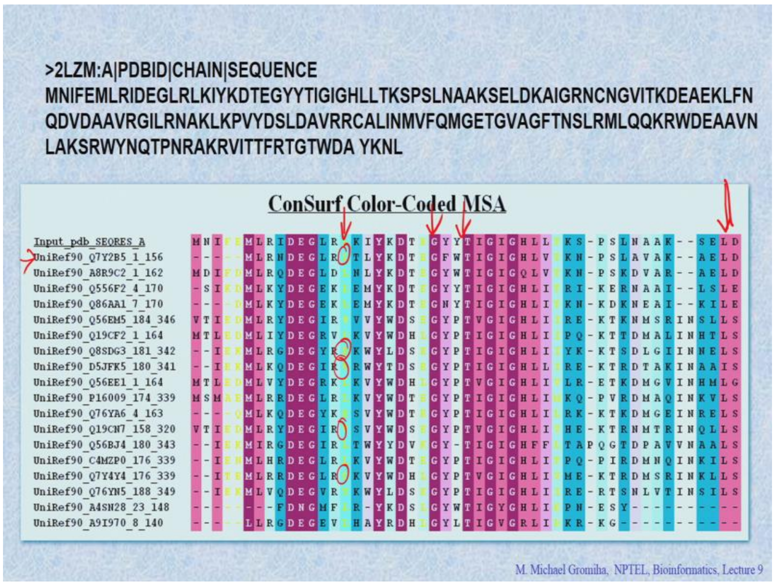
So, we discussed about AL2CO. So, now, there is another method this is called ConSurf. So, here it is easy to do that we do not have to give any sequences if the structure is known right you can give the just the PDB id if we give the identifier of the protein databank I will discuss about the protein databank in later classes right. So, then automatically this will get the sequence map to the PDB and it will get the homologous sequences from the UniProt database and they do the multiple sequence alignment and finally, we get the score the case of AL2CO what is input.

Student: Multiple sequence.

Multiple sequence alignment; so you have to work take your sequence and get the similar sequence homologous sequences and you have to do the alignment and you have to give the multiple sequence alignment as the input in this case if you give your PDB id it will automatically get the sequence right and find the homologous sequences right and then align the sequences get the multiple sequence alignment.

And finally, you will get the data. So, it is very simple just to give the id, but there are several issues; if your PDB id or sequence does not have sufficient number of homologous sequences you will you do not get any results it will tell you cannot calculate the conservation, but in the case of AL2CO you know that how many sequences you get depending upon the sequences if you give, you will get the results that is the difference between using different servers.

(Refer Slide Time: 18:01)



>2LZM:A|PDBID|CHAIN|SEQUENCE
MNIFEMLRIDEGLRLKIYKDTEGYYTIGIGHLLTKSPSLNAAKSELDKAIGRNCNGVITKDEAEKLFN
QDVDAAVRGILRNAKLKPVYDSLDAVRRCALINMVFQMGETGVAGFTNSLRMLQQKRWDEAAVN
LAKSRWYNQTPNRAKRVITTFRTGTWDA YKNL

ConSurf Color-Coded MSA

```
Input_pdb_SEQRES_A        M N I F E M L R I D E G L R L K I Y K D T E G Y Y T I G I G H L L T K S - P S L N A A K - - S E L D
UniRef90_Q7Y2B5_1_156     - - - - - M L R N D E G L R L T L Y K D T E G F W T I G I G H L V T K N - P S L A V A K - - A E L D
UniRef90_A8R9C2_1_162     M D I F D M L R Q D E G L D L N L Y K D T E G Y W T I G I G Q L V T K N - P S K D V A R - - A E L D
UniRef90_Q556F2_4_170     - S I E D M L K Y D E G E K L E M Y K D T E G Y Y T I G I G H L I T R I - K E R N A A I - - L S L E
UniRef90_Q86AA1_7_170     - - - - D M L K Y D E G E K L E M Y K D T E G N Y T I G I G H L I T K N - K D K N E A I - - K I L E
UniRef90_Q56EM5_184_346   V T I E D M L R Y D E G I R L V V Y W D S E G Y P T V G I G H L I T R E - K T K N M S R I N S L L S
UniRef90_Q19CF2_1_164     M T L E D M L I Y D E G R V L K V Y W D H L G Y P T I G I G H L I T P Q - K T T D M A L I N H T L S
UniRef90_Q8SDG3_181_342   - - I E K M L R G D E G Y R L K W Y L D S E G Y P T I G I G H L I T Y K - K T S D L G I I N N E L S
UniRef90_D5JFK5_180_341   - - I E K M L K Q D E G I R L R W Y T D S E G Y P T I G I G H L L T R E - K T R D T A K I N A A I S
UniRef90_Q56EE1_1_164     M T L E D M L V Y D E G R K L K V Y W D H L G Y P T V G I G H L I T L R - E T K D M G V I N H M L G
UniRef90_P16009_174_339   M S M A E M L R R D E G L R L K V Y W D T E G Y P T I G I G H L I N K Q - P V R D M A Q I N K V L S
UniRef90_Q76YA6_4_163     - - - - Q M L K Q D E G Y K L S V Y W D T E G Y P T I G I G H L I L R K - K T K D M G E I N R E L S
UniRef90_Q19CN7_158_320   V T I E D M L R Y D E G I R L S V Y W D S E G Y P T V G I G H L I T H E - K T R N M T R I N Q L L S
UniRef90_Q56BJ4_180_343   - - I E K M I R G D E G I R L T W Y Y D V E G Y - T I G I G H F F L T A P Q G T D P A V V N A A L S
UniRef90_C4MZP0_176_339   - - I E N M L H R D E G L R L K V Y W D T E G Y P T I G I G H L I T P Q - P I R D M N Q I N K I L S
UniRef90_Q7Y4Y4_176_339   - - I T E M L R R D E G L R L K V Y W D H L G Y P T V G I G H L I T M E - K T R D M S R I N K L L S
UniRef90_Q76YN5_188_349   - - I E K M L V Q D E G V R L K W Y L D S E G Y P T I G I G H L I T R E - R T S N L V T I N S I L S
UniRef90_A4SN28_23_148    - - - - - - - - F D N G M F L R - Y K D S L G Y W T I G Y G H L I T P N - E S Y - - - - - - - - - -
UniRef90_A9I970_8_140     - - - - - L L R G D E G E V L H A Y R D H L G Y L T I G V G R L I T K R - K G - - - - - - - - - - -
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

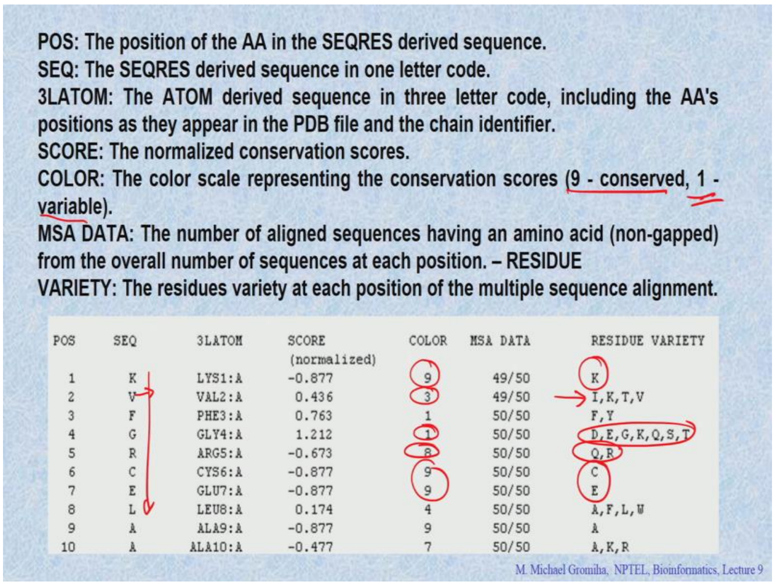
So, if we give this id. So, it gets the sequence from UniProt right from this database you get the sequences and then final it aligns from this one you can see different colors this is also easy to identify the residues which are highly conserved or which are highly variable right if you see these colors. So, can you see which residues are highly conserved which color.

Student: Magenta.

Magenta one right this is L right if you see this is T is a dark, right. So, dark color you can see this residues are highly conserved easily if you see this picture you can see that there is residues which are highly conserved and some cases it is highly variable. For example, if you see this one is occupied with the leucine, glutamic acid, threonine aspartic acid and valine right there are various residues occupies at this particular position. So, this will clearly tell you which residues are conserved which residues are variable.

Then we give the text file this will give complete details.

(Refer Slide Time: 19:04)



POS: The position of the AA in the SEQRES derived sequence.
SEQ: The SEQRES derived sequence in one letter code.
3LATOM: The ATOM derived sequence in three letter code, including the AA's positions as they appear in the PDB file and the chain identifier.
SCORE: The normalized conservation scores.
COLOR: The color scale representing the conservation scores (9 - conserved, 1 - variable).
MSA DATA: The number of aligned sequences having an amino acid (non-gapped) from the overall number of sequences at each position. – RESIDUE
VARIETY: The residues variety at each position of the multiple sequence alignment.

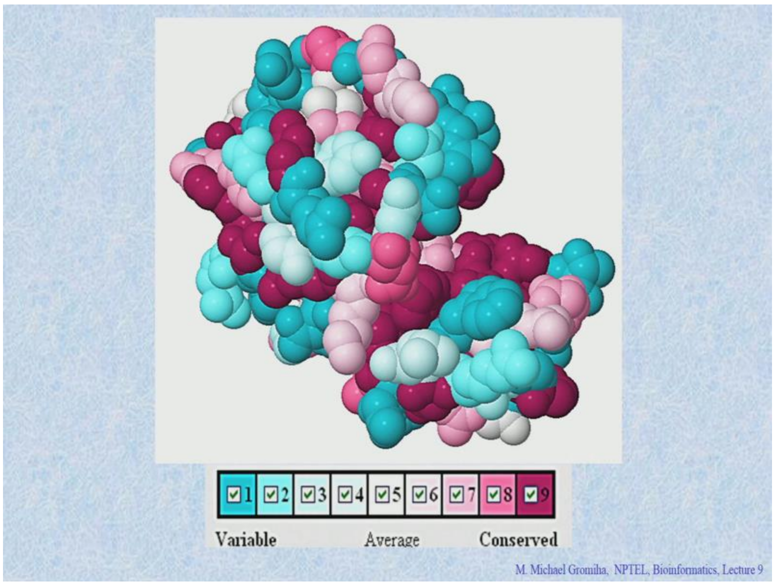
POS	SEQ	3LATOM	SCORE (normalized)	COLOR	MSA DATA	RESIDUE VARIETY
1	K	LYS1:A	-0.877	9	49/50	K
2	V	VAL2:A	0.436	3	49/50	I,K,T,V
3	F	PHE3:A	0.763	1	50/50	F,Y
4	G	GLY4:A	1.212	1	50/50	D,E,G,K,Q,S,T
5	R	ARG5:A	-0.673	8	50/50	Q,R
6	C	CYS6:A	-0.877	9	50/50	C
7	E	GLU7:A	-0.877	9	50/50	E
8	L	LEU8:A	0.174	4	50/50	A,F,L,W
9	A	ALA9:A	-0.877	9	50/50	A
10	A	ALA10:A	-0.477	7	50/50	A,K,R

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

Here this is your sequence. So, you get this is the atom in the PDB file and this is in normalize score and here they give the color code. So, color code they put the conservation is 9 and variable has 1. So, numbers varies where from 1 to 9; 9 is for conserved and 1 is for the variable this is a position in the derived sequence. So, you can see the data.

So, whether what is the variability first one this mainly by lysine and the second lysine by lysine. So, that means they are highly conserved and the second one valine is occupied by different residues isoleucine and lysine, threonine, valine. So, this way the color code is 3 that mean this is not conserved this is variable. So, here this is occupied to the lysine. So, it is the color code is nine. So, this highly conserved likewise if we see the same residues. So, this is highly conserved. So, there are only 2 residues. So, this conserve the color is 8 very highly variable right. So, you can see the color is one.

(Refer Slide Time: 20:19)



So, depending upon this variability they give color codes from 1 to 9, right and then you can see which one is conserved and which residues are variable then make this figure this is actual 3D structure the how the protein looks like right if we show different colors right can we see which positions or which location in the residues are conserved and which residues which positions they are highly variable.

Student: (Refer Time: 20:34).

If you look at this in these regions; so you can see some of them which are interior right they are highly conserved mainly because the residues are preferred to form the hydrophobic core. So, interior seeking residues are mainly hydrophobic residues. So, the variability is only among the hydrophobic residues. So, they prefer to be highly conserved right some cases mainly in the case of the surface.

So, you can see they are highly variable, they try to interact to other residues right may be different types of interactions in this case they have the variability to change the residues among the polar residues or charge residues and so on. So, when you have these 3D structures then also you can see which regions are highly conserved and where are the variable regions.

So, we get the conservation we will get a picture about this in the sequence level right which residues maintain to have the same position right in any homologous sequences. So, we can summarize again. So, what did we discuss today?

Student: Conservation score.

Conservation score, right. So, get the conservation score means you get the sequentially similar positions right among the homologous sequences the 2 steps the first is to get the frequency what are the difference ways to get the frequency.

Student: Unweighted.

Unweighted frequency.

Student: Weighted frequency.

Weighted frequency and independent counts to get the score.

Student: Entropy.

Entropy based method.

Student: Variance.

Variance based method and?

Student: Sum.

Sum of pairs method you can get different scores right you can use any of these methods to get this score. So, what are different algorithms we discuss online sources.

Student: AL2CO.

AL2CO; there is algorithm to conservation what is the input required for the AL2CO?

Student: (Refer Time: 22:16) multiple sequence alignment.

Multiple sequence alignment; so what is the another program we discussed.

Student: ConSurf.

Concept; what is the input for this ConSurf.

Student: pdbid.

pdbid right, it also it is also possible to give your sequences or the alignment right in the ConSurf also to get the conservation score. Finally, they give the numbers right from the variable to the conserve maximum 9, they give and the minimum they give 0 or 1. So, you can see in the PDB where you can see the conserved regions or where you have the flexible regions.

So, in the next class, then we will extend it right to see the how to construct phylogenetic trees right if you have the multiple sequence alignment how far the sequences are similar to each other whether 1 and 3 are similar or 1 and 5 are similar or 2 and 3 are similar. So, based on these variabilities we will try to construct trees from the tree construction. We will see what are the residues or which are the sequences right which are similar to each other which are closely related to each other and so on.

Thank you for your kind attention.