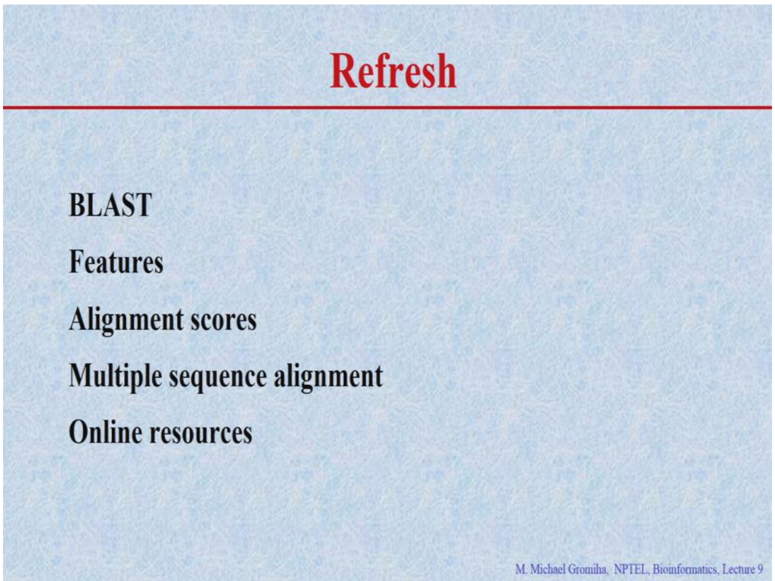
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 9a
Conservation score

In this lecture we will focus on Conservation Score. In the previous class we discussed about the multiple sequence alignment and if you align different sequences we will get an alignment and how to utilize this alignment to extract different features, how far we see that a particular residue highly conserved right in any positions.

So, what did we discuss in the last class.

(Refer Slide Time: 00:47)



Refresh

BLAST

Features

Alignment scores

Multiple sequence alignment

Online resources

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

We discussed BLAST, what is the BLAST?

Student: Basic (Refer Time: 00:51) local alignment.

Basic local alignment.

Student: Search tool.

Search tool right what is the software is called BLAST right, you can what are the various features of BLAST what applications of BLAST.

Student: Sequence to database we can align or two sequences we can align.

(Refer Time: 01:08) if you have a query sequence. You can identify the sequences which are similar to your query sequence. If you have two sequences you can see the identity or similarity between the sequences when you use BLAST identify any specific motifs or any specific patterns right inherently present in any sequence or any functionally important motifs.

So, they can use either DNA sequence or the protein sequence right for any sequence alignment. Then how to evaluate whether two sequences are similar or not, what are the various scores available to evaluate the alignment?

Student: Alignment score.

Alignment score depending upon the?

Student: Total (Refer Time: 01:53).

Total can do it the different aspects right one is you can with the matching pairs right. So, then second one you can do with the similarity, then we can do with the identity, then the query coverage, then p-value or the e-value right you can see probability of having that alignment with the specific threshold score right or what is the expected value.

Then we discussed about the multiple sequence alignment right how many sequences are required for multiple sequence alignment?

Student: More than 2.

More than 2 minimum we need 3 right 3 or more sequences if you have right what is the principle used in multiple sequence alignment, how they do the alignment?

Student: One by one.

Likewise they have the pairwise alignment, then see the similar sequences put together right and then they can align this distant related sequences. Finally, they make the complete the alignment right. Which is a program which can give the multiple sequence alignment.

Student: Clustal.

Clustal also we discussed few more software, what are the other online resources?

Student: MAFFT.

MAFFT.

Student: MUSCLE.

MUSCLE, PROMOL right this can you do that.

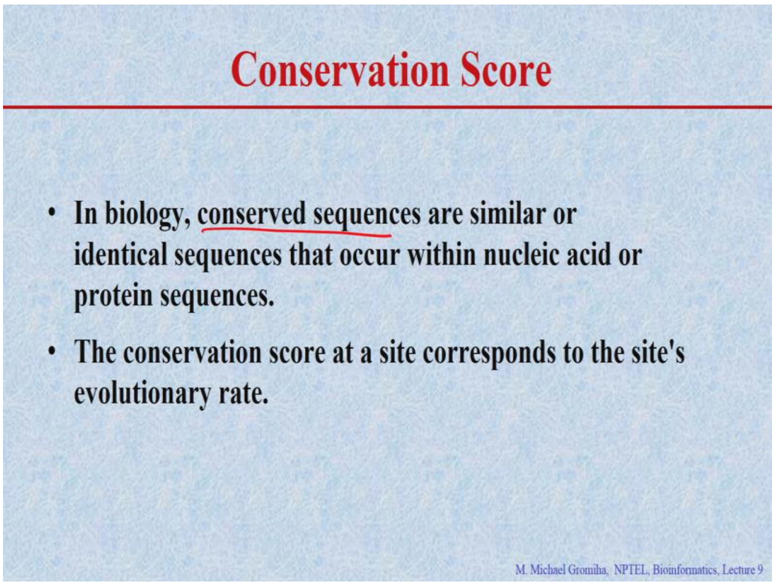
Then we discussed about psi-BLAST, what is psi-BLAST?

Student: Position specific.

Position specific right iterative BLAST so here you can make the position specific scoring matrices forward sequence right where running with the different sequences right. They can use different iterations with a threshold value. Finally, you can get the scoring matrix you can also call as profiles.

So, now if you have a multiple sequence alignment, then what are the applications what can we infer from multiple sequence alignment; in this several applications right. So, today we will discuss upon conservation. So, what do you think about conservation?

(Refer Slide Time: 03:31)



Conservation Score

- In biology, conserved sequences are similar or identical sequences that occur within nucleic acid or protein sequences.
- The conservation score at a site corresponds to the site's evolutionary rate.

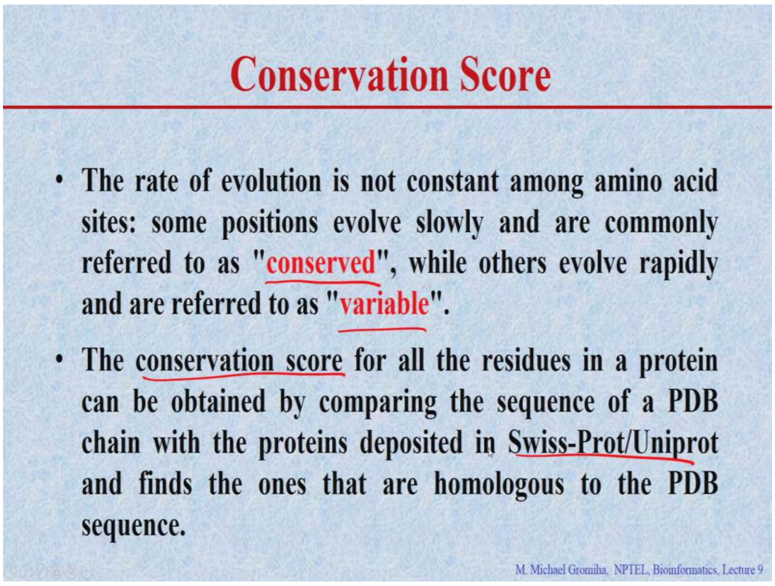
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

Right, if you think about the conservation if you have set of sequences some positions are occupied by same residue or similar residues right. In this case you can say that position that maintains a particular amino acid residue right either in protein sequences or nucleic acid sequences.

And these sequences also depends upon the evolutionary rate right how frequent your residue is mutated to different residues likewise when we discussed about the PAM matrix we have the alignment, several residues are the same and some residues change right during evolution.

So, likewise its core also depends on how far a specific residue is mutated right in the homologous sequences. So, whether this rate of evolution is same or different that depends upon the amino acid changes right. Some cases it remains the same, the PAM matrix also we discussed about the different amino acids right for example, cysteine or tryptophan, they get the high score and some positions which slowly to the different residues. Some cases they maintain to be constant and some positions they change from organism to organism.

(Refer Slide Time: 04:54)



Conservation Score

- The rate of evolution is not constant among amino acid sites: some positions evolve slowly and are commonly referred to as "conserved", while others evolve rapidly and are referred to as "variable".
- The conservation score for all the residues in a protein can be obtained by comparing the sequence of a PDB chain with the proteins deposited in Swiss-Prot/Uniprot and finds the ones that are homologous to the PDB sequence.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

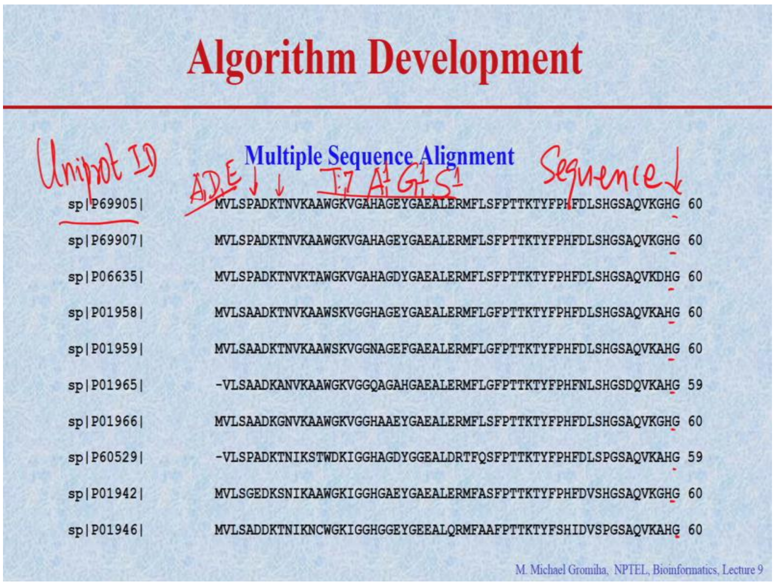
So, the positions which are the same right these are called the conserved. So, the ones which evolved rapidly that changes in different organisms they are called variable. So, if you have multiple sequence alignment, you can look at the three positions, in some positions you can see that is a residues they maintain at the same location they are called conserved, and some cases different organism they change the residues they are called variable.

Then how to get the conservation score right for example, if you have an amino acid sequence, we like to know which residues are conserve and which residues are variable how to do that? First we have to get the sequences right for the query sequence, you will get this similar homologous sequences where shall we get the homologous sequences? BLAST; where we can if you go to BLAST and then we get the homologous sequences and you get the sequences right and do the multiple sequence alignment and you get the score.

So, you have to compare the sequence of similar proteins right deposited in the Uniprot or Swiss-Prot database and which are homologous to this your own sequence right. Then you can compare the sequences and you will see where you have the variable residues and where you have the conserved residues.

So, now if you give multiple sequence alignment; for example, if you see position number 60.

(Refer Slide Time: 06:11)



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, if you look into the position number 60. So, this portion is conserved or variable?

Student: Conserved.

It is conserved because if you see the 10 different sequences, this is the Uniprot id ok. So, this is a sequence just for example, I kept from 1 to 60, we take the sixtieth position you can see all the residues right this positions right are accommodated with the residue glycine. If you look into this position right 1, 2, 3, 4, 5, 6, 7, 8, 9. So, this position is variable or conserved

Student: Variable.

Its variable and we can also say it is conserved if you see this position. So, what are the preference of amino acid residues at this position?

Student: T.

T.

Student: AGS.

AGS right TAGS, but we look into the all the 10 sequences, most of them have the residue threonine.

Now, if you have this alignment, how to give numbers how to get score. So, which one is highly conserved which one is highly variable right because some cases if you see, the position number 6 here also you see the position is accommodated with the alanine.

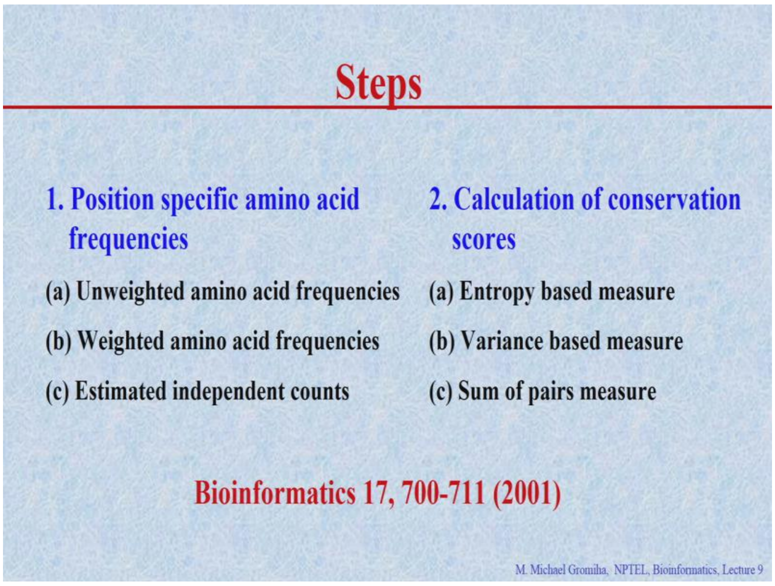
Student: D.

D.

Student: E.

E. So, here also three residues here also four residues right, which one is variable which one is conserved how far the difference between these two positions. So, we need to give a score. If you numerically get some numbers then you can compare this position a is more conserved than position number b. So, how to do that this involves two different steps.

(Refer Slide Time: 07:59)



Steps

1. **Position specific amino acid frequencies**
   - (a) Unweighted amino acid frequencies
   - (b) Weighted amino acid frequencies
   - (c) Estimated independent counts

2. **Calculation of conservation scores**
   - (a) Entropy based measure
   - (b) Variance based measure
   - (c) Sum of pairs measure

Bioinformatics 17, 700-711 (2001)

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

First step we need to get the amino acid frequencies; each position how far each amino acid residue prefers at a particular position, and then the second parts we convert this frequencies into score.

So, there are various ways to get the frequencies as well as to get the score. If you see the frequencies, I have listed the three different ways to get the frequencies, one is unweighted amino acid frequency. Here we do not give any weightage, all amino acids are treated same, all sequence are treated same. So, we do not give any weightage to any sequence or any positions or any amino acids and the second one we give weightage to some amino acids or to some sequences. For example, if you have particular sequence right we know that we need to maintain the residue positions, then we can give some preferences right we have to maintain that particular sequence.

Likewise say some positions for example, some cysteine this very highly conserved, we know that forming disulfide bridge. So, if you want to keep that we will give weightage to that particular residue right. See if it is cysteine then we give more weightage otherwise you will give less weightage you can do that.

Then third one you can compare with the random choice; you have your aligned sequences then if you shuffle randomly you will get the sequences and how far it will match, your aligned sequences as well as the random sequences right. Then this case you can say. So, this is your aligned sequence, which are really significant and you can calculate the frequency of that particular positions.

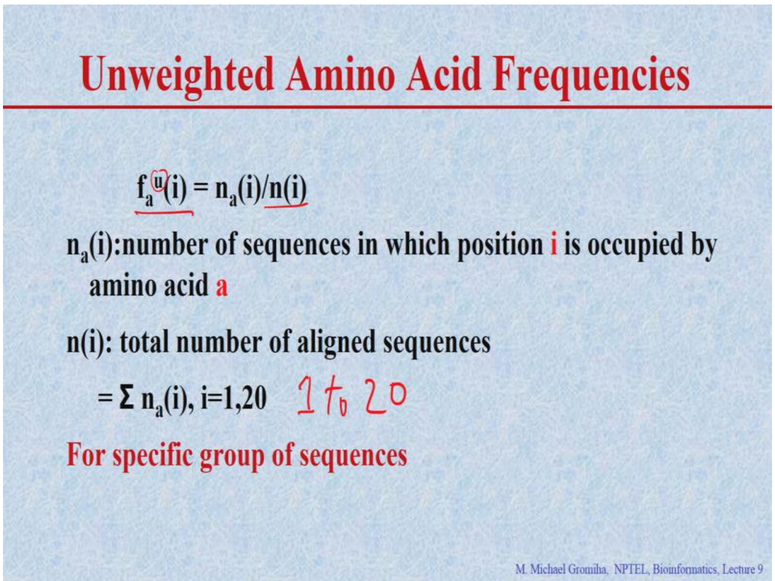
So, if we get this preference, then we can convert this in the score because we need a numerical values right. So, either we can use entropy based method considering the probability of a residue at the particular position, take the probability multiplied with logarithmic of probability you can get score. And the second one you can do the variance based measure right how far this positions vary with different positions for example, alanine in position number 3, how this differs from alanine in position number 10; how many alanines in position number 3 how many alanines in position number 10 and totally how many alanines in sequence compare with all these with positions and the residues you can measure a score.

Then you can also use sum of pair score for example, if a sequence a and you can the second sequence b. So, you have two different amino acids. So, you can use c matrix to see how far they are consistent, how far they are aligned right which matrix usually we use? PAM matrix.

Student: PAM.

or BLOSUM matrix, you can see and you put proper weightage and see whether which one is highly conserved which position is highly conserved.

(Refer Slide Time: 10:41)



Unweighted Amino Acid Frequencies

$\underline{f_a^{\boxed{u}}}(i) = n_a(i)/\underline{n(i)}$

$n_a(i)$:number of sequences in which position i is occupied by amino acid a

n(i): total number of aligned sequences

$= \Sigma\, n_a(i)$, i=1,20 1 to 20

For specific group of sequences

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

So, I will explain the steps one by one first we go with the unweighted amino acid frequencies here we do not give any weightage in this case u means unweighted. So, frequency of this a at the position i this is given us number of sequences having the residue alanine at position number i, normalized with total number of aligned sequences.

So, this i varies from 1 to 20 this is i equal to 1 to 20 for the 20 different amino acid residues. So, if you see this alignment what is the frequency of occurrence of these amino acid residues it position number 9, how many times T?

Student: 5 6 times.

6 times T 6 Ts.

Student: 7 7 times.

7 times T 7 times A.

Student: 1 time.

1 time G.

Student: One time.

One time S.

Student: One time.

One time right if you have another position for example, alanine 7 times, and any of the other residues 1 1 1. So, you will get this similar score or a different score.

Student: (Refer Time: 11:56).

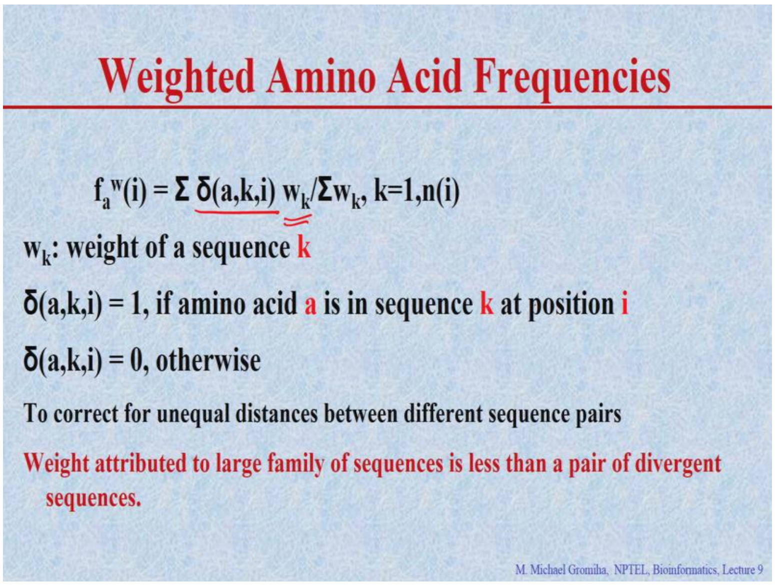
Similar score because here we do not use any way to any different amino acids, we give equal weightage to all amino acids. Even in this case, if it is instead of threonine 7 if it is valine 7 times and also alanine one tryptophan one.

So, then we will get the same score, but if you look into this alignment sometimes if we have the threonine it won’t match with the hydrophobic residues. But even in this case you will get a same score, because we do not give weightage to any amino acid residues.

So, here fa of i is the each residue i right any residue for a in the any position i, normalized to with the total number of sequence at a particular position fine.

So, normally you use this for a specific group of sequences with highly homologous any specific group, then we expect similar residues at similar positions. So, in this case we use unweighted amino acid frequencies for any specific group of sequences.

(Refer Slide Time: 12:51)



Weighted Amino Acid Frequencies

$f_a^w(i) = \Sigma\ \underline{\delta(a,k,i)}\ \underline{w_k}/\Sigma w_k$, k=1,n(i)

w_k: weight of a sequence k

$\delta(a,k,i)$ = 1, if amino acid a is in sequence k at position i

$\delta(a,k,i)$ = 0, otherwise

To correct for unequal distances between different sequence pairs

Weight attributed to large family of sequences is less than a pair of divergent sequences.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

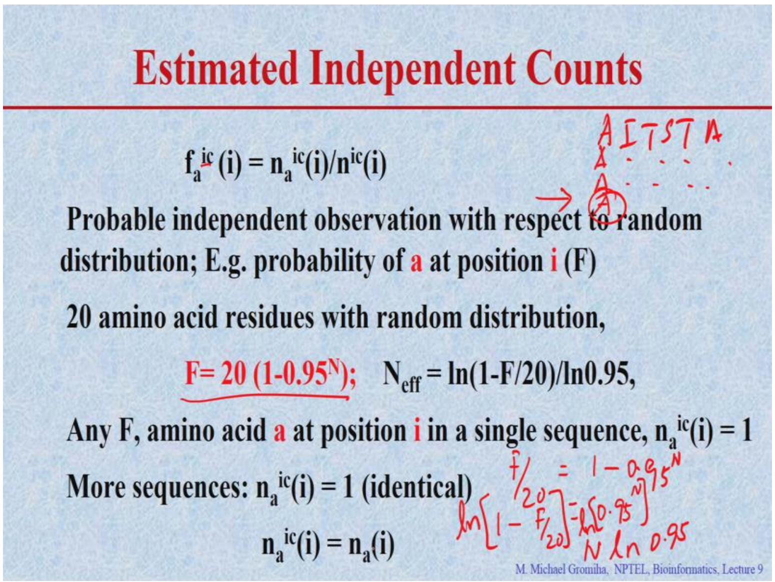
So, the second one this is a weighted amino acid frequencies, here we give weightage. If we give weightage to any specific sequence or we weightage to any given residues right to give preference. Here we introduce this term right delta aki . So, this equal to 1 if a in the sequence k at position i, if you decide position right any position i in the sequence k. So, amino acid a in the sequence right and then you give the weightage for that particular sequence, otherwise this equal to 0 is wk is the weightage of a sequence k right and delta aki equal to 1, if the amino acid a is in the sequence k at position i.

If it is not then you can put 0 otherwise. So, the k equal to one to number of sequences. So, here why we need to do this weightage? Position conserved, also we give some preference to some specific amino acid residues, if threonine 6 times and 4 hydrophobic residues, this different from threonine 6 times and serine 4 times your similar type of residues you can have more score see this case we gave weightage.

Then also you can give some sequences more weightage; For example, if you have several sequences right and some of them are highly related and some of them are distant related. If you want to include the distant related sequences for example, divergent sequences, if you want to include in this score then you have to give weightage to the sequence, otherwise if nine sequence are homologous and one is a distant related one.

So, the score will be always high because based on the 9. So, if you want to include the preference of these divergent sequences, then we need to give a weight. So, in that case you can include the information from these divergent sequences. Then the third one, we can also use this frequency based on independent counts how far you can get these counts in random distribution.

(Refer Slide Time: 14:48)



Estimated Independent Counts

$f_a^{ic}(i) = n_a^{ic}(i)/n^{ic}(i)$

Probable independent observation with respect to random distribution; E.g. probability of a at position i (F)

20 amino acid residues with random distribution,

$F = 20\,(1-0.95^N)$; $N_{eff} = \ln(1-F/20)/\ln 0.95$,

Any F, amino acid a at position i in a single sequence, $n_a^{ic}(i) = 1$

More sequences: $n_a^{ic}(i) = 1$ (identical)

$n_a^{ic}(i) = n_a(i)$

$F/20 = 1 - 0.95^N$

$\ln\left[1 - F/20\right] = \ln\left[0.95^N\right]$

$N \ln 0.95$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

So, independent count they put ic, the same the frequency of occurrence of any residue a at position number i, which is given as number of residues of a at position number i normalized with the total number of residues at any position.

So, if you all the 20 residues or at the random distribution, then you can see assume that F equal to 20 into 1 minus 0.95 into N, if N equal to 1 what is the value?

Student: (Refer Time: 15:19).

1 minus 0.95.

Student: 9 (Refer Time: 15:22).

Is equal to 0.05.

Student: (Refer Time: 15:24).

0.05 into 20 that is equal to.

Student: 1.

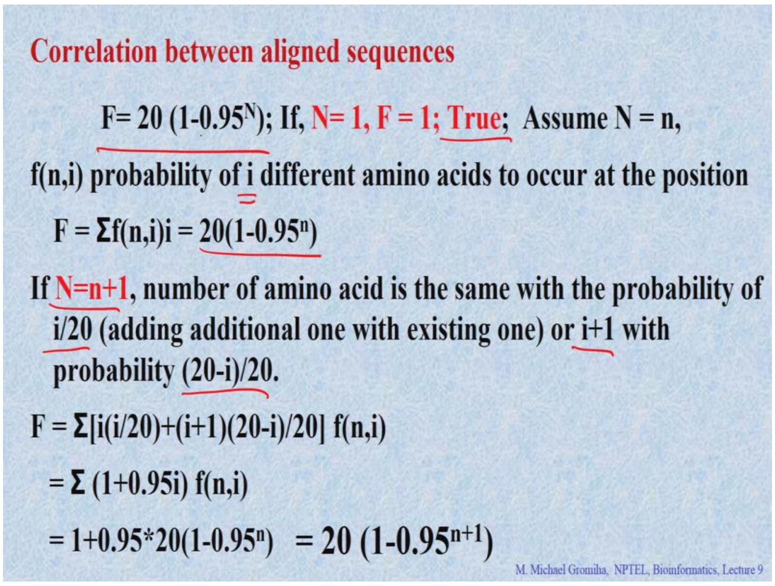
One right; so now, you can see for us you can assume that, we can fit this function F in this equation such a way that if it is totally random completely 5 percent. So, again if you do this a equation, you can get the equal to one F equal to 1 this is expected that is fine.

So, from this one you can calculate the effective of this n you have the actual values and you have the random values and then you can fit with the some function right in this case you can fit with this N as N effective. So, use this equation then if you change this equation to get the N effective right. So, you move this 20 here; that means, F by 20 equal to 1 minus 0.95 into N power N, in this case 1 minus F by 20 this equal to 0.95 power N, then take the logarithm ln right ln then you can move this N here that is N into ln of 0.95 from this you can get this N effective right this equal to ln (1 – F) / 20 divided by ln (0.95). So, you can get this number.

So, for any F take amino acid a in position I in a single sequence this is equal to 1 if you take that then if you have more sequences for example 20 sequences. If you have one more sequence there are two possibilities; one possibility is that could be the same amino acid right for example, if you have a sequence right A I T S T A right this is a conserved here right then we add new sequence, there are two possibilities one possibility is equal to be A, then this case $N_a^{ic}(i) = 1$ because this is identical, even if add more, there is no difference because average is the same.

And second possibility is other than a whatever it is, because if it is other than a it is not identical if it is not a then $N^{ic}(i) = N_a(i)$ because depends upon the amino acid.

(Refer Slide Time: 17:45)



Correlation between aligned sequences

$F= 20\ (1-0.95^N)$; If, $N= 1, F = 1$; True; Assume $N = n$,

$f(n,i)$ probability of i different amino acids to occur at the position

$F = \Sigma f(n,i)i = 20(1-0.95^n)$

If $N=n+1$, number of amino acid is the same with the probability of $i/20$ (adding additional one with existing one) or $i+1$ with probability $(20-i)/20$.

$F = \Sigma[i(i/20)+(i+1)(20-i)/20]\ f(n,i)$

$= \Sigma\ (1+0.95i)\ f(n,i)$

$= 1+0.95*20(1-0.95^n) \quad = 20\ (1-0.95^{n+1})$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 9

So, now, the question is if you add one sequence what is the probability of this function F right whether this will fit with this equation or not.

So, in this case imagine the equation with two conditions, one if N = 1 or N = n + 1. In both the cases if you are able to fit in this equation, then you can say that you can use this to get the independent counts. If N = 1 then it will fit to this equation N = 1, F = 1 that we discussed earlier. So, that is true then assume N = n where, f of N_i is a probability of i right different amino acids to occur at the position.

So, now you can see this for if N = n you will get this equation f (n,i) that is equal to 20 * 1 - 0.95 * n right that is fine because we assumed it that is that is fine.

Now, if N = n + 1 they add one more sequence, then there are two possibilities one is its i / 20 adding the same one or you can take the i + 1 with the probability of (20 – i) / 20. We have 20 possibilities not the same one. So, I have to subtract. So, (20 – i) / 20, we have 20 different possibilities.

Then if you add this in this equation F = i * (i / 20) right and (i + 1) * (20 – i) / 20 in the function of n,i. So, if you simplify this equation, you will get this 1 + 0.95 * i with the function of n, i and this equal to 20 (1 - $0.95^{n + 1}$) because we get derive these numbers, and if this is equal to zero point this one if you divided by 20 right 0.05. Then finally, if you take this out 20 out then we will get 20 (1 - $0.95^{n + 1}$). So, we look into these two equations, if N = n or N = 1 it read obeys this equation. If it is n + 1 the same instead of n we use n + 1.

So, in this if you prove this one then we can use this particular equation to see because the actual and the random distribution and you get the independent counts.