

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

**Lecture - 11a**  
**Protein sequence analysis**

In this lecture, we will discuss about the features or the parameters or the properties which can be derived from protein sequences.

(Refer Slide Time: 00:25)

**Refresh**

---

Phylogenetic tree  
Methods  
Tree construction: UPGMA method  
Phylip

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 11

In the last class; what did we discuss?

Student: Phylogenetic.

Phylogenetic tree construction right, what is a tree?

Student: It's a relationship.

It will give a representation of relationships, right. So, we can give the relationship among different sequences right, for example, we use the protein sequence or the DNA sequence and also we try to see the time right, to evolve from one sequence to another sequence.

So, the different ways to construct phylogenetic trees right; what are the different; what is the very commonly used method to construct phylogenetic trees?

Student: UPGMA.

UPGMA method right, we also discussed how to construct a tree for a sequence set of sequences, we used 5 different sequences right and we constructed a tree right. So, in this case, A and C are common to each other and D and E are common to each other right and then B and A C are common to each other. We construct a tree and accordingly we assign weights, it may be also possible to add another one internal node at (AC, B) at this point and you can recalculate the weights, then what are other methods we discussed for the tree construction?

Student: neighbor

Neighbor joining method and maximal likelihood methods and so on, then we discussed about a program which can be used to construct trees. So, what is the name of the program?

Student: Phylip.

Phylip; what is the input for the Phylip?

Student: Multiple sequence alignment

There is multiple sequence alignment, right you can use MAFFT to get the multiple sequence alignment; you can directly give this input in Phylip and you can get the trees.

Then we discussed about the bootstrapping. That is one of the methods to assess the performance of this method as well as significance, we will be making this as several sets of random sampling.

(Refer Slide Time: 02:09)

**Protein sequence analysis**

```
>sp|P01966|HBA_BOVIN  
MVLSAADKGNVKAAWGKVGGHAAEYGAELERMFLSFPT  
TYFPHFDLSHGSAQVKGHGAKVAAALTKAVEHLDDLPGA  
LSELSDLHAHKLRVDPVNFKLLSHSLLVTLASHLPSDFT  
PAVHASLDKFLANVSTVLTSKYR
```

**What can we do with sequence?**

M. Michael Gromiha, NPIEL, Bioinformatics, Lecture 11

So, in this lecture we mainly focus on this protein sequence analysis, right where shall we get the sequence? UniProt database, right. So, I show a sequence, this is for the hemoglobin a chain from bovine. So, and this is amino acid sequence right, I give the amino acid sequence in single letter code right. You are all familiar with the single letter code and 3 letter codes. So, here I use mainly single letter codes for each amino acid residue.

So, we have the sequence. So, what can you do with the sequence, what are the information you can derive from the sequence? So, look at the sequence; it is the combination of different amino acid residues in a specific combination right, sequence.

So, whether any differences or any similarities among different sequences in UniProt database? If there are similarities or if there are differences, whether we are able to use these differences to infer the function, or to infer the 3D structures, or to infer any binding sites and so on, right. Now, we have the sequence right, there are various parameters you can calculate. What are the easiest parameters you can calculate from an amino acid sequence.

Student: Frequency of occurrence.

The frequency of occurrence right; so easily you can count, how many times A is present, how many Ds how many Cs and so on.

(Refer Slide Time: 03:31)

**Amino acid occurrence**

It is the number of amino acids of each type present in a protein.

E.g. THISISAPEPTIDE

A: 1; C: 0; D: 1; E: 2 etc.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 11

For example if I show a peptide, this is not a real sequence, this is peptide. So, if I give this one, you can easily count the number of times each amino acid residue occur in this particular peptide.

So, How many times A present in this peptide?

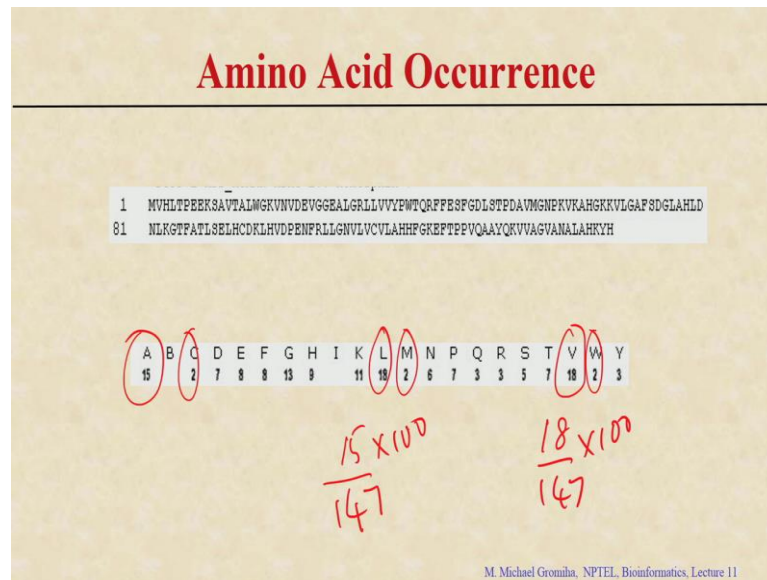
Student: one. once.

There is only once; A is a present once; how many times E?

Student: Twice.

Twice: so one here and one here; 2 Es. So, we give any sequence, you can get this number.

(Refer Slide Time: 04:10)



And then see whether these numbers resembles anything right, we will discuss about that.

So, now this is a real amino acid sequence right, this is a protein; so here if you see this amino acid sequence. So, how many times A occurs in this sequence?

Student: 15.

15 times A occurs, right. If you look into this occurrence of different amino acid residues, some amino acids occur very frequently or many more times. For example, what other amino acids, which have high preference, high occurrence?

Student: Valine, leucine

Leucine 18 times.

Student: Valine

Valine 18 times. So, these are the amino acids which occur very frequently in protein sequences right. So, there are some amino acids right, which are rarely occurring. For example if you see cysteine, only 2 times, tryptophan 2 times, for this case, methionine also 2 times.

So, if you consider several sequences you can see the variation in the patterns. For this particular sequence, for mainly the hydrophobic residues, alanine, lysine, valine right, they occur predominantly, that's 15 times, 18 times in this sequence, right you can use this information; for example, if you have 2 sequences, one is the short sequence, one with the long length; for example, one with the 100 residues one with 1,000 residues.

So, how far this occurrence vary? For the small sequence, we get the less numbers right because only 100 residues. If it is 1,000 residues in a sequence then the numbers will be high, at least 10 times, because the length is more. So, directly if you compare these 2 residues right what will happen?

Student: (Refer Time: 05:51).

Naturally the longer ones have more number of amino acid residues. To normalize this, we have to directly compare, then we have to normalize with something right. What is the factor we need to normalize, in this case?

Student: Length.

Length, because first case 100 residues and the second one 1,000 residues. So, if you normalized with the length then you can see whether any preference of amino acid residues in any particular protein of different lengths.

(Refer Slide Time: 06:21)

**Amino acid composition**

Occurrence

It is the number of amino acids of each type normalized with the total number of residues.

It is defined as:

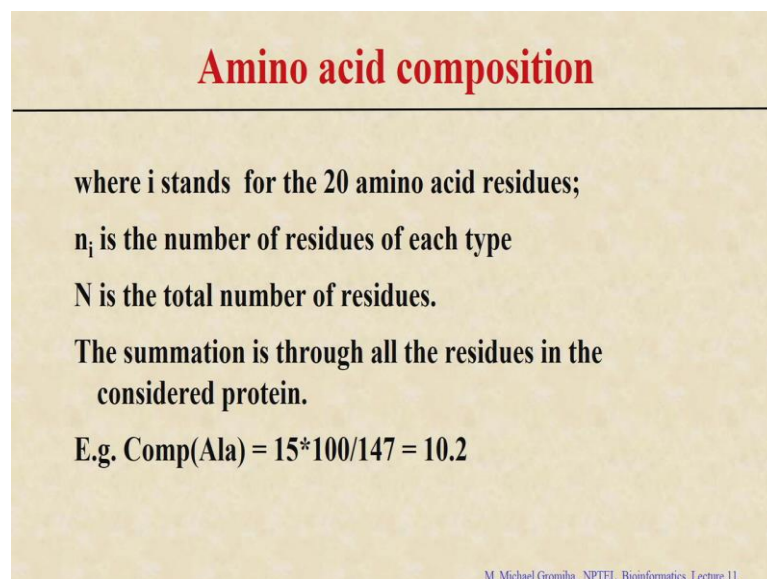
$$\text{Comp}(i) = \sum n_i * 100.0 / N$$
$$i) = \frac{n(i) * 100}{N} \quad \text{Comp(Ala)} = \frac{n(\text{Ala}) * 100}{N}$$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 11

So, in this case, we use a term, composition, right. So, composition gives you the normalized value of amino acid occurrence. So, essentially this is the number of amino acids of each type, this is called occurrence right, this we call as occurrence, and divided by total number of residues N. So, we can define amino acid composition as composition of I, this is equal to  $\sum n_i$  into 100 divided by N right, this is equal to  $i$ ;  $i$  means for each type of  $i$  you have to count how many times  $i$ . You can for example, take single residue, you can also see  $n_i$  multiplied by 100 to get the percentage, and divided by the N for any specific residue  $i$ .

For example, alanine composition of alanine equal to number of alanine, in the whole sequence divided by n, if you want to percentage, you can multiplied by 100. So, here  $i$  stands for the 20 different residues right. For each residue alanine, aspartic acid and valine. So, we will get for the 20 different residues.

(Refer Slide Time: 07:36)



**Amino acid composition**

---

where  $i$  stands for the 20 amino acid residues;  
 $n_i$  is the number of residues of each type  
 $N$  is the total number of residues.  
The summation is through all the residues in the considered protein.  
E.g.  $\text{Comp}(\text{Ala}) = 15 \cdot 100 / 147 = 10.2$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 11

So,  $N$  is a total number of residues which we use for normalizing the occurrence right. Summation you can take do it for all the residues, but if you take the total that will be 100 percentage right, if you take the total we will get the 100 percentage.

So, you can take the composition of alanine which can calculate as 15 residues for example, into 100 divided by 147, that is equal to 10.2 right; this is if I take this example; here how many alanines in this protein?

Student: 15.

15. So, 15 divided by;

Student: (Refer Time: 08:08).

147 right if you consider 147 residues, multiplied by 100; what is the composition of valine?

Student: 18 (Refer Time: 08:16).

18.

Student: Divided by (Refer Time: 08:18).

Divided by 147 multiplied by 100; right you will get the composition of any specific residue in this protein. If this is the case, if there are 100 residues in a protein and 1,000 residues in your protein, right you will get the composition and total will be?

Student: 100.

100, right. So, even the 100 residues present or 1,000 residues present, when we normalize with chain length we will get total value 100. In this case, you can directly compare the composition of each amino acid residues right. Even if one is 50 or another one is 150. So, we can compare when we normalize with the chain length. So, this will give you some information other than just compare in the occurrence. Then how to write a program, how to calculate the composition?



(Refer Slide Time: 09:06)

**Algorithm**

1. Read 20 amino acid residues: E.g. aa(i), i=1,20
2. Read the sequence: seq(i), i=1,n
3. Normalize number of residues: no(i)=0
4. Compare each residue in the sequence with standard 20 residues

```
do i=1,n
do j=1,20
if (seq(i).eq.aa(j)) no(j) = no(j)+1
```

Handwritten notes:  $n(A)=0+1$ ,  $n(D)=0$ , AIKTLTVART, Input: Sequence List of amino acids (20) ACDEF...

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 11

If you give any sequence right, for example, I have a sequence like this. So, I want to calculate the composition. What are the input we require?

Student: Sequence.

We need a sequence; this is a sequence we require then what else we need?

Student: List of amino acid.

List of amino acid we require right, how many amino acids?

Student: 20.

20. So, if you represent the sequence in single letter code right. So, we have to list the amino acid residues in single letter code, right this is because then only, we can easily match ok.

Now, we can have the 20 residues. So, we have a sequence, then we need to get the occurrence first. For getting the occurrence before you count, first we need to normalize right for 20 residues, we need to normalize all the numbers into 0. So, initially we can now see that number of alanine equal to 0, number of aspartic acid is equal to 0, for all the cases we put 0.

Now, what we have to do? We have to compare. Take the sequence first for this A; we have to compare this A with the list of amino acids, okay here you can see the list of amino acids; if you compare this and this if it matches, then we add in this array number of A equal to one. Second one, if you take I, it matches A? It will not match, it goes on to the 20 residues, when you have I, it will match. So, when you pass all the amino acid residues in the sequence, so where you have match then you can count.

So, now we have the numbers for all the 20 residues, this will give you the amino acid occurrence.

(Refer Slide Time: 10:55)

The slide is titled "Algorithm" in red text at the top. Below the title, there are three numbered steps:

5. Count the number with respective residues for each match
6. Result will give the occurrence of each residue
7. Normalize with total number of residues

Below the steps, the formula  $comp(i) = no(i)/n$  is written in blue. A red checkmark is next to the denominator 'n', and a red " $\times 100$ " is written below the formula. The word "occurrence" is written in red cursive next to step 6. At the bottom right of the slide, there is a small text: "M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 11".

So, now we have 20 numbers; for the 20 amino acids right as an occurrence. So, now, we have the occurrence; now what we have to do?

Student: Divide by length.

So, we have the occurrence for the 20 residues, then normalize with number of residues. If you take each residues right 1 to 20, normalize with the n, for each case then we will get the composition.

If you want to get the composition in percentage, then you have to multiply this with 100 right, very simple. So, we take the sequence and get the 20 different amino acids and do a matching. If it matches then increase the number as +1 right, because already we initialized, then complete it till the end of the sequence right, then we have 20 numbers

for the 20 amino acids, then normalize with n. So, then, we will get the composition right, this easily you can calculate the composition.

(Refer Slide Time: 11:49)

<u>ADCEFGHIKLMNPQRSTVWY</u>		
seq.dat		
79		
MALLPAAPGAPARATPTRWP	A	17 21.52
VGCFNRPWTKWSYDEALDGI	D	2 2.53
KAAGYAWTGLLTASKPSLHH	C	1 1.27
ATATPEYLAALKQKSRHAA	E	2 2.53
	F	1 1.27
	G	5 6.33
	H	3 3.80
	I	1 1.27
	K	5 6.33
	L	8 10.13
	M	1 1.27
	N	1 1.27
	P	8 10.13
	Q	1 1.27
	R	4 5.06
	S	4 5.06
	T	7 8.86
	V	1 1.27
	W	4 5.06
	Y	3 3.80

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 11

So, now again example I have a sequence. This is amino acid sequence, how to get the composition? So, here you have 20 residues, you compare with these residues and then you can find the number of residues in each 20 residues.

For example, if you see, how many times G occurs?

Student: 5.

5; 1, 2, 3, 4, 5, right; G occurs 5 times, D?

Student: 2; 2.

D 2 times right for the 20 residues right, we have the values for the 20 residues, if you add up everything you will get 79, this is occurrence right, to get the composition what we have to do?

Student: Divided by N.

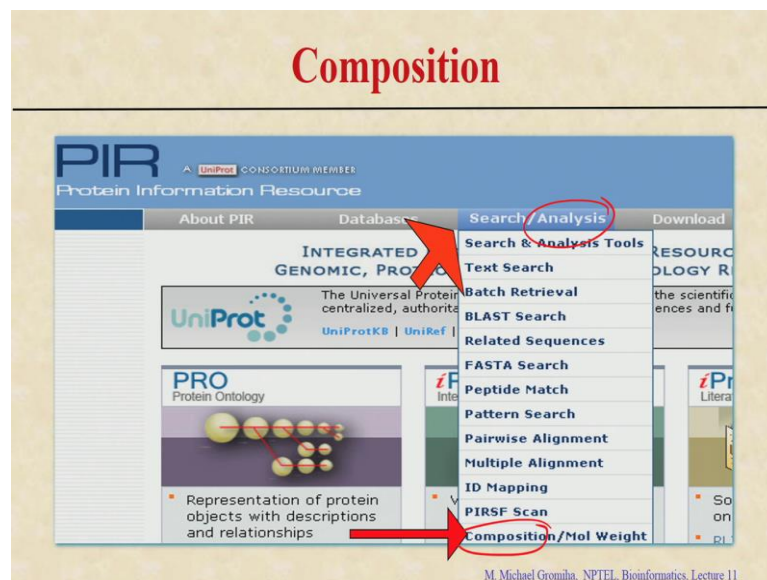
You have to divided by N. So, in each case, you will normalize with the N. What is N here?

Student: 79 (Refer Time: 12:31).

N equal to 79, we divide with the 79, then we will get this number and multiplied by 100 you will get in percentage. So, now, if you have set of sequence, if I give you can calculate right, you can calculate the occurrence, as well as you can calculate the composition.

So, you can also write the program and this same information you can also obtain from different resources right. There are several software available in the literature, just you give the amino acid sequence right then you can get the composition occurrence.

(Refer Slide Time: 13:02)



So, one of the software you can see is PIR and we discussed earlier; what is PIR?

Student: Protein information resource.

Protein information resource right; so what is it mainly develop for what?

Student: Sequence database.

Sequence database right, they collected the or amino acid sequences and they developed database for sequences. And when they make the database, it is very important to do some analysis and to get some information, otherwise adding up the data one by one to many data, it will not make any sense. So, you need to analyze the data and we have to extract some information which are hidden in the sequences.

So, they developed several tools: you can search with a text, you can search with BLAST, you can search with the composition, or many things you can calculate. So, if you want to get it to calculate, go with this analysis, if you see a list of tools available, and here you can see the tool which can calculate the composition.

(Refer Slide Time: 13:58)

**Composition**

HOME / Search / Composition/Molecular Weight Calculation

Composition/Molecular Weight Calculation Form

Enter any UniProtKB identifiers:  
(separated by a space)

and/or  
Insert your sequences below using the single letter amino acid code:  
(separate sequences by an empty line)

sp|P68871|HBB\_HUMAN Hemoglobin subunit beta OS=Homo sapiens GN=HBB  
PE=1 SV=2  
MVHLTPEEKSAVTALWGKLVNDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK  
VKAHGKKVLGAFSDGLAHLNLRGTFATLSELHCDKLVDPENFRLLGNVLVCLAHHFG  
KEFTPPVQAAYQKVVAGVANALAHKYH

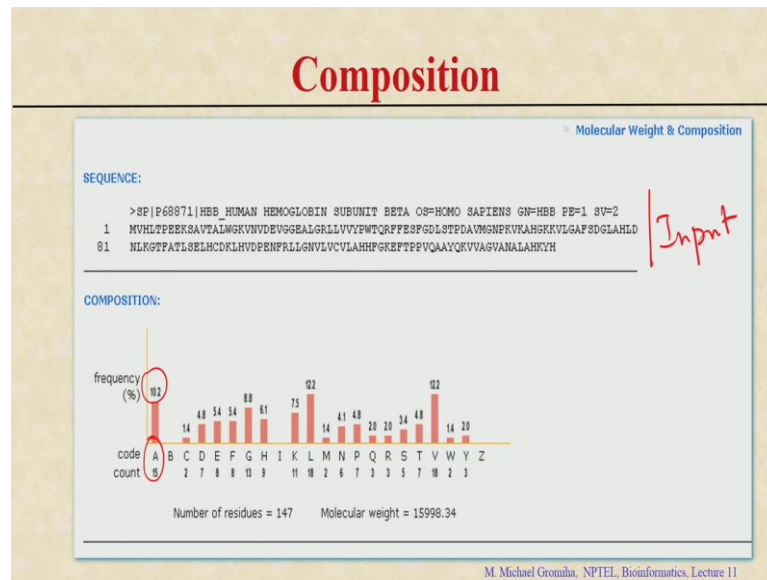
Submit Reset

Example: P53039 (sample output/annotated output)

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 11

So, go ahead, click on this composition and we get this window. Either you give the UniProt identifier or you can give your sequence. So, if you give this, your amino acid sequence right; so, the insert your amino acid in a the single letter code, right you can give single letter code right. So, then if you click on submit.

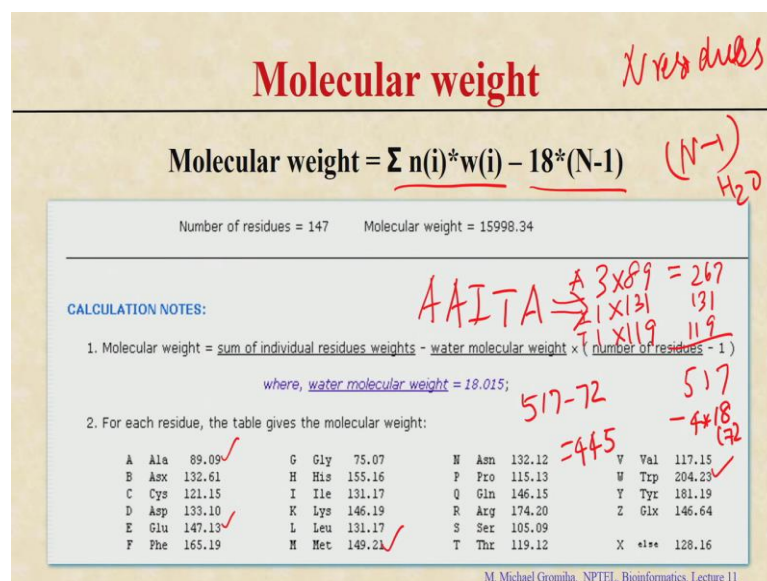
(Refer Slide Time: 14:21)



You will get the sequence, this is your input, okay here this is your sequence occurrence as well as the composition.

Run the amino acid residues here. So, as say I showed earlier, alanine occurs 15 times and the composition is 10.2%. So, you get the sequence in the X axis and Y axis, you can get the composition right. For any sequence, just you paste the sequence and you get the data.

(Refer Slide Time: 14:49)



Then the next one you can calculate the molecular weight, the simplest one. So, how to get the molecular weight?

Student: Multiply.

So, because if you have the type of data, for example, if you have one amino acid and 2 amino acids, 3 amino acids right, if you join together, by means of?

Student: Peptide.

Peptide bond. When you make a peptide bond, so, we will by the elimination of?

Student: Water.

Water molecule, right; so in this case, when you combine amino acids, 2 residues we will eliminate one water molecule right, if there are N residues, how many water molecule we will eliminate?

Student: N-1 (Refer Time: 15:24)

N-1 water molecules, we will eliminate. So, in this case, if you want to calculate the molecular weight, right in a sequence, for each residue, we know their molecular weight. For example, alanine is 89, and glutamic acid 147 or methionine is 149, tryptophan is 204 and so on; we know the values. So, we have a sequence; for example, this is a sequence. So, to calculate the molecular weight right, What we have to do? First we have to assign values for each of the residues. So, how many alanines here?

Student: 3.

3. So, 3 multiplied by;

Student: 89

89, then what I is 1 time, 1 multiplied by.

Student: 131.

131 and 1 T, multiplied by.

Student: 119.

119. So, this equal to?

Student: 267.

267.

Student: 1 (Refer Time: 16:25).

131, 119; so total will be?

Student: 1..

One.

Student: 1.. 7, 517.

517.

Student: Yes sir.

Right. So, this is the molecular weight?

Student: No.

No, right because this is the molecular weight of this 5 amino acid residues, but when we combine 5 amino acid residues we will eliminate?

Student: 4

4 water molecules right. So, you have to subtract.

Student: 4.

4 multiplied by.

Student: 18.

18; what is 4 multiply by 18? 72. So, now, the answer will be 517-72, this equal to?

Student: 445



445. So, if we have any sequence, we can calculate the composition, occurrence and you get the molecular weight. If we have the occurrence you can use the occurrence to get  $n(i)$ . So, multiplied with the weight for each residue  $w(i)$ . So, this will give you the full sequence, when forming the sequence, you will eliminate  $N-1$  water molecules. So, you subtract with the 18 in the  $N-1$  because 18 is a molecular weight for water and then we get the molecular weight of that particular protein, fine. I will give you a sequence ok.

(Refer Slide Time: 17:38)

### Molecular weight

**seq. dat**

79

**MALLPAAPGAPARATPTRWPVGCFNRPWTKWSYDEALDGI**

**KAAGYAWTGLLTASKPSLHHATATPEYLAALKQKSRHAA**

*Occurrence*

*A D C E F*

*w(i)*

*$\sum_{i=1}^{20} n(i) \times w(i)$*

*- 18(n-1)*

A	Ala	89.09	G	Gly	75.07	M	Asn	132.12	V	Val	117.15
B	Asx	132.61	H	His	155.16	P	Pro	115.13	W	Trp	204.23
C	Cys	121.15	I	Ile	131.17	Q	Gln	146.15	Y	Tyr	181.19
D	Asp	133.10	K	Lys	146.19	R	Arg	174.20	Z	Glx	146.64
E	Glu	147.13	L	Leu	131.17	S	Ser	105.09	X	Other	128.16
F	Phe	165.19	M	Met	149.21	T	Thr	119.12			

79

→

8129

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 11

Now, I give a sequence; the same sequence I gave right. What all information do you need to calculate the molecular weight?

Student: Sequence.

Sequence; we need and then we need the 20 residues, but we got the 20 residues and the sequence you can calculate the occurrence. So, when we get the occurrence then what to do?

Student: Multiply

Multiplied by weight; so for the 20 residues we have the weight right, so  $w(i)$ . So,  $n(i)$  we have right, we already got the occurrence. So, for each occurrence is multiplied by weight of  $i$  and you get the summation; summation equal to,  $i$  equal to 1 to 20. So, you get the weight, then you have to subtract with  $18(N-1)$  right. So, we have, we get the 4

Gs and 2 Ds and so on multiplied with the proper appropriate weights and subtract with  $18(N-1)$ . So, we could do this, you will get this number.

So, how to calculate the molecular weight?

Student: Sir composition.

Right, we need the occurrence. So, we get the occurrence. We know the values for each amino acid residues, the molecular weights. So, we calculate the  $n(i)$  multiplied by  $w(i)$  right, and you have to subtract the molecular weight of the water right, this is equal to  $18(N-1)$ ; so if you assign the values for all the residues and then calculate, and finally you get the number 8129.

(Refer Slide Time: 19:13)

**Amino acid property**

Total hydrophobicity =  $\sum n(i) \cdot hp(i)$

A, C, G, M, Y:	1
F, I, L, V, W:	2
D, E, H, K, R:	-2
N, P, Q, S, T:	-1

Seq 1: AILVA  
Seq 2: RSTVTS

Seq 1:  $1 + 2 + 1 - 2 + 1 = 3$

Seq 2:  $-2 - 1 - 1/2 - 1 = -4$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 11

Now, you can calculate the average properties, right just we get the composition and occurrence and the molecular weight, if we have the occurrence it has influenced various properties.

For example, if you have more number of charged residues like aspartic acid or glutamic acid or lysine or arginine, you can say the protein is highly charged right, whether it is acidic or it is basic. If we have more number of hydrophobic residues, then we can say this protein A is highly hydrophobic than protein B right, if we compare different sequences, we can use various properties of the individual amino acids right and use this

value to get the average values for the whole protein and this will tell you whether the protein is based on any property, it is high or low.

For example if you have a sequence 1 and sequence 2, AIKT here RSTV how for these sequences behave based on hydrophobicity. So, we can put here again VA and TS and if you see this sequence number 1; so we have values for the 20 residues right, for example, if you have hydrophobicity, 20 different amino acid residues, they have 20 different values. Or polarity or charge right, or tendency to form hydrogen bonds which residues have higher tendency to form hydrogen bonds. Among the 20, which is a polar residues? They have high preference to form hydrogen bonds right, likewise 20 residues, they have 20 values.

How far the residues are flexible? Some residues are highly flexible, some residues are rigid. How many rotatable bonds each residues can have? Right, likewise there are various properties right for each amino acid residues. So, if you have the values then sum up the numbers in a full sequence, that will tell you the characteristic features of the complete protein, based on the particular property. For example, if I have the sequence 1 and sequence 2 and I want to see which residue is highly hydrophobic.

So, if you see I have values for 20 residues. I give one example. For example, I classify the amino acids into the 4 groups right, one is hydrophobic, one is highly hydrophobic and one is the polar and one is charged. For the hydrophobic one, so, you put 1. If it is highly hydrophobic like isoleucine, leucine, valine or tryptophan I put 2. For the charged ones I put -2 and the polar ones I put -1. If this is the case, what is the value for sequence 1? A is equal to 1, plus I equal to.

Student: 2.

2; K equal to.

Student: -2.

-2; T equal to -1, V equal to 2, A equal to 1. So, this equal to 3-3; 0, this equal to 3, you take second one; what is second sequence, R equal to - 2 - 1 - 1 + 2 - 1 - 1; this equal to.

Student: - 4.

- 4; so compare these 2 sequences. You can see first one is highly hydrophobic, right than the second one, right. So, if we have any protein sequences right you can calculate different properties; here I give the numbers 1, 2, -2, -1, but we have the actual values for the 20 amino acid residues, which can be obtained by experiments, either a octanol experiment, water experiments, or the ethanol/water experiments by getting the relative solubilities, or computation derived scales. There are several computationally calculated values in the amino acid residues.

We can use exact values to see the average property of any protein.

(Refer Slide Time: 23:19)

**Amino acid property**

---

Average hydrophobicity =  $\frac{\sum n(i) \cdot hp(i)}{N}$

↓ occurrence      → no. of residues

↓ hydrophobicity

1. AMENLNMDSR  
2. LLYMAAAVMM

Compute total and average hydrophobicity

*Handwritten calculations:*

For sequence 1: n(A)=3, n(L)=2, n(E)=1, n(N)=1, n(M)=2, n(D)=1, n(S)=1, n(R)=1. Total = 13. Average = 1.3.

For sequence 2: n(L)=3, n(Y)=1, n(M)=3, n(A)=4, n(V)=1, n(M)=1. Total = 13. Average = 1.3.

Example calculation:  $\frac{7 \cdot 3 - 6 \cdot 3}{10} = -0.4$

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 11

So, I have the 2 sequences; now you can calculate the values, right using this equation or what is n(i)?

Student: Composition.

n(i) is the occurrence of each residues, right the occurrence of each residues, hp(i); that means, you see hydrophobicity of residue i right, hydrophobicity, you get the average you have to normalize with N; what is N?

Student: (Refer Time: 23:48).

Number of residues, right; so you take this equation. So, how many As here?

Student: 1 (Refer Time: 23:58).

Number of A equal to 1, M?

Student: 2

2, then what else, E?

Student: 1.

1; N?

Student: 2.

2; L?

Student: 1.

1; D?

Student: 1.

1, S 1, R 1; so we group the numbers right, if you see back right NPQST, right S, N right, that is equal to -1. So,  $-2 + -1 - 3$  and then see this group D E H K R; D E, H is not there, right. So, 1, 2, 3; 3 into equal to -6 and then what else you have? ACGMY is 1, this is 2, this equal to 3, L equal to 1; that is 2. So, total value equal to -4. So, we take the second equation, this sequence and apply the same equation here, right, this equal to this 4 is the total value you want. The average, 4 divided by 2, 3, 4, 5, 6, 7, 8, 9, 10; this equal to -0.4; you take the second sequence here; how many times A?

Student: 3

1, 2, 3.

Student: 2

L?

Student: 2

2, Y?

Student: 1.

M?

Student: 3.

1 2 3 and V?

Student: 1.

1. So, now, if you see these ACGMY; ACGMY 3, 6, 7, 7 into 1; this equal to 7; right. So, 1 multiplied by 1, equal to 7 and then L and V are 2; so  $2+1$ ;  $3$ ;  $3 \times 2 = 6$ , this will be?

Student: 13.

13; total will be 13. So, average equal to.

Student: 1.3.

1.3, right, 1, 2, 3, 4, 5, 6, 7 8, 9, 10, right 1.3; so in the 2 sequences one sequence is -0.4; one sequence is 1.3. So, what can we infer from these values?

Student: second sequences is..

Second sequence is more hydrophobic than the first sequence, right. So, we have the some sort of proteins which are highly prefer to be highly hydrophobic environment for example, transmembrane helical proteins. So, if you have some sequences which are predominantly hydrophobic then you can see that protein could be a transmembrane protein. So, likewise if you have different sequences, you calculate various features, not just hydrophobicity, various features you can calculate, and then you can compare with respect to the different features, and see whether you can get the function of any particular protein, if it is completely not annotated, you can see these could be probably a transmembrane protein or DNA-binding protein or whatever probable functions and so on.

So, in this case, you can, these properties will be very useful, we can get all the information from the sequence, you don't need the structure, in a sequence, we can calculate these things.