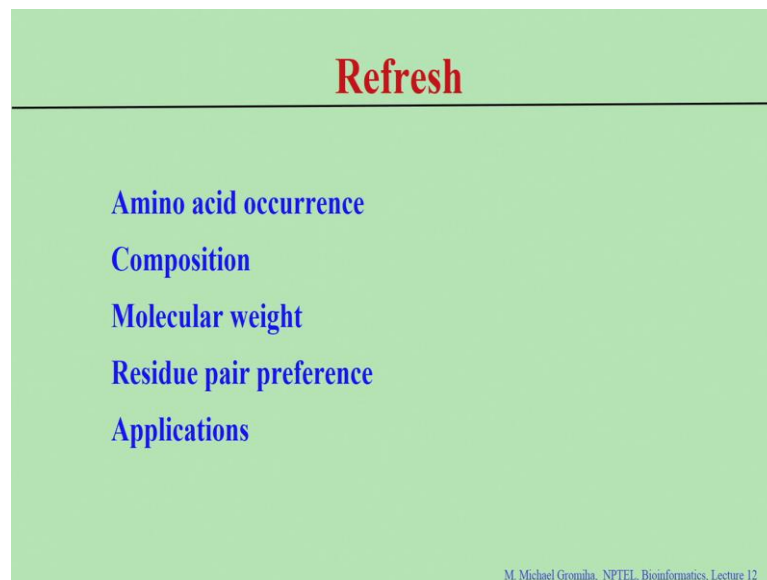


Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 12a
Hydrophobicity profiles

In this lecture, we will mainly discuss about the construction of hydrophobicity profiles and the applications. Earlier we discussed all different types of applications of these amino acid sequences. So, what are the various parameters or features or properties we discussed in the previous class?

(Refer Slide Time: 00:37)



Student: So, amino acid composition.

Amino acid occurrence, amino acid composition.

Student: Molecular weight.

Molecular weight, pair preference and so on. So, what is amino acid occurrence?

Student: So, number of time.

It is essentially.

Student: Occurrence.

Number of times each amino acid occurs in a protein sequence, then composition.

Student: Normalized

Normalized with the chain length. So, a difference between amino acid occurrence and composition is normalization with the number of residues in a sequence. Then we discussed about molecular weight. So, you can calculate the molecular weight?

Student: Hm.

By substituting appropriate weight for each amino acid residues and subtracted with.

Student: water.

The water molecules, $18(N-1)$ water molecules, because when you form peptide bonds, it's elimination of one water molecule. Then we discussed about the pair preference, how far the residues occur next to each other: A with A, A with D and so on. Then we will discuss about couple more applications. One major application we discussed, is how to distinguish between different types of proteins. We have proteins, with different functions, we have different structures. So, whether it is possible to distinguish between these 2 types of proteins based on the amino acid sequence information.

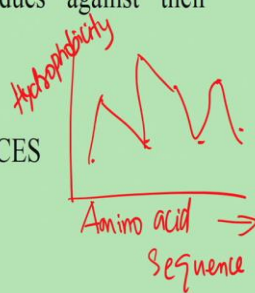
So, we discussed based on the parameter or based on the property, amino acid composition, we used amino acid composition for the 2 sets of proteins and for unknown ones, we compared with the known ones, and then we decide depending upon the deviation. We can also do with the correlation the also we can use different properties. So, also we discussed about the amino acid properties for a average property for any given sequence, each amino acid have the unique values. So, add up together normalized with the chain length, that will give the average amino acid property values.

(Refer Slide Time: 02:31)

Hydrophobicity profile

Hydrophobicity profile is simply the plot of the hydrophobicity indices of the residues against their sequence numbers.

E.g. SAMPLEDATAWITHHYDINDICES

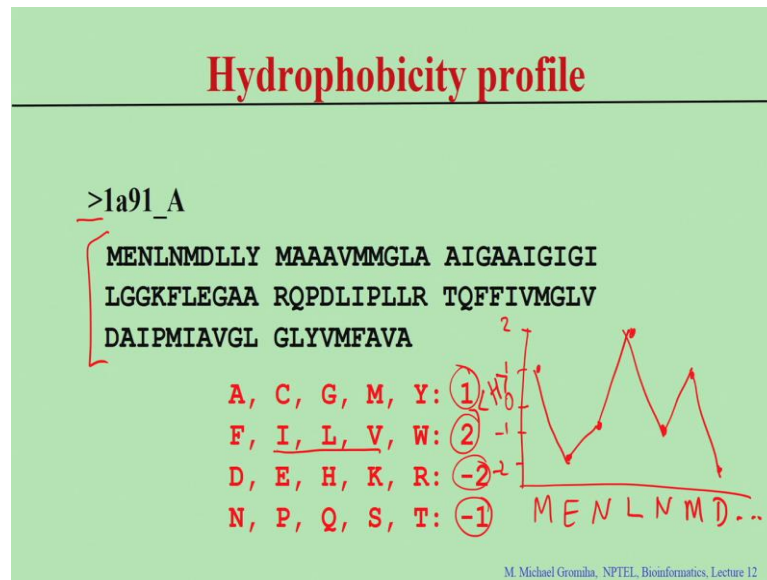


M. Michael Gromiha, NPTEL Bioinformatics, Lecture 12

So, now what is hydrophobicity profile? Because name itself tells, it is the plot of the hydrophobicity indices versus amino acid sequence. For example, if you take any protein sequence. So, we can make a 2D plot. So, X axis, we give about the amino acid sequence. Y axis - you write the hydrophobicity. Each amino acid has a value, depending upon the hydrophobic character of the residues.

So, depending among the amino acid sequence, the residues in the sequence, we can make a plot for each value. When you connect then we will get a plot. So, we will see how to construct a plot with an example.

(Refer Slide Time: 03:28)



We have the sequence, for example; this is the amino acid sequence in which format?

Student: Fasta format.

Fasta format because you can say they started with the > symbol, here we have the amino acid sequence. So, I made some numbers for the 20 amino acid residues, based upon their hydrophobic behavior. For example, if you take the highly hydrophobic residues like isoleucine, valine, leucine, phenylalanine, I put value of 2 and less hydrophobic, I put the value of 1, and the polar residues I give -1 and the charge residues I give value of -2.

So, now you take the amino acid residues versus these hydrophobic indices, what is a hydrophobicity profile? It is a plot connecting?

Student: Sequence.

A sequence versus the values right. So, here what is the sequence?

Student: M.

M.

Student: E.

E.

Student: N.

NL.

Student: N.

N M D and so on right. So, here you can put the hydrophobicity value. So, first one is M, what is the value for M?

Student: 1.

1. So, you can conclude here this is 0. So, here you can put 1, 2 -1 -2. So, M equal to 1. So, I plot here and the second one is E, what is the value of E?

Student: -2.

-2 here and N.

Student: -1.

N -1, L +2, this N -1, M +1, D -2 and so on. Then we connect. We connected. So, when you construct a plot, we can see some sort of patterns. This hydrophobicity profile will provide you a specific pattern which will be helpful to identify some secondary structures, or you can see any specific motifs, or you can see any segments which traverse the membrane and so on.

So, there are various software available to make a plot right. So, one of this is available in the literature, that is called 3DInsight, this will consider various properties, it will take the hydrophobicity values along with other properties to make a graph. X axis is the amino acid index and Y axis is a hydrophobicity profile or any different properties. In this slide, when I made this profile, I use a value of 1 2 -2 -1. Actually all the 20 amino acid residues contain specific values right.

(Refer Slide Time: 06:14)

Sample data			
Nozaki-Tanford-Jones (H_p)			
A: 0.87	D: 0.66	C:1.52	E: 0.67
F: 2.87	G: 0.10	H: 0.87	I: 3.15
K: 1.64	L: 2.17	M: 1.67	N: 0.09
P: 2.77	Q: .00	R: 0.85	S: 0.07
T: 0.07	V: 1.87	W: 3.77	Y: 2.67

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 12

So, I give couple of examples, one is the Nozaki-Tanford-Jones scale this is the first scale derived for the 20 different amino acids, you see experimentally. Directly we cannot calculate the hydrophobicity values, so they did indirect way, that is a relative solubility of each amino acid residues in water as well as in ethanol. So, they did the relative solubility and converted the solubilities into hydrophobicity. So, if you look into this 20 different amino acid residues we can see specific residues which are high values and some of them may have less values.

So, if you see this one can you name few residues, which are highly hydrophobic?

Student: isoleucine, tryptophan.

Right tryptophan, isoleucine.


Student: Phenylalanine.

Right phenylalanine right. So, these residues are hydrophobic because they contain aliphatic groups or aromatic groups. So, they are hydrophobic. So, in the partition coefficient, even the relative solubility, they also showed that these residues are highly hydrophobic. On the other hand, if you look into the charged residues or polar residues for example, you have a serine or you have aspartic acid or you have the glutamic acid, you can see that the values are less, compared to the hydrophobic residues.

So, if you make an average value. So, you can easily discriminate the hydrophobic and hydrophilic residues with these numbers. This is one example which you obtain experimentally. Likewise there are several scales available, I show another example.

(Refer Slide Time: 07:42)

Sample data			
Ponnuswamy-Gromiha (H_{gm})			
A: 13.85	D: 11.61	C: 15.37	E: 11.38
F: 13.93	G: 13.34	H: 13.82	I: 15.28
K: 11.58	L: 14.13	M: 13.86	N: 13.02
P: 12.35	Q: 12.61	R: 13.10	S: 13.39
T: 12.70	V: 14.56	W: 15.48	Y: 13.88

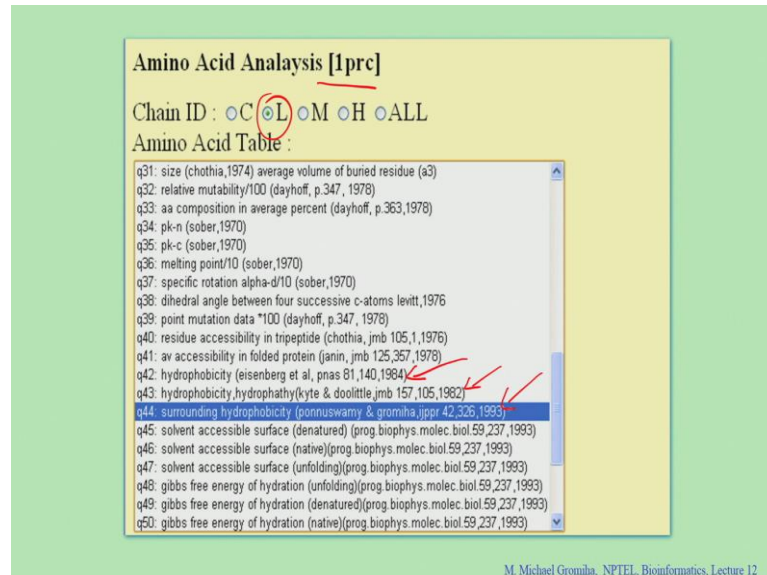


This we obtain from computation analysis. What we did here? Here we considered different dataset of proteins and assign the values for each residue which are influenced with the neighboring residues. They identified the residues which are occurring within the limit of any specific radius and then see; what are the residues which are within this limit. And they assigned the experimental value and then finally, they derived these final scales, but I will explain about the development of the scale in later classes, fine.

So, if you see in this scale here also you can see the real situation. As we discussed earlier, they have some of the hydrophobic residues, see like cysteine, tryptophan, isoleucine they are highly hydrophobic and several polar residues and the charge residues like lysine and you can see the case of aspartic acid and glutamic acid and these residues, they are polar in nature. And major difference you find from these 2 scales, that is mainly proline, proline they did the relative solubility, proline as a ring. So, you can see this is hydrophobic in this case, but if you look into the location of this proline, it is mainly surrounded with polar residues, this is the reason why proline is polar in this scale. Many scales, they assign proline as polar residues.

So, you can see some differences and similarities and most of the scales, currently we have more than 100 indices available. So, they have qualitatively, they have similar behavior. So, I show one example where we can construct the plot.

(Refer Slide Time: 09:21)



The screenshot shows a software window titled "Amino Acid Analysis [1pre]". At the top, it displays "Chain ID : C L M H ALL". Below this is a section labeled "Amino Acid Table :" containing a scrollable list of properties. The list includes:

- q31: size (chothia,1974) average volume of buried residue (a3)
- q32: relative mutability/100 (dayhoff, p.347, 1978)
- q33: aa composition in average percent (dayhoff, p.363,1978)
- q34: pk-n (sober,1970)
- q35: pk-c (sober,1970)
- q36: melting point/10 (sober,1970)
- q37: specific rotation alpha-d/10 (sober,1970)
- q38: dihedral angle between four successive c-atoms levitt,1976
- q39: point mutation data *100 (dayhoff, p.347, 1978)
- q40: residue accessibility in tripeptide (chothia, jmb 105,1,1976)
- q41: av accessibility in folded protein (janin, jmb 125,357,1978)
- q42: hydrophobicity (eisenberg et al, pnas 81,140,1984)
- q43: hydrophobicity,hydrophathy(kyte & doolittle, jmb 157,105,1982)
- q44: surrounding hydrophobicity (ponnuswamy & gromiha,jjppr 42,326,1993)
- q45: solvent accessible surface (denatured) (prog.biophys.molec.biol.59,237,1993)
- q46: solvent accessible surface (native)(prog.biophys.molec.biol.59,237,1993)
- q47: solvent accessible surface (unfolding)(prog.biophys.molec.biol.59,237,1993)
- q48: gibbs free energy of hydration (unfolding)(prog.biophys.molec.biol.59,237,1993)
- q49: gibbs free energy of hydration (denatured)(prog.biophys.molec.biol.59,237,1993)
- q50: gibbs free energy of hydration (native)(prog.biophys.molec.biol.59,237,1993)

Red arrows point to the entries for q42, q43, and q44. The entry for q44 is highlighted in blue. At the bottom right of the window, the text "M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 12" is visible.

So, here is a list of several properties and here we you have several hydrophobicity values for example, the hydrophobicity values given with Eisenberg also the Kyte and Doolittle scale, they developed in 1982 then this is the hydrophobicity scale. So, you have different scales, if you want to make a plot, you first you take the PDB code or amino acid sequence, this is the PDB code and here we take the L chain.

(Refer Slide Time: 09:51)



So, we selected this hydrophobicity scale and if you click, then we will get the plot. What is the X axis?

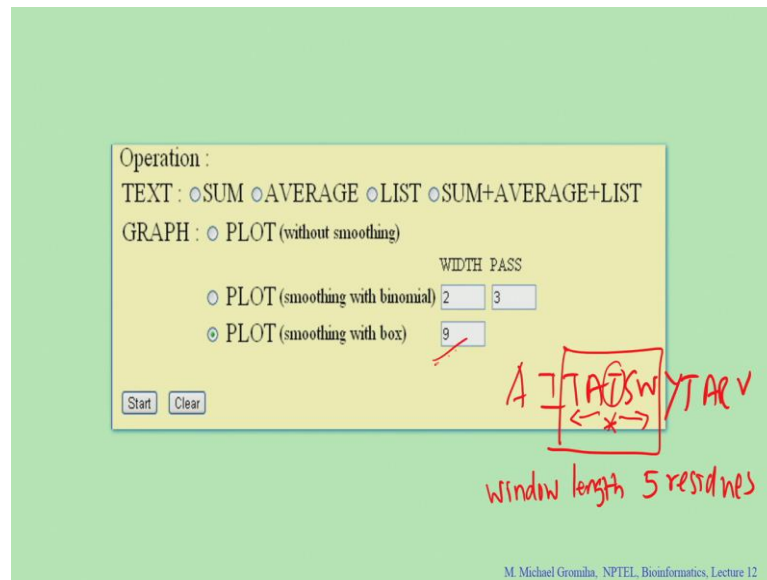
Student: Chains.

Amino acid sequence, what is the Y axis?

Student: Hydro.

Hydrophobicity value right. So, you can see the plot connecting the sequence and the plot. Some cases you can infer the information, some cases, it is not. If you see here, it is kind of clusters. So, we can see the zigzag positions, up and down on this case. So, if you cannot infer anything from the single residue plot, most of the cases, you can find some patterns, then you can also try to get some window average.

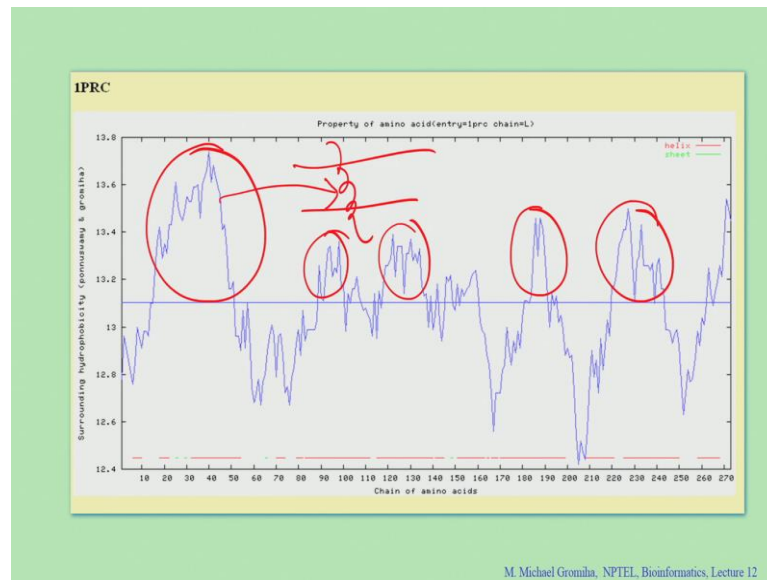
(Refer Slide Time: 10:25)



In this case it is possible to get the average. So, for example, if you click on the plot with smoothing with box, in this case you can smooth for any window length, for example, if you can amino acid sequence like this.

So, if you take the single residue plot, they plot for each residue. For each residue they have the hydrophobicity value, they plot. We carry window length for example, any residue if you take a window length of 5 residues for example, window length 5. So, what they do? So, they make a window, considering 5 residues, 2 from left side, 2 from side, where this is the central one. So, they have make a window 5. So, what they do? They calculate the average value and plot for the central one. For the first residue we do not have the left side residues. First 2, if you have a window length of 5, and all other residues take a 5 residues length.

(Refer Slide Time: 11:36)



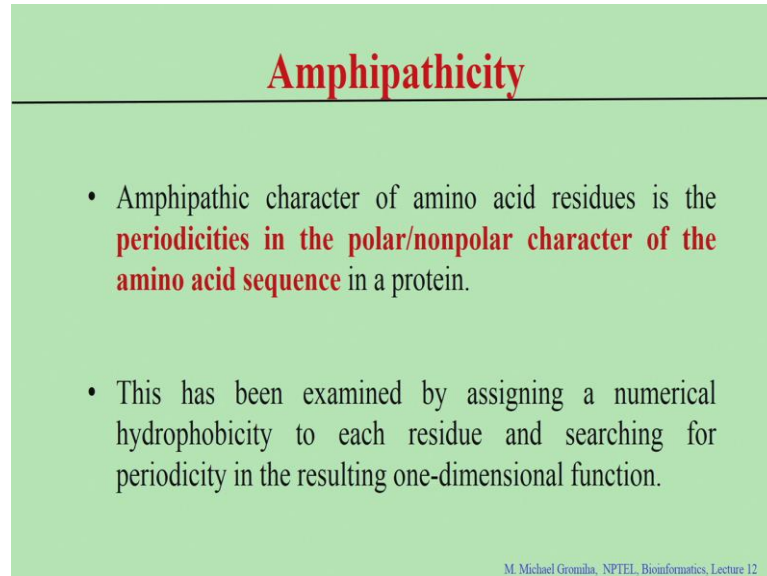
So, now if you do this for example, if we take the smoothing box of 9, then will get the plot like this. Now can you see the difference between the previous plot and this plot? Yes. Here you can easily identify some regions which are highly hydrophobic in nature and some regions which are less hydrophobic in nature for example, if you see here all the residues are highly hydrophobic and you can see somewhere here, someplace here, someplace here, and someplace here, here you can see these completely polar charged residues or polar residues.

So, will you have this one, this will tell you that there is a stretch of residues, but there highly hydrophobic in nature with one or 2 polar residues in between. Because of that one or 2 residues, in the first plot we can see a zigzag, look at that polar residues has less hydrophobic is down. Here because of the average, that compensates maybe somewhere here, the residues are highly hydrophobic and interestingly if you see this protein, is a membrane protein as I discussed earlier, the proteins which are embedded in membranes right.

So, these residues, they span in a membrane, if you see there is a membrane. So, here this is the protein right. So, the region is a membrane. So, this response resembles this one. So, you can use you can make a plot. This plot has some applications to identify the regions, which are inserted in the membrane right. So, if you can use this hydrophobicity profiles to get several other information.

So, this is another behavior which are related with hydrophobicity, there is called amphipathic character.

(Refer Slide Time: 13:05)



Amphipathicity

- Amphipathic character of amino acid residues is the **periodicities in the polar/nonpolar character of the amino acid sequence** in a protein.
- This has been examined by assigning a numerical hydrophobicity to each residue and searching for periodicity in the resulting one-dimensional function.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 12

What is a meaning of amphipathic?

Student: has both charge.

Both right. 2. *Amphi-* means 2 right. So, where this is high or low. So, this will give you the information regarding periodicities of the polar and non-polar characteristics of sequence. How do get that? You have the numerical hydrophobicity values, we assign the values and then we see if there are there any periodicity in one dimensional plot. How to do this? For example, if we take alpha helical segment. How many residues per turn?

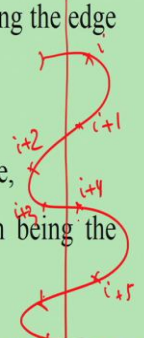
(Refer Slide Time: 13:41)

Amphipathicity: α -helices

The residues of an α -helical segment are considered on four adjacent edges along the direction of the helical axis. The average hydrophobicity of the residues constituting the edge i ($i = 1,4$) is given by

$$\alpha_i = (\sum h_{i+j})/n,$$

where n is the total number of residues in the edge, j increases at an interval of 4 from 0 to m , m being the number of residues in the helix; h is the hydrophobic index of the residue.



M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 12

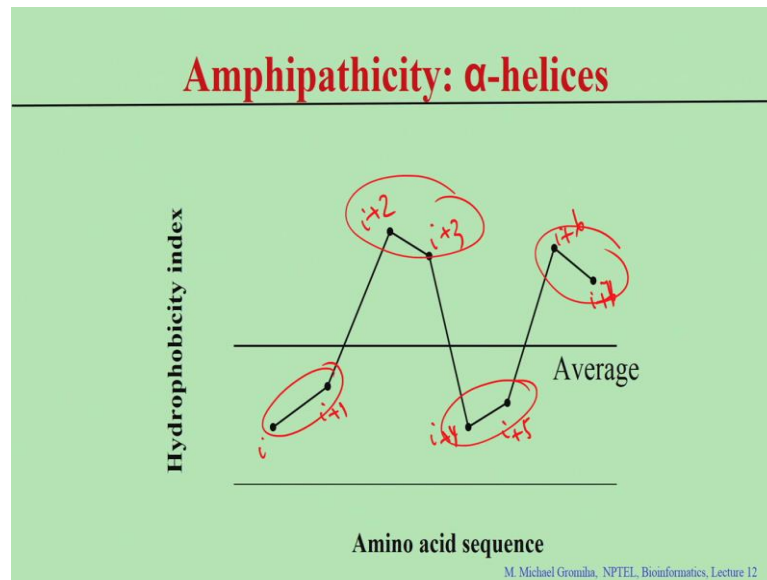
Student: 3.6.

3.6 residues per turn. So, if you have a helix, it start from here right. So, one turn is up to here, this way and this way right. So, you have a helical axis, you can put the residues 1 2 3 4, the same way 1 2 3 4. So, you can make it. So, in this case you can see the 2 residues are on side and 2 residues on the other side. So, they show a periodicity of these residues at different positions, this is i , for this is i , this is $i+1$, $i+2$, $i+3$, $i+4$. So, you can see the periodicity $i+5$. So, between this i and $i+4$.

So, these 1 2 3 4 and this 5. 1 and 5, there would be on the same place, likewise you can see the 2 and 6. So, if you have a helical segment, you can make on 4 adjacent edges you can put 1 2 3 4 in the direction of the helical axis, then you can calculate the average hydrophobicity of each edge for example, we take this, these plus same is here, same is here.

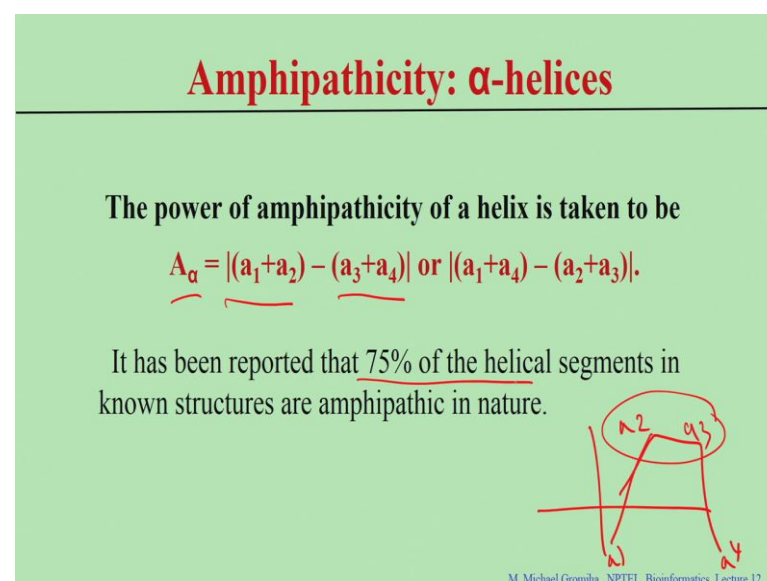
So, you can see $i+j$ with the interval of 4, and see how many residues in that particular place.

(Refer Slide Time: 15:06)



So, for example, if you see this periodicity this is i , this $i+1$, $i+2$, $i+3$, $i+4$ and $i+5$. You can see a periodicity, 2 residues are less hydrophobic here and 2 residues have high hydrophobicity and then again 2 less and 2 high. So, if you have this type of patterns, then we can say these residues are part of alpha helix, or these residues can suit alpha helix right. So, if you have a sequence and if you make a plot with hydrophobicity and if you see any of these patterns, then you can say that these residues can form alpha helix. Then the second is, how to calculate the power of amphipathicity. For example if you have 2 helices, they are amphipathic and which one is more amphipathic.

(Refer Slide Time: 15:59)



For in this case, you can calculate the power of amphipathicity alpha.

So, if 1 and 2 are high and 3 and 4 are less then take the first 2 and the second, third and fourth and get the absolute difference. In this case, here take these numbers as 1 and these numbers as 2, this is $i+6$ and this is $i+7$. So, then take the difference, this will give you the power of amphipathicity, you can see the A_α , either $a+1$, $-(a_3 + a_4)$ or if $a_1 a_4 - (a_2+a_3)$, this is in this case for example, if it is like this right.

So, this is 1 2 3 4; $a_1 a_2 a_3 a_4$, if this is the case, this will come together $a_2 + a_3$ and this will come together, you can see difference. So, from this one, we can see whether they are amphipathic or not, but if you look in to the literature about 75% of the helical sequence of known structures were amphipathic in nature. This case, if you use this type of profiles, we are able to predict to some level of acquires at least to 70, 65 to 70% of accuracy in any sequence. This is for alpha helix. How about in the case of beta sheets? In the case of beta strands right.

(Refer Slide Time: 17:27)

Amphipathicity: β -strands

A β -strand segment is considered to have two faces and the average hydrophobicity of residues constituting the face i ($i = 1, 2$) is given by

$$\beta_i = (\sum h_{i+j})/n,$$

where n is the total number of residues in the face, j increases at an interval of 2 from 0 to m , m being the number of residues in the strand;

M. Michael Gromiha, NPTEL Bioinformatics, Lecture 12

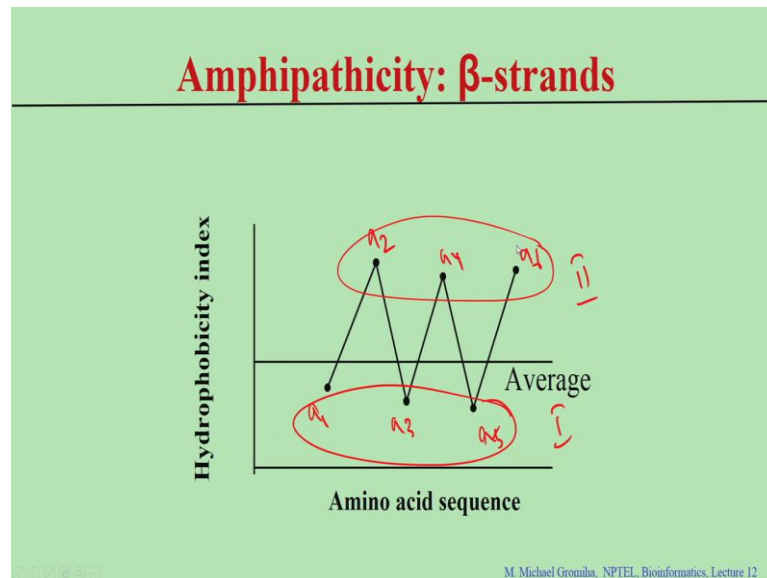
The slide includes a diagram of a beta-strand with two faces, one on each side, indicated by red arrows and circles. The formula for β_i is underlined in red.

So, you can see there are 2 faces; you have 2 faces, one is here, one is here. The alternate faces. So, one is having high, another is low hydrophobicity.

So, in this case you can calculate β_i , this is a summation h_{i+j} by n , where here these intervals of 2, the helix we have interval of 4, in the beta sheet interval of 2, with up to n ,

where n is the number of residues in the strand. So, to calculate the β_i is equal to Σh_{i+j} divided by n , for n is the number of residues in each face, either here or here.

(Refer Slide Time: 18:04)



So, now we have the patterns, you can see the X axis is amino acid sequence, Y axis hydrophobicity index. So, here these are the hydrophobic values of different amino acids. So, this is the average value. So, this a_1 , a_2 , a_3 , a_4 , a_5 , a_6 . So, if you see the power of amphipathicity, you can calculate, may be this is one side, this is another side, this is one face, second face, then add up all these together, take the average and add here everything together, and take the average finally, you can get the values right.

(Refer Slide Time: 18:37)

Amphipathicity: β -strands

The amphipathicity index of a strand is computed using the equation,

$$A_{\beta} = |\beta_1 - \beta_2|.$$

The structural analysis showed that about 65% of the β -strands possess amphipathic character.

M. Michael Gromiha, NPTEL Bioinformatics, Lecture 12

You can get $\beta_1 - \beta_2$. So, they get the amphipathic behavior of this particular beta sheet.

So, the structural analysis also showed that about 65 percent of the beta sheets, they possess amphipathic in nature. If you take the all the beta sheets in the known structures and construct the plots and you can see about 65 percent, they are amphipathic in nature. So, now, what are other applications of this hydrophobicity profiles, here we discuss 2 cases, one is the alpha helices. What is periodicity in alpha helix?

Student: 1 + 4.

Yeah 2 on one side, and 2 on other side. What is the periodicity in beta sheets?

Student: 2 1 1 2.

1 1, you have the alternate high and low hydrophobicity patterns.

(Refer Slide Time: 19:19)

Patterns

Identify the pattern of hydrophobic residues for membrane spanning helical proteins

E.g. AILVGYWFFVVA

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 12

So, some cases for example if you get the hydrophobic behavior, this is the average value. For example, this is the average value if you take, and all the residues are highly hydrophobic, For example, I put this sequence, A is hydrophobic, I is hydrophobic, L is hydrophobic. So, all residues are hydrophobic. So, we have the pattern of highly hydrophobic residues. If this is the case, this will resemble a type of protein, a type of segment, which type of segment?

Student: (Refer Time: 19:49).

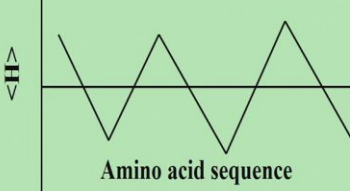
This is membrane spanning alpha helices right. So, this is membrane spanning alpha helices. So, if you plot, make a plot, you can get some information from primary sequences right. So, without having any information, if this character provides a specific information, at least you can develop from that point of view. So, constructing hydrophobicity profiles, will give some information from the sequence.

(Refer Slide Time: 20:14)

Patterns

Amphiphathic character of β -strands by alternative hydrophobic-hydrophilic residues

E.g. **A**K**I**N**I**H**V**T**F**K**I**K**L**P



Amino acid sequence

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 12

This is another pattern, just I discussed earlier which is the beta sheets, amphipathic character. So, because you can see the red one, is the hydrophobic, and the black one is the hydrophilic. So, you can say alternating hydrophobicity. So, this will give the pattern for the beta strands.

(Refer Slide Time: 20:29)

Pattern Definition

1 2 4 5 6 7 8 10 14 16 17

[LIVM]-[VIC]-x(2)-G-[DENQTA]-x-[GAC]-x(2)-[LIVMFY](4)-x(2)-G

1. Use capital letters for amino acid residues
2. Use "[...]" for a choice of multiple amino acids in a particular position. [LIVM] means that L, I, V, or M can be in the first position

DRY GXXG

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 12

So, now in the amino acid sequence, it is also possible to identify any specific motifs are patterns. In this case we define a specific definition for defining patterns for example, you take any sequence like for example, 5 residue segments or 10 residue segments

whether they possess any specific patterns. So, in some cases we discussed earlier about conservation. So, what is conservation?

Student: Residues conserved

Yeah may position a particular residue occupies the same position in all the homologous sequences. In this case they residue is said to be conserved, if your residue is certainly conserved then you have to keep that particular residue in all the positions for example, if you have specific motifs GXXG motif, or any specific motif, DRY motif, this is for something for the GPCR, this is for a disulphide bridge forming enzymes. So, if they have some specific motifs then we can search any specific motifs in the whole database and see whether there is any characteristic pattern for that particular set of proteins.

So, in this case we define patterns using some sort of notations. So, here we show one sequence, we use some residue names for example, here this means the residue glycine is conserved at that particular position, that maintains the same residue at the particular position. So, we cannot change this, for this place as well as this place. They put x, what is the meaning of x?

Student: Any residues.

Any residues. You can put all the 20 amino acid residues, any residue is allowed at this particular position, then we have these numbers. Number means?

Student: 2 time.

Number of times, with 2 means 2 times. So, we can follow the same notation. Then we use the square brackets for example, if here we put L I V M, this will give you the choice of amino acids, any of these 4 residues. Among these 4, either L or I or V or M, any residue can accommodate at the position number one.

(Refer Slide Time: 22:46)

Pattern Definition

3. Use "{...}" to exclude amino acids. {CF} means C and F should not be in that particular position
4. Use "x" or "X" for a position that can be any amino acid.
5. Use "(n)", where n is a number, for multiple positions; x(3) is the same as "xxx"

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 12

Then we have this curly brackets, if this is brackets.

Student: exclude.

This is excluded, that this is not allowed for example, if you put CF, C and F should not be in that particular position, they give the exception that these are not allowed. Then you put x for any amino acid, n means number of times. If I show this one, how many amino acids in this pattern? This is 1 2, here 2, 2 plus 2, 4 5 6 7 8 10 14 16 17; 17 amino acids in this pattern.

(Refer Slide Time: 23:31)

PIR: Pattern Definition

V V VVG Q P A IA LMFY AA G
[LIVM]-[VIC]-x(2)-G-[DENQTA]-x-[GAC]-x(2)-[LIVMFY](4)-x(2)-G

Illustrates a 17 amino acid peptide that has:

- L, I, V, or M at position 1;
- V, I, or C at position 2;
- any residue at positions 3 and 4;
- G at position 5 and so on

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 12

So, how to write a pattern in this case, if you take this one? So, position number 1, which residue can come in position number 1?

Student: Anything

Anything, for example I put V, for the second position again we put V, third and forth?

Student: Anything

Anything, you can put a 2 times V again. And this position?

Student: GG.

G because here we cannot do anything. So, you have to use G because G is conserved. And in this here?

Student: N

NQ you can put Q. So, where is x, x we can put P right. So, here.

Student: A.

A then 2, any residues?

Student: 2 3.

I, A, I A and here.

Student: L.

L.

Student: M.

M.

Student: F.

F.

Student: V.

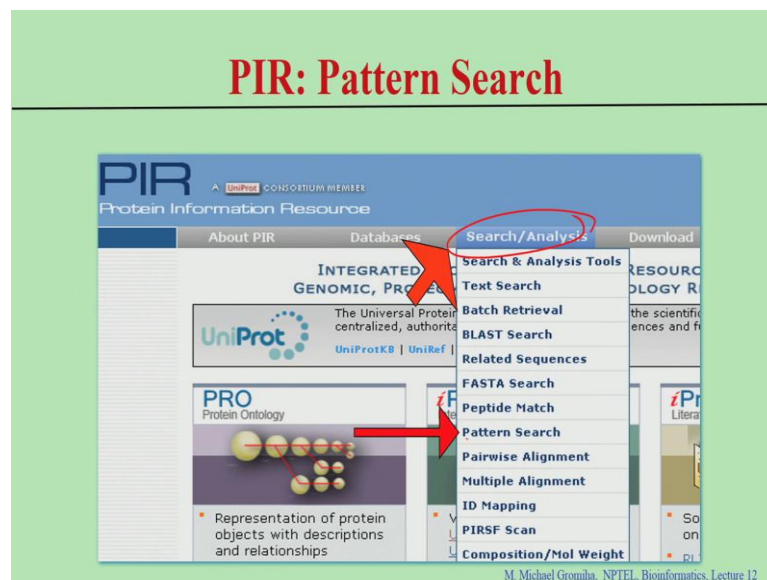
V. So, because L is allowed, M is allowed and F is allowed and V is allowed 4 times because 4 times you put 4, then 2x, AA and finally, G. So, we can write a pattern and then we can use this pattern to see whether you can see this type of patterns in all the sequences of particular type. I will explain one with an example.

So, now this we can write a code to get the pattern, and also this tool is available in the PIR; what is PIR?

Student: protein information resource

Protein information resource, as they first initially developed for the protein sequences right.

(Refer Slide Time: 25:00)



And then they develop some tools for analyzing sequences and here, if you go the search analysis, there is a tool called this pattern match.

(Refer Slide Time: 25:11)

PIR: Pattern Search

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 12

So, we go with this pattern match. So, we can give the pattern, I give a same pattern here LIVM same, and if you submit.

(Refer Slide Time: 25:18)

[LIVM]-[VIC]-x(2)-G-[DENQTA]-x-[GAC]-x(2)-[LIVMFY](4)-x(2)-G

Protein AC/ID	Protein Name	Length	Organism Name	PIRSF ID	Match Range
096777/D96777_HELVI	Cyclic nucleotide and voltage-activated ion channel	678	Helicoverpa zea (Tobacco budworm moth)		477-493
097119/O97119_LIMBO	Cyclic nucleotide-gated ion channel LCN1	900	Limulus polyphemus (Atlantic horseshoe crab)		532-546
P04175/SPAQ_SALTY	Surface presentation of antigens protein spaQ	86	Salmonella typhimurium	PIRSF004669	4-20
P04182/SPAQ_SALTI	Surface presentation of antigens protein spaQ	86	Salmonella typhi	PIRSF004669	4-20
P04342/SPAQ_SALSE	Surface presentation of antigens protein spaQ	86	Salmonella senftenberg	PIRSF004669	4-20
P04343/SPAQ_SALTF	Surface presentation of antigens protein spaQ	86	Salmonella typhisus	PIRSF004669	4-20
P04C03/PCP_EC03	Catabolite gene activator; (AltName: Full-cAMP receptor protein; AltName: Full-cAMP regulatory protein)	210	Escherichia coli (strain 1210)	PIRSF003111	38-46
P04C03/PCP_EC06	Catabolite gene activator; (AltName: Full-cAMP receptor protein; AltName: Full-cAMP regulatory protein)	210	Escherichia coli O6	PIRSF003111	38-46
P04C03/PCP_EC07	Catabolite gene activator; (AltName: Full-cAMP receptor protein; AltName: Full-cAMP regulatory protein)	210	Escherichia coli O157:H7	PIRSF003111	38-46
P22972/CNGAL_HUMAN	CGMP-gated cation channel alpha-1; (AltName: Full-CNG channel alpha-1; Short=CNG-1; Short=CNG1; AltName: Full-cyclic nucleotide-gated channel alpha-1; AltName: Full-cyclic nucleotide-gated channel, photoreceptor; AltName: Full-cyclic nucleotide-gated cation channel 1; AltName: Full-foot photoreceptor CGMP-gated channel subunit alpha)	690	Homo sapiens (Human)	PIRSF002402	506-522
P22974/CNGAL_MOUSE	CGMP-gated cation channel alpha-1; (AltName: Full-CNG channel alpha-1; Short=CNG-1; Short=CNG1; AltName: Full-cyclic nucleotide-gated channel alpha-1; AltName: Full-cyclic nucleotide-gated channel, photoreceptor; AltName: Full-cyclic nucleotide-gated cation channel 1; AltName: Full-foot photoreceptor cGMP-gated channel subunit alpha)	694	Mus musculus (Mouse)	PIRSF002402	498-514
P36600/KAPK_SCHPO	CAMP-dependent protein kinase regulatory subunit; (Short=PKA regulatory subunit)	412	Schizosaccharomyces pombe (Zelensky strain)	PIRSF000549	171-187; 305-321
P49050/KAPK_YSTMA	CAMP-dependent protein kinase regulatory subunit; (Short=PKA regulatory subunit)	525	Yarrowia lipolytica (Yeast)	PIRSF000549	241-257; 375-391

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 12

And you will get the sequences which contain the particular pattern. So, if you click on this, any of this sequence if you click.

(Refer Slide Time: 25:25)

Length = 210 Click on a bar to show its sequence; to copy and paste it, press ctrl then ctrl-z.

POACJ8 1 210

PF00027

PF00325

1 MVLGKPKQTDPTLEWFLSHCHIHKYPKSTLIHQGEKAETLYYIVKGSVAVLIKDEEGKEM

61 ILSYLNQGDFIGELGLFEEGQERSAVVRAKTACEVAEISYKFRQLIQVNPDLMLRLSAQ

121 MARRLQVTSEKVGNLAFLDVTGRIATLLNLAKQPDANTHPDGMQIKITRQETIGQIVGCS

181 RETVGRILKMLEDQNLISAHGKTIIVYVTR

Pattern: LIHQGEKAETLYYIVKG

[LIVM]-[VIC]-x(2)-G-[DENQTA]-x-[GAC]-x(2)-[LIVMFY](4)-x(2)-G

New β -signal motif : $P_o x G h_y x H_y x H_y$

[K,R,H,Q,N,S,T]G[I,V,L,F,M,Y,W,A,C].[I,V,L,F,M,Y,W].[I,V,L,F,M,Y,W]

Algorithm

K. Imai, M.M. Gromiha and P. Horton (2008) Cell

M. Michael Gromiha, NPEL, Bioinformatics, Lecture 12

So, will get this full sequence and you can see the pattern in this sequence, where does start from? Here right VKG, from here to here. So, if you check this is L, L is here and the I is here and 2 residues here and then G, this is a conserve G it is here, and E is here and then K and then A and then E and T and L Y, L is here Y is here and I right, then 2x, V and K and final G, is final G, that is fine, this works, right the program works.

So, you can get this type of patterns and then see whether each specific pattern is important for a particular set of proteins. I explain with one example. So, this is your specific motif for the beta signal; that means, for the insertion and assembly of beta barrel membrane proteins, this is a type membrane protein I discussed earlier right. So, this type of motif is essential. Experimentally they observe, computationally we can do this analysis using this search.

So, here P_o means polar. So, any polar residues, and here there is one x. So, here put x, here dot, and G. So, G is conserved and any hydrophobic residues is H_y and any x and hydrophobic here and x and hydrophobic. There is a difference between this small h_y and capital H_y ; small h_y includes these small residues, alanine, cysteine, but capital H_y , it does not include this A and C.

Now, the question is why experimentally they showed that this specific motif that is important, that is true or not, how to verify?

Student: If it is present in.

Right if it is present or not. So, first we take all proteins of beta barrel membrane proteins, and then easily we can write a pattern because if you go here and if we write this pattern then I will get the list of proteins, or other way you can do it, first download all the beta barrel membrane proteins and then see whether this pattern is available or not. The 2 options, one is the whole sequence, we can search or we can see where this is important, of the N-terminal or the C-terminal; mainly C-terminal for the insertion, you can see, take this C-terminal forty residues and see whether the pattern is available or not, fine.

So, if you find this pattern then you are happy. Then next question is how to verify, this is important for this particular type of proteins? You have to compare. So, this should be present in this type of proteins, and should not be present in other type of proteins. So, we can construct different datasets; where shall we get the sequences?

Student: UniProt.

UniProt database right. So, then you get the different sequences like normal globular proteins, we can see the inner membrane proteins, all beta proteins, different type of proteins you can construct and see whether is the pattern is available or not, if available?

Student: There is not.

That is not specific, if it is not available, very specific. When we do this, we can find this particular motif, only for the beta barrel proteins; that means, this signal is important for the insertion assembly, then we can find something. Likewise you can define several motifs and you can see some novel patterns and you can explain why this is important and what is the main use of these particular motifs in any sequences, you can do lot of analysis with different types of proteins.