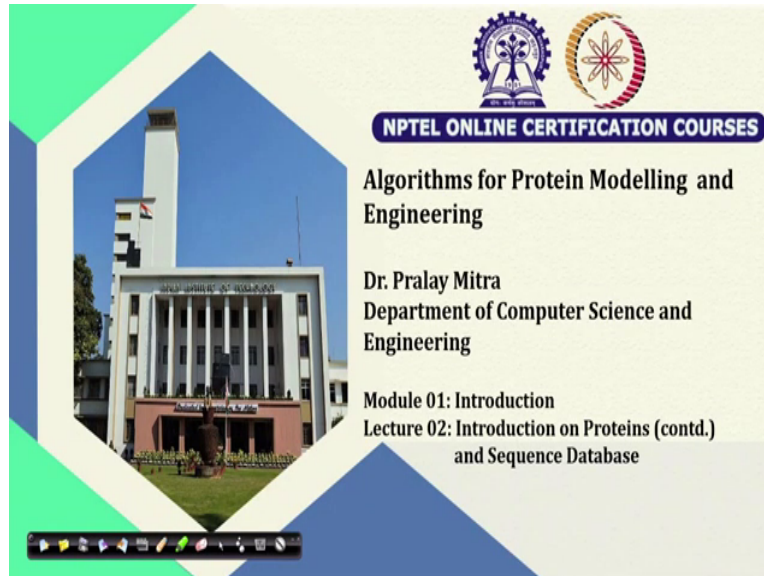


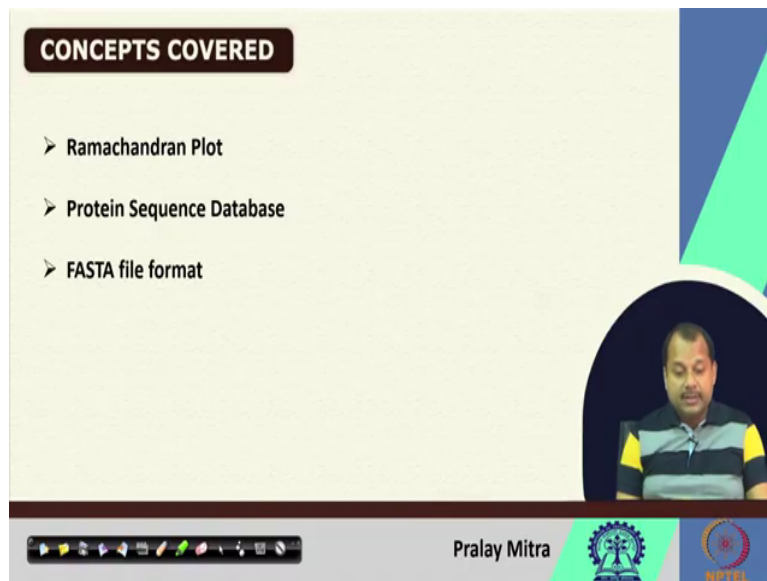
Algorithms for Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture: 02
Introduction on Proteins (Contd.) and Sequence Database

(Refer Slide Time: 00:15)



Welcome back. We shall continue the introduction module - introduction to protein and after a few introductory slides, then we shall discuss the sequence database of the protein. Because these databases will be useful for us when we shall develop algorithms or test/benchmark designed algorithms. The sequence database, as well as the protein structure database, will be useful for us. So, today in this lecture, I shall cover the sequence database then we shall also cover the structure database.

(Refer Slide Time: 00:49)

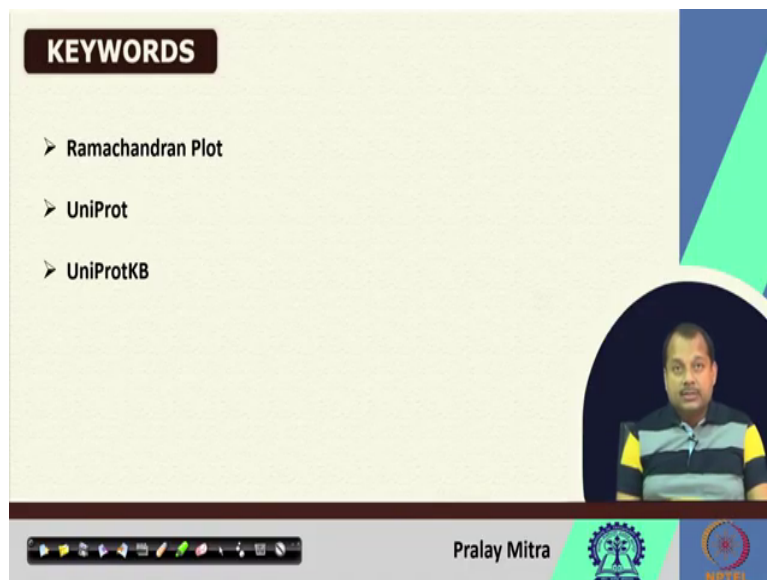


CONCEPTS COVERED

- Ramachandran Plot
- Protein Sequence Database
- FASTA file format

Pralay Mitra

The slide features a dark blue header with the title 'CONCEPTS COVERED' in white. Below the title, three bullet points are listed. A circular inset in the bottom right shows the speaker, Pralay Mitra, wearing a grey and yellow striped shirt. The bottom of the slide has a navigation bar with icons and logos for Pralay Mitra, IIT Bombay, and NPTEL.



KEYWORDS

- Ramachandran Plot
- UniProt
- UniProtKB

Pralay Mitra

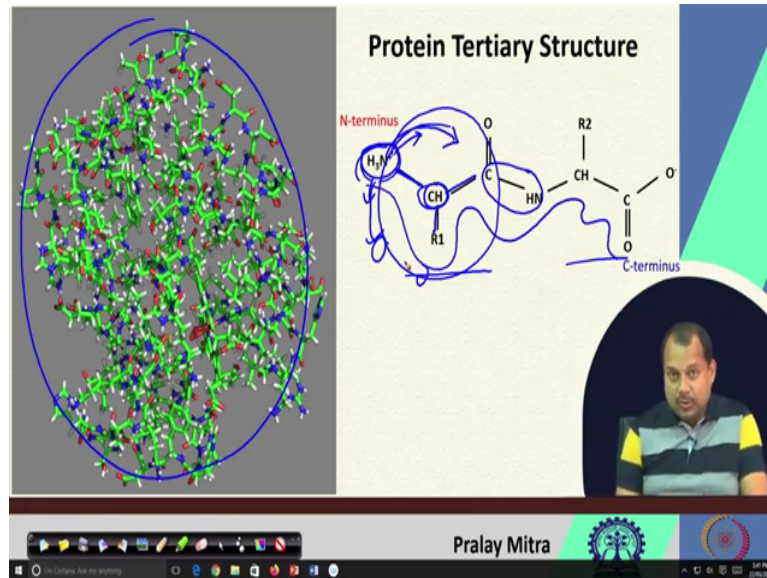
The slide features a dark blue header with the title 'KEYWORDS' in white. Below the title, three bullet points are listed. A circular inset in the bottom right shows the speaker, Pralay Mitra, wearing a grey and yellow striped shirt. The bottom of the slide has a navigation bar with icons and logos for Pralay Mitra, IIT Bombay, and NPTEL.

Today, the concepts that I am planning to cover are Ramachandran plot, protein sequence database, and FASTA file format. Thus, the keywords are Ramachandran plot, UniProt, and UniProtKB. Ramachandran plot, in the context of protein structure, is useful to recognize the quality of the protein structure.

Specifically, when we shall model protein molecules. For model construction, we use computational techniques or an algorithm or an implementation, where if you give input, as output you will get the protein structure. After this model construction, it is important to know whether the model structure is biologically relevant, or it is physically stable, etcetera.

Thus, several checkings need to be done before we send it to the experimentalists or the clinicians for their test. To do that, the first checking or one of the primary checks, which is well accepted is using this Ramachandran plot.

(Refer Slide Time: 02:01)



Let us start with the last slide on the protein tertiary structure that we have discussed. As you remember that on the left-hand side is the protein tertiary structure with one protein molecule, the sequence of the protein molecule. Now, corresponding to each amino acid, I replace it with their corresponding atoms that constitute the amino acid molecule.

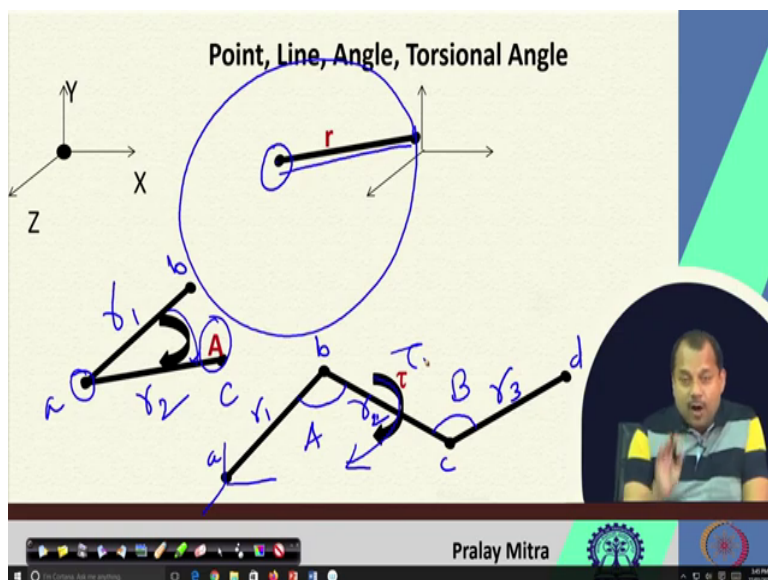
If I place those amino acids one after another and between two consecutive amino acids if I connect the amino and carboxylic group then a peptide bond is formed at this position. So, one peptide bond is formed at this particular position, OK! This is the connection between, N-terminus and C-terminus since there are only two amino acids. One and two that I have shown you, but you may have to say 100 or 150 or 200, etcetera. That is also possible.

Therefore, you just keep on adding amino acids and corresponding to each amino acid or the biomolecule if you replace them with their corresponding molecular structure, where the atoms are connected with the covalent bonds, then you will get one stable structure like this in 3-dimensional space. Okay!

In the process, you see that this is one structure - a schematic draw. And, these are the covalent bonds, so I know exactly what is the length of that bond. Also, it is the fact that given this is one position and this is the covalent bond, this particular atom can be here or

here. Centering at this position this particular atom-like nitrogen or say this NH_3 , this part can rotate anywhere like this and can accordingly be placed here or here or here or here. So, finally where? What is the correct position? Regarding this, we have to have some rules and indeed there are some rules. But, before going to that rule, let me go to some basic things.

(Refer Slide Time: 04:32)



If there is only one single point in 3-dimensional space, then you need to mention three things, X, Y, and Z coordinate. Using that one, you can mention its exact place in 3-dimensional space. If you have two points and you know that what is the distance between these two points, then what do you need to know? Along with the coordinate of one particular point, you need to mention what is the length, that is r .

Using X, Y, Z, and r , you can fix that position. It is true that if X, Y, Z are given for this then it can rotate and it can have anything and as such, there is no problem with that one because they are stable. Therefore, it does not matter whether it is on this side or on that side, or at those sides. If these two atom positions are there and only one covalent bond, what is the length of the covalent bond is also mentioned.

The moment you will have the third point, I mean three points are there then you need to incorporate one more parameter that is the angle parameter A. Given one position if you know X, Y, Z coordinate, so you know r_1 and r_2 (two covalent bonds) between my three points a, b, and c. Then between a and b the covalent bond distance is r_1 , between a and c it

is r_2 if you know the coordinate of X, Y, Z of A. OKAY! And then using the angle of A you can also fix that position.

Next, if you have four points a, b, c, and d, then for this one you may have X, Y, Z coordinate assuming that is your reference frame or if it is not then also you can have X, Y, Z coordinate system for that one. Then you have r_1 covalent bond here, r_2 covalent bond here, r_3 covalent bond here. Now, you know what is angle A, you know what is angle B.

Since it is in 3-dimensional space, you can keep a, b, c in one plane and b, c, d in another plane, and you want to rotate the plane. Because of that flipping, one more angle you need to know or fix is that dihedral angle. I can denote that one by this. Okay!

(Refer Slide Time: 07:51)

	N	CA	C
	11.751	37.846	29.016
	12.501	39.048	28.539
	13.740	38.628	27.754
	14.235	39.531	26.906
	15.552	39.410	26.282
	16.616	38.913	27.263
	16.789	39.630	28.369
	17.791	39.281	29.375
	17.598	37.844	29.863
	16.368	37.519	30.261
	16.004	36.186	30.742
	16.371	35.097	29.741

Bonds, Angles and Dihedral Angles

N-terminus

H₃N⁺

R1

R2

C-terminus

O

C

H

HN

CH

O

O

Pralay Mitra

Bonds, Angles and Dihedral Angles

N	11.751	37.846	29.016
CA	12.501	39.048	28.539
C	13.740	38.628	27.754

N	14.235	39.531	26.906
CA	15.552	39.410	26.282
C	16.616	38.913	27.263

N	16.789	39.630	28.369
CA	17.791	39.281	29.375
C	17.598	37.844	29.863

N	16.368	37.519	30.261
CA	16.004	36.186	30.742
C	16.371	35.097	29.741

Pralay Mitra

Bonds, Angles and Dihedral Angles

N	11.751	37.846	29.016
CA	12.501	39.048	28.539
C	13.740	38.628	27.754

N	14.235	39.531	26.906
CA	15.552	39.410	26.282
C	16.616	38.913	27.263

N	16.789	39.630	28.369
CA	17.791	39.281	29.375
C	17.598	37.844	29.863

N	16.368	37.519	30.261
CA	16.004	36.186	30.742
C	16.371	35.097	29.741

Pralay Mitra

Now, if you look at the protein structure on the right-hand side. Left-hand side, I shall come later. On the right-hand side, what do you have? You see that this is one, sorry, this is one amino acid AA1. This is another amino acid AA2. So, on this left-hand side whatever is there, on the right-hand side the same thing is repeating except for these terminal cases.

You know that those terminal cases will occur for only one N-terminus and another C-terminus in the entire protein sequence. But for the rest, it is repeating. What is it? N CH CO, then CO N CH CO, this way it is keep on repeating. For the time being, I am not considering this $r1$ and $r2$. As I mentioned earlier that this is a sidechain and by varying this sidechain different amino acids will be generated.

Now, if I look at this one, this path, is also called the main chain. Sorry! I am not visible here. This is called the main chain - which one? This one and this is my side chain. If I look at the main chain, then you will see that this is basically repeating, and when it is repeating then the different torsional angles or at the main chain, different angles are formed.

This will be also CH and C. Okay! So this CH and C and this CH and C is the same bond-wise and angle-wise also the name is same, the angle will be different, I agree, but the name is same. Here, the angle is between N C and this C, here N C and this C. Similarly, if you look at the variations at the torsional angles, then you will see that how many variations are there. C, CH, CO, NH, C, and, this is one.

Let me erase this one and draw again. Now, this is 1, so 1, 2, 3, 4, this is one and then say starting from here. So, this is one, OKAY! And starting from here 1, 2, 3, 4 this is one. So, these three dihedral angles or the angles between two planes will keep on iterating. Those are named omega which is around the peptide bond. You remember that this is carboxylic group, this is carboxyl group CO, this NH is amino and when they are forming a bond, so that is the peptide bond. Then across these CH and C, so, N CH C NH that is my psi and then phi that is CN CH C.

Hence, phi, psi, and omega, those three angles are keeping on changing. Now, if you have multiple amino acids, then you can consider that. OK! Here there will be another phi, I did not draw why I will come to that one. Here there will be another. I will come to that. This is i , this is i , this is i , and then it will be i plus 1 corresponding to i th amino acid i plus one amino acid, that way it will keep continuing. Hence, the situations are.

Here, the point is, why I did not include is because if I consider this one, then you see that for this phi. So, this phi is supposed to be C N C C but here I am getting N C C, the previous C is missing because it is the starting point. That's why there is an assumption, you can assume 180° or so as one and that is the same for here where I will be missing one psi value.

And that for that psi also, I will be having some assumption like 180 or so, but for the rest, I shall have the exact value of that one if I have the coordinate of every atom here and it is. We will get that one if we have the structure through X-ray diffraction or we model in 3-dimensional space, then corresponding to each atom we should have one particular coordinate X, Y, Z, that coordinate is now on the left-hand side as you can see.

Here, the atom is N, N indicates this N, then CA. CA is nothing but the Carbon with A stands for alpha. This is also called the C alpha which is at the heart of the amino acid. You remember that at the center position there was one C, so something like this. On one side is H another side is R, on the right-hand side there is carboxyl group C with double O, and on another side, there is an amino group.

(Refer Slide Time: 14:22)

Bonds, Angles and Dihedral Angles

N	11.751	37.846	29.016
CA	12.501	39.048	28.539
C	13.740	38.628	27.754

N 14.235 39.531 26.906
CA 15.552 39.410 26.282
C 16.616 38.913 27.263

N 16.789 39.630 28.369
CA 17.791 39.281 29.375
C 17.598 37.844 29.863

N 16.368 37.519 30.261
CA 16.004 36.186 30.742
C 16.371 35.097 29.741

N-terminus

C-terminus

R1

R2

Ψ

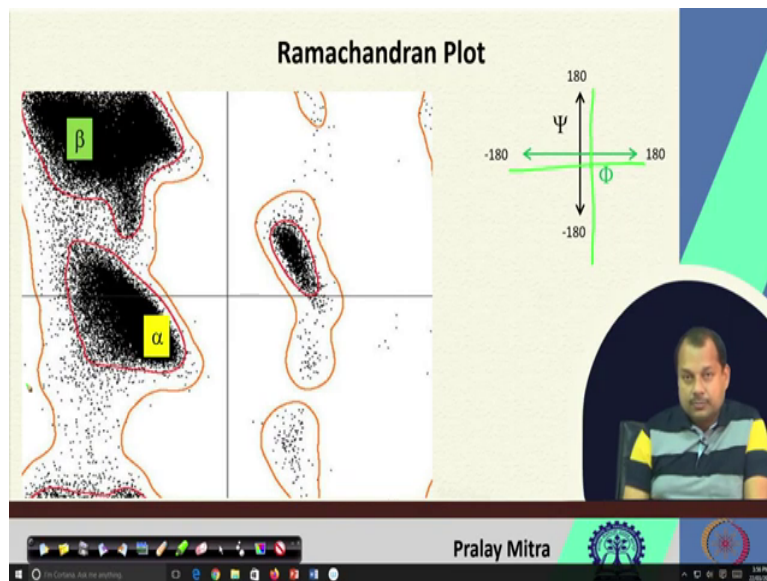
Φ

Pralay Mitra

Now, for this particular structure, here it is called the C alpha. That is why its name is CA. The question may come to your mind if there is alpha then is there beta? Gamma? Delta etcetera? Yes, those are there as part of the sidechain. If you consider this is alpha; the next carbon will be beta; the next carbon will be gamma like that way it will go on.

Right now, we are not going into that. What we are trying to demonstrate is if I consider this N here, C alpha here, then C only. The rest of the atom, I am not considering now. When we will go into a detailed discussion with a PDB structure then I will show that to you then N, sorry this is not H, N then C then C. That way if I go N CA C, N CA C, N CA C and using this one I will get the coordinate, and possibly I can calculate the phi-psi angles. Now, I am interested in phi and psi. Okay!

(Refer Slide Time: 15:35)



Now, we shall look at the work of one scientist GN Ramachandran. He mentioned that if you can calculate the phi and psi of all the amino acids in your protein structure. Definitely the terminal situations you are not including, except those terminal situations. And if you make a plot, what plot?

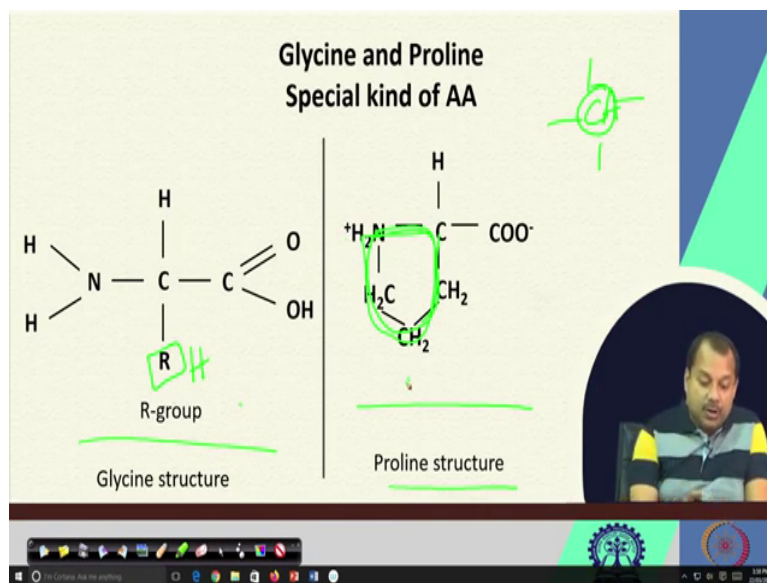
Along the X-axis, if you make phi, and along the Y-axis if you make psi since it is angle both varies from minus 180 degrees to plus 180 degrees which means covering 360 degrees. Then you will get a plot as per the name of the discoverer, the name is called the Ramachandran plot which will look like this on the left-hand side. Each dot in the scatter plot indicates one point corresponding to one amino acid with its phi and psi value.

This plot is drawn based upon the experimentally validated protein structures which are stable. Okay! Now, you see that there are dense regions, there are regions that are a little sparse and some regions are there which are completely white which means there are no points except few small dots that you can consider as the aberrations or the exceptions.

Professor Ramachandran mentioned that if the phi-psi value of all the amino acids in your protein structure is within this limit - what is the limit? The dense region he divided into two parts - you see that one is within the red region, another is orange. Red is strictly following the plot and orange is loosely following the plot. If it is within that plot means the contour or boundary which is given here. Then particular protein structure is a valid one and it is a stable one.

But if it is not then that is not a correct structure, it does not matter whether you got that structure computationally or experimentally. OK! We should remember this because when we are modeling some protein structure or getting the protein structure experimentally. Therefore, the fastest test we can perform is using this Ramachandran plot. There exist several software which can do that one. So, you need not have to worry about it – just make use of those.

(Refer Slide Time: 18:49)



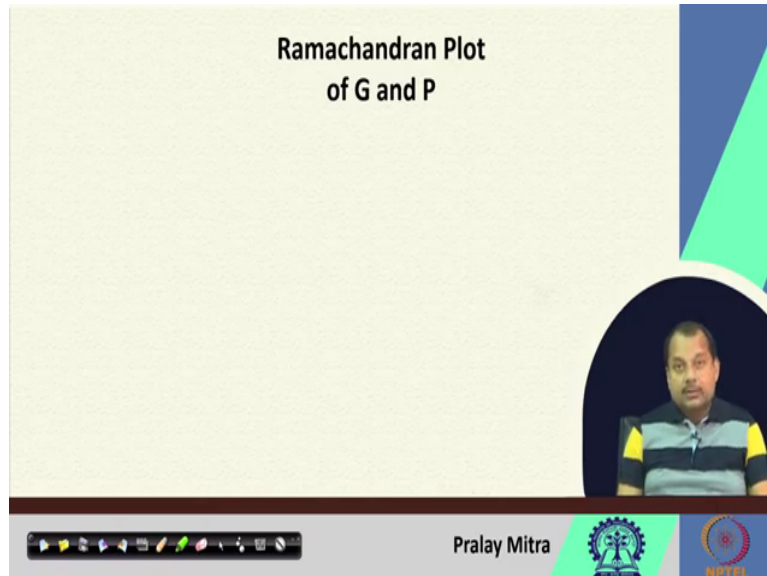
But these two guys are special. Special in the sense, if you look at their structure glycine and proline then what you will see is that for glycine the side chain (at the place of R) there is only one hydrogen. OK! As you know hydrogen is the lightest and occurs at the first position of the periodic table. On the other hand, for the proline, it is not a 4-handed structure.

By 4-handed, what I wish to say is that the generic structure that I draw is something like this. This particular C alpha which is also I am calling that CA or C alpha has kind of four hands that hold four different atoms. But for this particular proline structure, an interesting fact is that this is connected indirectly with N. Hence, there is a cycle it is forming. In one situation, I am sorry for this typo, it will be H, in one situation it is H, only H for the glycine. In another situation, here there is a binding.

You see if it is the light most only one and that is the minimum atom you can think of as R that is H. Because of that one, it is highly flexible. Isn't it? This glycine is highly flexible whereas, this proline structure has some constraints and that constraint is because they are

connected. So, freely they cannot move too much. That's why if you think about the Ramachandran plot, then these two guys may not follow the Ramachandran plot that I have shown to you.

(Refer Slide Time: 21:03)



There is an option for these two also. I am drawing that one you can think of. Since glycine is very flexible, that's why it can cover a lot of space in the phi-psi region. On the other hand, if I am thinking about proline since it is very restricted, it is strict. Hence, the phi-psi map, it is covering a small region, not much region is occupied by it. This makes the glycine and proline different from others.

When you are thinking about the Ramachandran plot after modeling your structure then you should keep these two things also in your mind that in your structure if there is glycine and proline separate treatment is required for them. You may think that non-glycine and non-proline are one data, glycine is one data, proline is one data. Separately you have to test. Otherwise, despite a correct protein structure, you may get some wrong results out of the Ramachandran plot.

(Refer Slide Time: 22:20)

Protein Sequence Database

UniProt (<https://www.uniprot.org/>)
The mission of **UniProt** is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB (<https://www.uniprot.org/uniprot/>)
UniProtKB consists of two sections:

- Reviewed (Swiss-Prot)** - Manually annotated
Records with information extracted from literature and curator-evaluated computational analysis.
- Unreviewed (TrEMBL)** - Computationally analyzed
Records that await full manual annotation.

Pralay Mitra

NPTEL

About the protein sequence database, it is a good thing that in protein most of the data which are required for us as a sequence or structured data is freely available. The primary database, which we will be used for extracting protein sequences, is known as the UniProt.

The homepage of the UniProt says the mission of the UniProt is to provide the scientific community with a comprehensive, high quality, and freely accessible resource of protein sequence and functional information. If you browse that one you will get a lot of information, but we will be taking only the sequence information at times if it is required to extract some other functional information then I will introduce that in that particular context.

Inside UniProt, UniProtKB is also introduced which consists of two components, one is called the Swiss-Prot and the other is called TrEMBL. The Swiss-Prot is reviewed and is manually curated whereas, TrEMBL is computationally analyzed and un-reviewed. If you are interested to use only the reviewed and manually curated part of the UniProt, then you should go for the Swiss-Prot. OK!

TrEMBL mostly we will not be using. Instead of TrEMBL you can use the whole UniProt itself - that is my point. By this time, if I am talking about the protein sequence, then you know what a protein structure indicates, and what a protein sequence indicates. Protein sequence means that I am writing each amino acid using English characters, A, C, etc. A for alanine, C for cysteine, D D for aspartic acid, E for glutamic acid, using those characters, collectively I am representing one protein sequence.

Based upon the component amino acids, I am writing those characters one after another. While I am writing one after another, then it is forming one protein sequence. You can also consider that as a kind of a string consisting of those characters and that is also true. That is a string where 20 alphabets are used and it is true. But is there any format, specific format that is followed for this protein sequence?

(Refer Slide Time: 25:11)

```
>>p|P68307|NU3M_BALMU NADH-ubiquinone oxidoreductase chain 3
MNLLLTLLTNTTLALLLVFIAFWLPQLNVVYAEKTSPEYECGFDPMG SARLPFSMKFFLVAI
TFLFLDLEIALLLPLPWAIQSNNLNTMLTMALFLIQLLAASLAYEW TQEGLEWAE
```

FASTA File Format

Pralay Mitra

Yes, there is one format, which is called FASTA. OK! The FASTA file format consists of two parts. The first one is for your comments or to make some identity or to give the primary key or ID. As you know that whenever you have some database then corresponding to that database there is one primary key so that uniquely you can identify one record or you can discriminate the records based upon that primary key. One primary key I already introduced to you and that is the PDB ID.

Where a PDB stands for Protein Data Bank, and ID is the identifier. Corresponding to the Protein Data Bank there is one identifier corresponding to one data. Now, for UniProt also corresponding to that UniProt, there is one particular ID that is called the UniProt ID. Along with the sequence in the FASTA file format, there is an opportunity to include any identifier if you have one.

Since I am extracting from the UniProt, then definitely it will have the UniProt ID. So that at a later point of time, if I go back, and I wish to do some more calculation or modification of my data, or if I say compare data, for all those analyses, mostly I will be using that ID. and,

that is my UniProt ID and it is given here. This is my UniProt ID. This green is not prominent very much; let me change it to some other color, blue. This UniProt ID is provided as the first line.

Along with that one, if you have any comments, you wish to write something regarding the molecule or your comment, you can write that one. But that should occur in a line where it starts with this greater than (>) symbol. This is very important. If it starts with greater than then in the FASTA file format, the general assumption is that - a line whose first column or which starts with a greater than then that line is used for the comment. It is not part of the sequence, try to understand PDB ID or UniProt ID or in your comment or any ID is not part of the sequence. Part of the sequence will start with an English character or the one-letter character of valid amino acids from the first column.

Usually, the FASTA line goes up to 60 columns, and then it goes to the next line. But this restriction is not maintained always. The reason is very simple and you can understand that. If I write in one line, then what will happen - that it will go on. So, you have to use a horizontal scroll to read, to see that sequence. What is the length or what is the sequence?

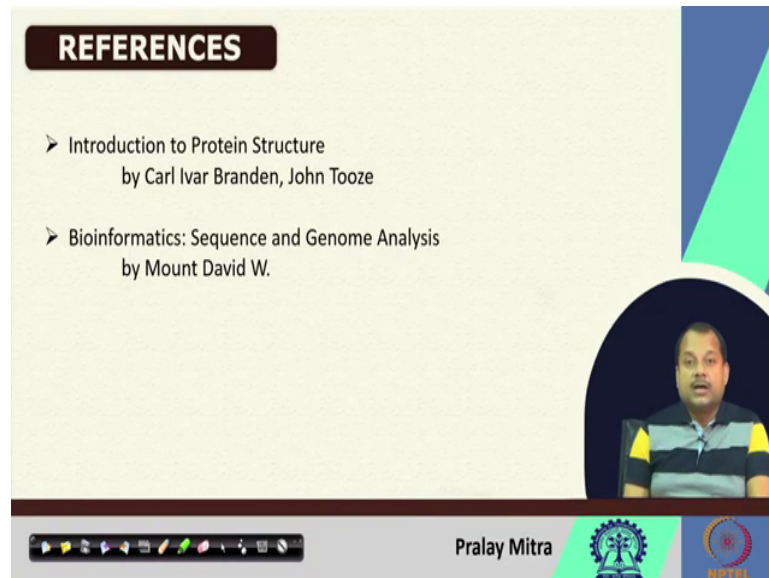
But if it is within say 60, then most of the text editor supports columns up to 60. That is why this restriction is there. But if you write in one column that is in one row, or in one line that is not a problem that is for this sequence that the sequence starts from the second line. And please note that it is not starting with greater than. If you start with this greater than symbol then during the processing universally it is accepted that it is a comment line not related to protein sequence. The protein sequence will start with an English character indicating amino acids. And if it is multi-line - no problem you just keep on concatenating.

After this, T will go here, if there is one more line, then this next line will go at this end that way it will keep on concatenating. OK! And you will get a complete protein sequence. When you are processing - up to 60 limits on the column is purely based on your visualization, but for processing, you need the complete sequence. Therefore, you have to take it as a whole sequence. This is one line; this is another line for your processing purpose.

This comment part, how do we interpret that is up to you. There is one keyword. You can store this sequence with this keyword, you can have one structure - data structure basically

where this ID will be stored and this sequence will be stored for processing. Again, it is up to you. But the format says it is the FASTA file format.

(Refer Slide Time: 30:27)



Along with the previous reference, that I have mentioned “Introduction to protein structure”, you can also look at the book on “Bioinformatics: Sequence and Genome analyses by Mount David”. Summarily, what we discussed in this lecture is the Ramachandran plot, its necessity in the context of the protein structures to know or to test the goodness of a protein structure specifically when you are modeling or engineering some protein structure - it is very relevant.

And then we also introduced to you that protein sequence. We talked about the UniProt, UniProtKB, Swiss-Prot, and TrEMBL, not in detail, but I just introduced them to you. Whenever it will be required in that particular context we will discuss it in detail. And also, I mentioned that sequence although can be taken as one single string, for the visualization purpose, chopping at column number 60 may be done. And there is a particular format also which is called a FASTA file format to store a sequence where any line which starts with a greater than symbol (>) indicates that particular line is a comment line.

That particular line consists of information apart from the amino acids. Now, you can use that part to store the ID of the sequence, to store the information regarding the protein molecule. I will show some examples also on the next class when we will discuss the PDB. And after that one, the line which is not starting with the greater than is part of the protein sequence.

It will consist of 20 English characters consists of essential amino acids. There should not be any space in between. If there are multi-lines then you can concatenate those multi-lines until you will reach either end of the file or another greater than symbol indicating that another FASTA sequence is going to follow. We can keep on concatenating that one and store that one for processing purposes. Those things we completed. In the next lecture, we will start the protein structure database. Thank you.