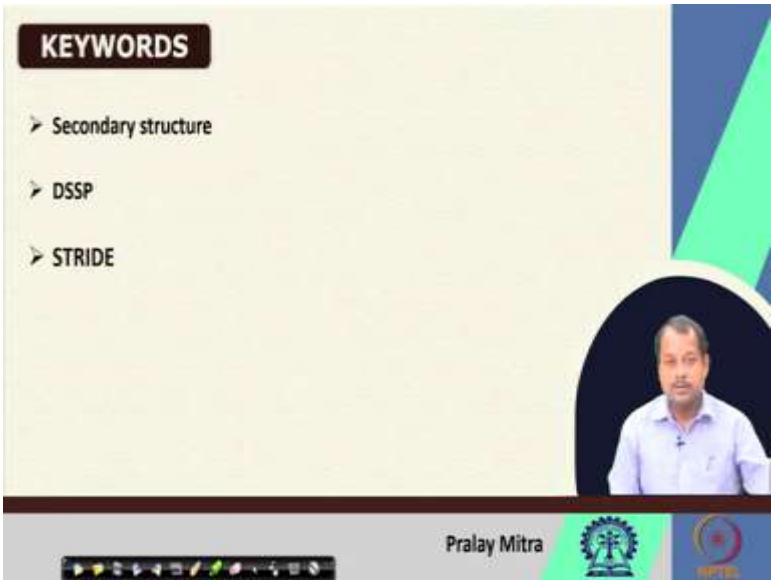
Algorithms for Protein Modelling and Engineering
Professor Pralay Mitra
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur
Lecture 52
Assigning Secondary Structure to Protein Sequence (Contd.)

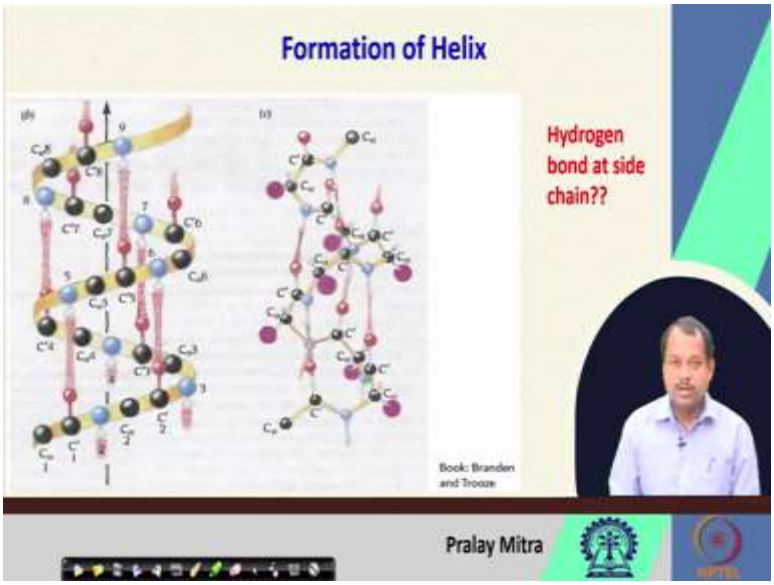
Welcome back so we are continuing on the hydrogen bonding and then we started to look at the pattern the hydrogen bond creates in the helix formation and the sheet formation.

(Refer Slide Time: 0:26)



So, that we will continue in this lecture also, the concept we are planning to cover the same, secondary structure assignment, DSSP STRIDE. And the keywords are also kept as same.

(Refer Slide Time: 0:37)

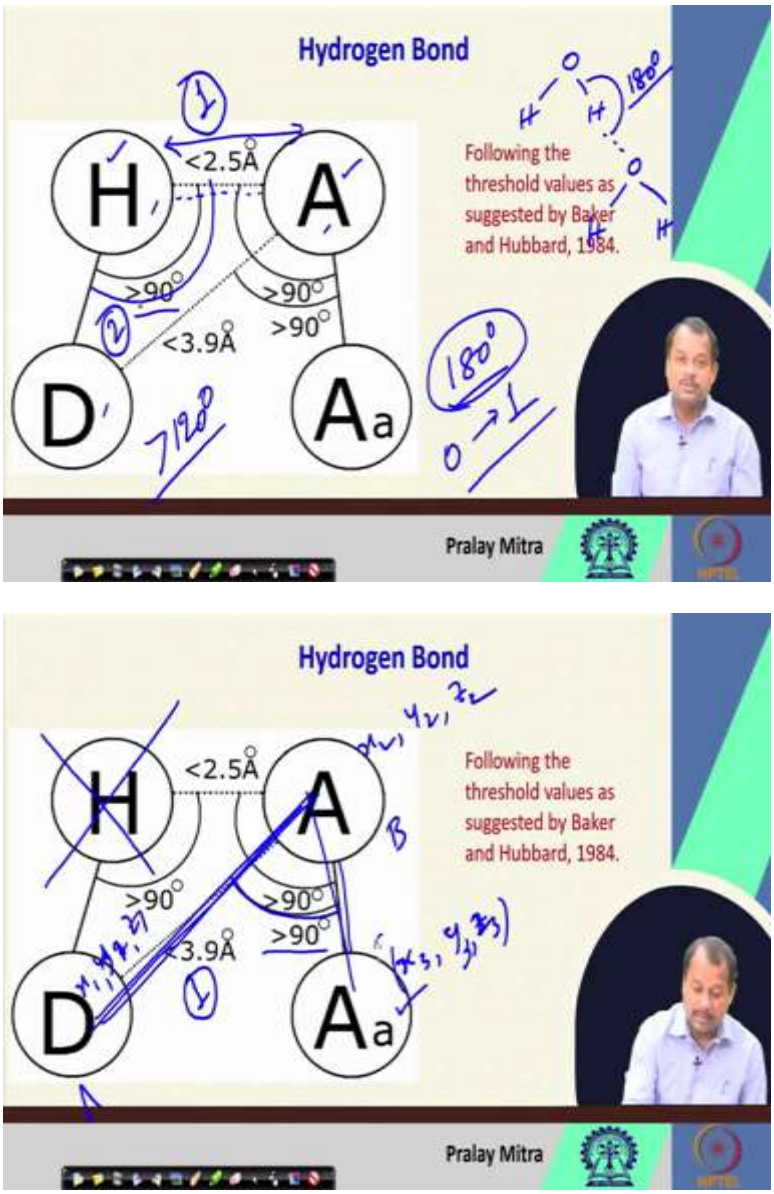


Now, we stopped our discussion on the last lecture, where we extensively discussed the hydrogen bond formation at the main chain. And when I call it is a helix, then there is a specific pattern that we discussed the pattern regarding the hydrogen bond formation. Now, there is a specific pattern and that pattern is for this hydrogen bond formation.

Now, just to complete the discussion, I also asked you to look at the hydrogen bond formation at the side chain, because when it is not only the secondary structure assignment. But say we wish to look at the total energy contributed due to this hydrogen bond formation, in protein folding, or in protein-protein docking, or interaction, or even for the protein design also that how much say energy we are losing, because of the mutation, or deletion at the protein interface, or at the protein overall.

Then definitely we need to calculate the hydrogen bond. So, that is why we will first focus on calculating the hydrogen bonds and then we will go back to our actual problem of assigning secondary structure to the protein structure, protein structure means atomic level structure is known to you.

(Refer Slide Time: 1:57)



So, let us make our life easy, so on the on the last lecture it was a bit chemistry, now it is just a geometry. So, this particular threshold, the particular threshold, I am following using the most widely concept taken by Baker and Hubbard, in 1984.

So, he has 2 different proposals, so first of all by looking at this diagram, you can see that hydrogen bond is present that is H, and apart from hydrogen bond D, indicates the donor with which H is making , covalent bond, solid line, solid straight line.

Now, on the other hand side, there is one acceptor. So, indicated by A, within the circle and there is acceptor antecedent. So, in the descendancy, so who is holding the acceptor. So, why that acceptor antecedent is required, I will go to that one, because as of now while we are discussing the hydrogen bond, we are focusing only the donor acceptor and the hydrogen atom.

But sometimes it may possible specifically, when the protein structure is solved using extra crystallography, that you may not get the information of hydrogen bond, either hydrogen sorry, hydrogen atoms, either all the hydrogen atoms are missing, or partly it is missing. So, considering all those things, it might be a good idea, that you have an alternative option, where hydrogen atom is not present, that is why we are having acceptor antecedent also.

Now, let us look at the geometry first, when the hydrogen atom is present. If the hydrogen atom is present, then the geometry says, so compute the distance between H and A. Now, our input is the atomic level coordinate of the atoms.

So, it will be very easy and simple for you to calculate the Euclidean distance between the atom A and atom H, once you will calculate, let us check whether the threshold of 2.5 angstrom holds, or not, which means in order to form a hydrogen bond, the hydrogen atom and the acceptor between which the hydrogen bottom atom hydrogen bond this dotted line actually will be formed must be within 2.5 angstrom.

So, that is our criteria number 1, our second criteria is that the angle between D, H, and A will be greater than 90 degree. So, this is our criteria number 2. Ideally it should be 180 degree as you have noted for the hydrogen bond structure, that we discussed. So, for water molecule, it was something like this. So, it was something like this.

So, here the angle is 180 degree, but we will give some flexibility and we will consider greater than 90 degree. Now, in some cases some people consider, let it be not greater than 90 degree, let it be greater than 120 degree, that is up to you and up to your implementation. But you should remember in the context of hydrogen bond, we also mentioned that the energy associated depends upon the angle also, this angle.

So, when it is 180 degree, it is the complete correct interaction, otherwise one angular term needs to be multiplied, so the cos theta and in that cos theta part, so it will vary from 0 to 1. So, that

way you will consider a fraction, thus you should remember although you are following this geometry, because of several reasons like so structures are not properly refined, or say hydrogen bond can move and when say it was crystallized during that moment I got specifically one snap.

So, because of that snap, so it was moved away, so I am not getting 180 degree, etcetera, etcetera, there may be several reasons. But you should remember from count point of view although you can report that 3 hydrogen bonds are there.

But if I say that what will be the total energy contributed, you may not multiply 3 with the individual contribution of 1 hydrogen atom, but rather you have to add them by going for that angle into your consideration. So, whether it is 180 degree, or what is the actual. Now, this is about the scenario, when hydrogen atom is present.

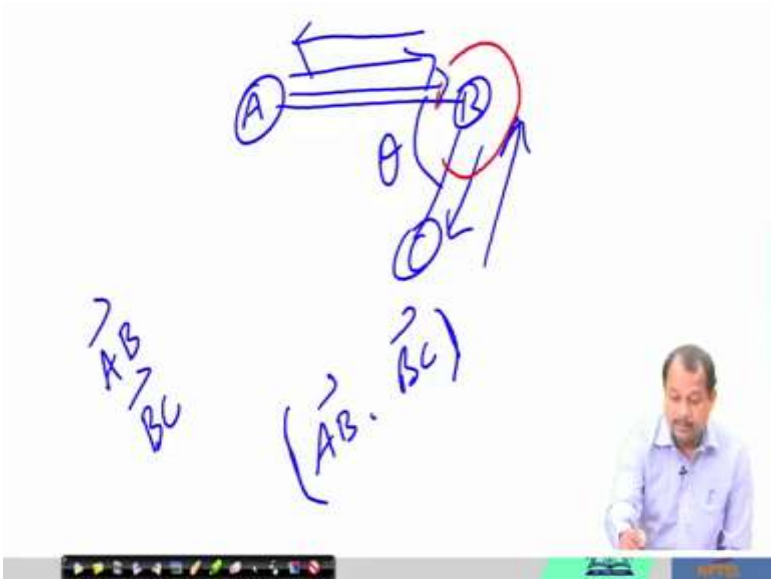
Another scenario when hydrogen atom is not present also is placed intelligently, then this atom is not present at all. So, but it is not only the distance I am calculating, but angle also I need to calculate and you know that, we discussed at the introductory classes, in order to calculate the angle also from your basic knowledge of the geometry, at least 3 atoms are required, they must be non collinear.

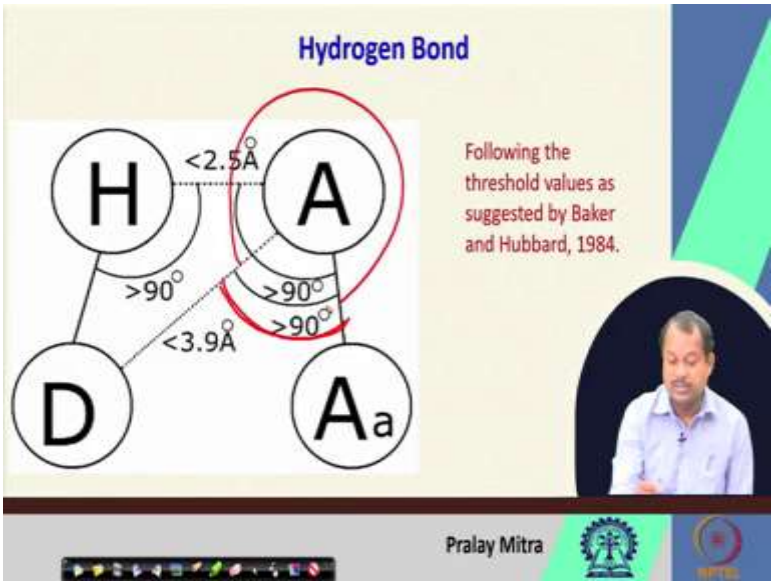
If they are collinear its fine in this context that we can also accept that one that is going to be your 180 degree angle. But anyway 3 atoms are required for us, if it is not hydrogen atom I mean hydrogen atom is absent, then I will consider accept and accept an antecedent.

And you see that, there is a covalent bond between acceptor and acceptor and antecedent, in that case the criteria, I am giving the first criteria the distance between donor and acceptor will be less than 3.9 angstrom, sometimes some people consider 4 angstrom, some people consider 3.5. So, in between 3.5 and 4, you can have your own value. Next the angle it will consider is greater than 90 degree between D A and A, Aa.

So, donor acceptor and acceptor antecedent between these 3 basically you will compute the angle and this angle must be greater than 90 degree. As I mentioned the calculating the Euclidean distance is very easy, that you can calculate directly from here and the angle calculation is also very easy, because you know 2 coordinates x 1, x sorry, x 1, y 1, z 1, and x 2, y 2, z 2, here x 3, y 3, and z 3, three angles. So, say A, B and C.

(Refer Slide Time: 8:56)



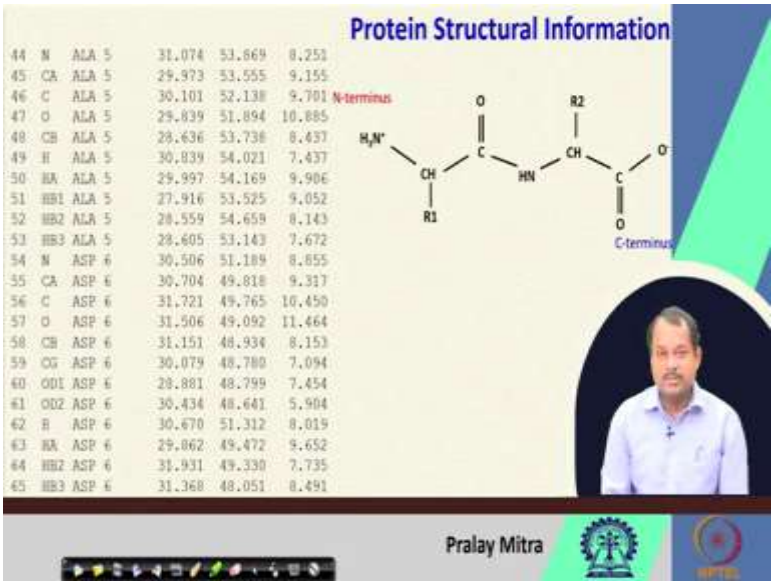


Now, what you need to do, in order to calculate the angle. So, you have 3 point A, B and C, so you are calculating this angle. So, you draw one vector, now these two vector you compute the dot product of this and if you take basically the which one, the cos inverse, basically the dot product of this one will be multiplied with the cos inverse, multiplied with the cos theta. So, that cos inverse will give you the angle.

Now, what is your direction? So, whether this one to this one, or say this one and this one, or say this one, this one, this one, based upon that one, either you will get this angle, or you will get this

angle. So, be careful about this. Although for our example in this case, it does not matter much, but definitely you should be careful for that. But in this case basically the angle must be here, you should not compute this one mistakenly and then infer.

(Refer Slide Time: 10:22)

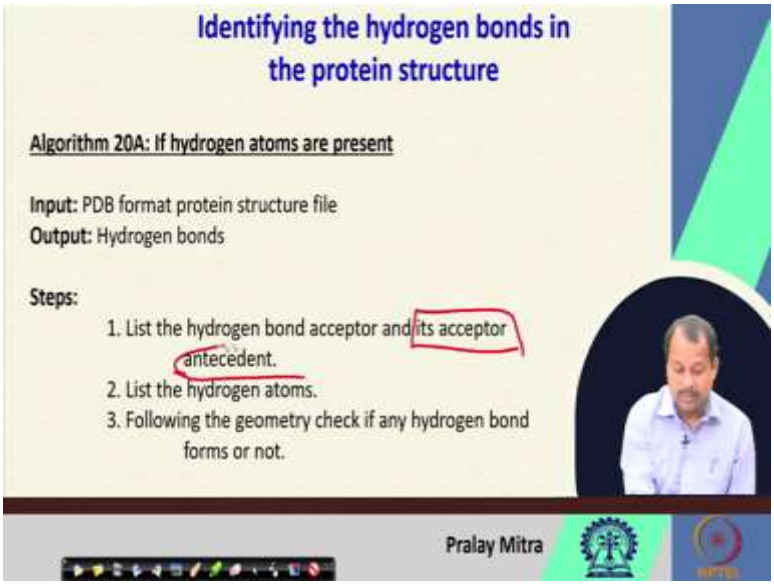


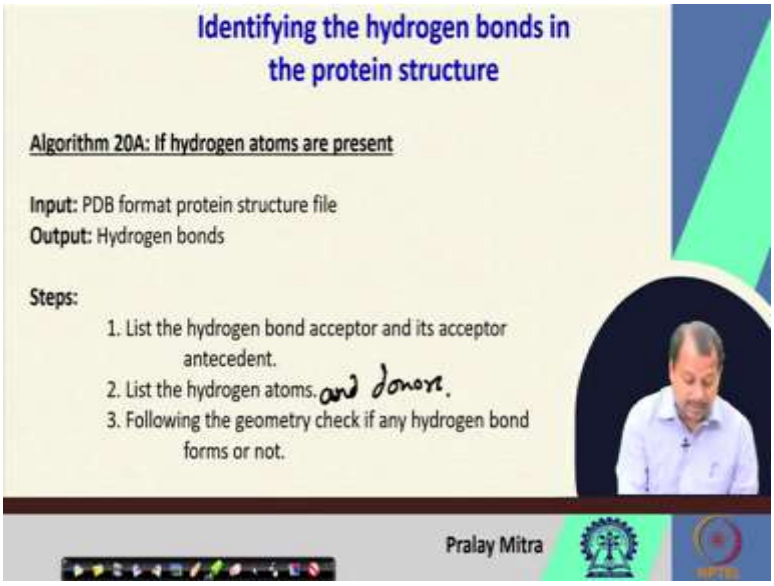
So, we are ready to compute actually the hydrogen bond, if there is any present. Now, this slide I am taking from our introductory class, here you can see on the left hand side the PDB information is given to you, the first column indicates the atom number, followed by the atom name, then residue number, then residue name, then residue number, then x, y, z coordinate, in real number.

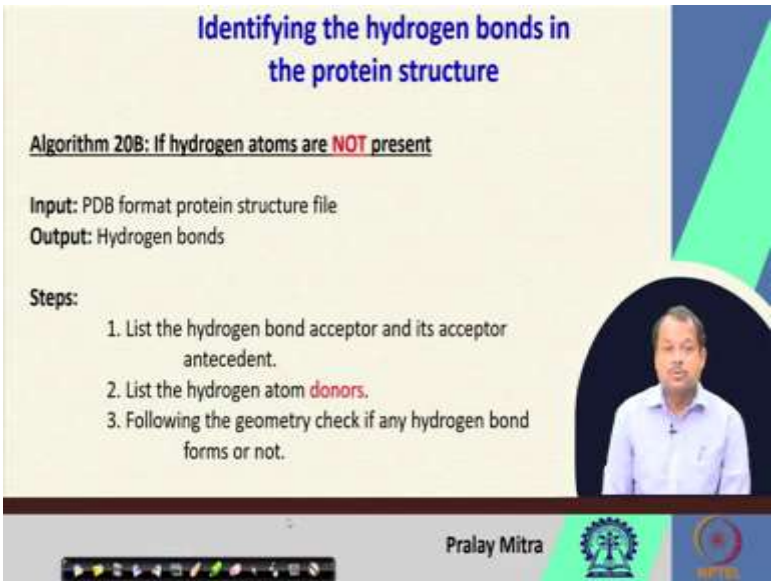
Now, incidentally in this case hydrogen atoms are present. So, we will use the first criteria, which means the angle between hydrogen atom and the donor and the acceptor actually is going to be the less than 2.5. And the angle between the donor, hydrogen and acceptor will be greater than 90 degree. But if this hydrogen atoms are absent, then also we know the angle between the acceptor and the donor is going to be less than 3.9 angstrom, or something in between 3.5 to 4.

And the angle between donor, acceptor and acceptor antecedent is going to be greater than 90 degree, better to have the angle threshold as 120 degree. But 90 degree greater than 90 degree is also fine. But do not forget the contribution towards the energy function will vary based upon that what is the angle.

(Refer Slide Time: 11:48)







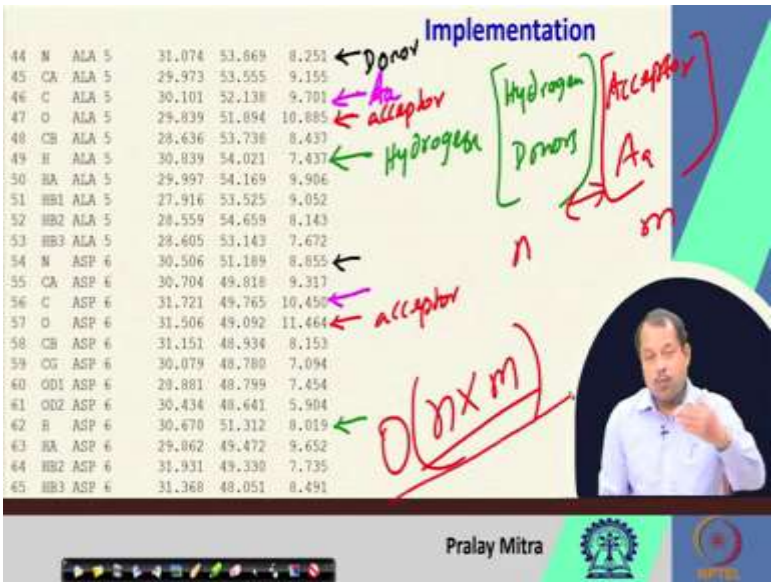
So, with this let us propose the algorithm. So, in this case we have two variations, if hydrogen atoms are present, then input is PDB format protein structure file and output is the hydrogen bonds and the steps is least the hydrogen bond acceptor and its acceptor antecedent. List the hydrogen atoms following the geometry, check if any hydrogen bond forms, or not.

So, the geometry extensively we discussed. So, it is only to calculate the Euclidean distance and the angle between 3 atoms, that we discussed either donor, in this case basically donor, angle, donor and hydrogen and the acceptor. Now, first step in first step list all the hydrogen bond acceptance acceptor antecedent. So, that way you have to keep one list for acceptor and acceptor antecedent. But if hydrogen atoms are present, then this particular information is irrelevant.

So, you may exclude that one, you may exclude that one. But for the next version, when hydrogen atoms are not present, then specifically you have to have the acceptor antecedent information. Look at the step 2 last time, it list the hydrogen atoms, but since hydrogen atoms are not present, then I have to list the hydrogen atom donors. Now, definitely in this case...

So, in this case since you need to calculate the angle. So, one situation is that you calculate the angle from hydrogen acceptor and antecedent, another is that you list the hydrogen atoms and donors. Then from there also you can calculate. But in the next case, since hydrogen atoms are not present basically. So, you have to work with the hydrogen atom donor.

(Refer Slide Time: 14:01)



Now, if I look for the implementation, it is very simple to implement, you remember at the main chain the structure we have considered based upon that one, who is going to be your say donor. So, this nitrogen, so in black color I am mentioning the donor, at the main chain hydrogen bond I am talking about, then this nitrogen nobody else here.

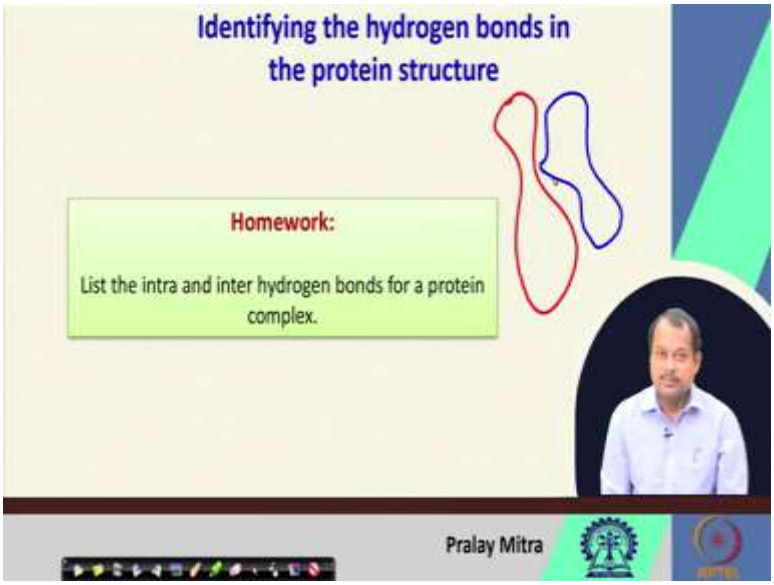
Now, who is going to be the acceptor in red color. So, this and then here I will have this, and who is going to be the acceptor antecedent. So, it is the oxygen of the carboxylic group. So, acceptor antecedent is going to be the carboxylic carbon sorry, yes carboxylic carbon. So, let me give a pink color say, then carboxylic carbon means, this is my Aa in sort I am writing acceptor antecedent. And in this case, it is going to be this Aa.

And the hero, the hero for our discussion is the hydrogen, which one, this one, in this case this one. So, you got so donor, acceptor, acceptor antecedent, or black, pink, and red will be there always, the green color hydrogen may, or may not present, depending upon whether that particular file have that one, or not. So, what you need to do? You have to make a list. So, list of donor and hydrogen atoms in one side.

So, hydrogen this is one list, another list acceptor and Aa. Now, for all such pairs belongs to each amino acid, you need to check. So, since you are comparing with the all pair, so if you have say n number of hydrogen and the donor and m number of acceptor antecedents. So, it will be m

cross n order algorithm, or steps that we have discussed on the last class sorry last slide. Now, these are the number of steps. So, you need to calculate.

(Refer Slide Time: 17:08)

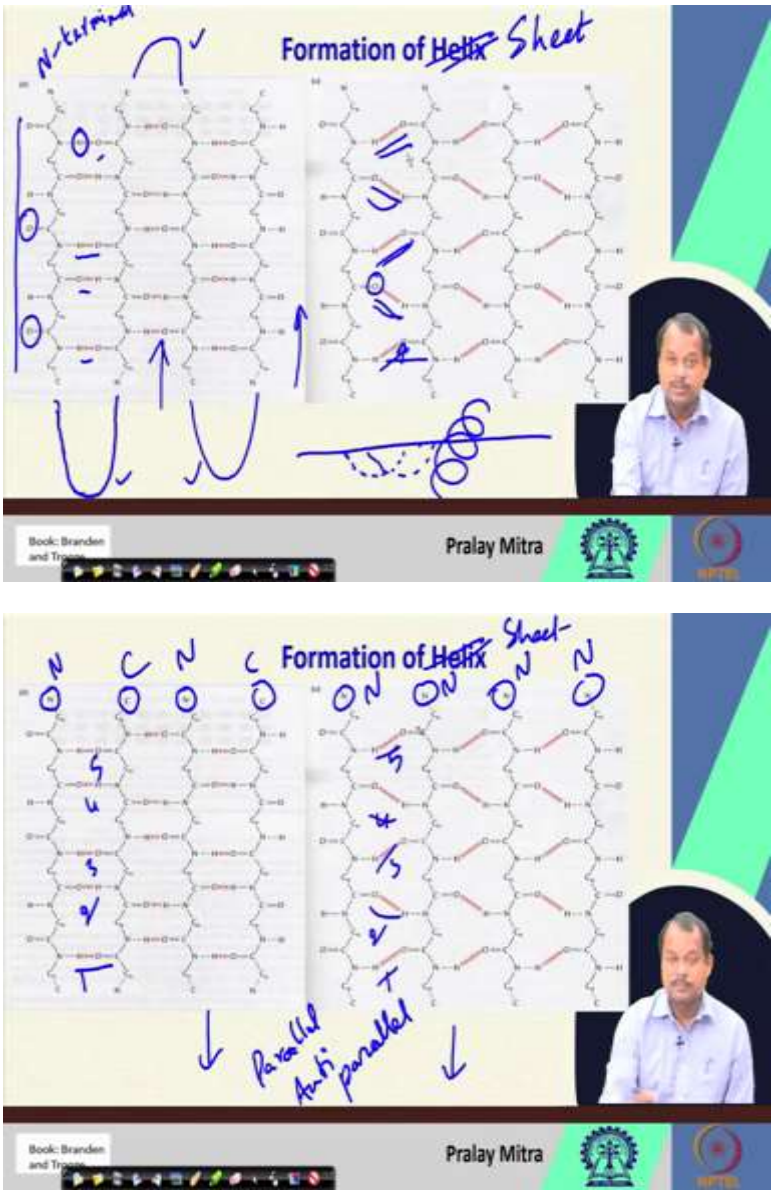


While you are doing that one, then one question I can ask to you and also I can give that to you as a homework given one protein complex, which means red and blue, I need to know how many intra, how many inter hydrogen bonds are there and list them?

Which means I need the number of hydrogen bonds, which are forming inside the red, inside the red subunit, inside the blue subunit and across the red and blue subunits, that way if you can calculate that one, then specifically this, because of this intra calculation, you can check that what is the stability contributed, because of the complex formation.

What is the stable, what is the contribution of the hydrogen bonding, or towards the stability of the protein complex formation that you can estimate? And of course on the stability of individual protein, that you can compute based upon the intra hydrogen bond inside the red and inside the blue. So, you can try to implement this one, so algorithm is very simple. So, I do not need that algorithm the same algorithm you can use.

(Refer Slide Time: 18:32)



So, again another slide, I have taken from the book of Brandon and Trooze, that is a very good book you should read that book. So, it is formation of the sorry, there is a typo, it will be sheet, it will be sheet. So, here what you can see that, again there is a pattern, if I look at on the left hand side, then you can see that, this hydrogen coming from nitrogen basically, this hydrogen coming from nitrogen is forming hydrogen bond with oxygen, it is same because I am considering on the main only the main chain.

Now, let us assume one hypothetical situation, that as if this is my N terminus, or n terminal. Now, the chain is going like this, like this and then going back, then it is going like this, coming back going like this, coming back. So, here this blue, this blue, this blue, as if my coil and this stretch, this stretch, beside this stretch, this stretch, they are basically sheet.

Now, the sheet is forming between the in the same chain. So, when they are taking some shape like this. So, it is not inside the shape, inside the chain. So, one situation was that, when there is a chain say it is the chain, the sequence, then with this there is a hydrogen bond here, or let me, let me give it a dotted way and since these dotted lines are basically within 2.5 angstrom.

So, it provides some restriction on the structure. So, the structure cannot be straight line like this, it will take some spiral shape, because of that pattern. Another way in the formation of the sheet, it says that, the chain, or the main chain will go like this and when it will come back and moving just beside this previous chain, then between these two chains, there will be hydrogen bond formation.

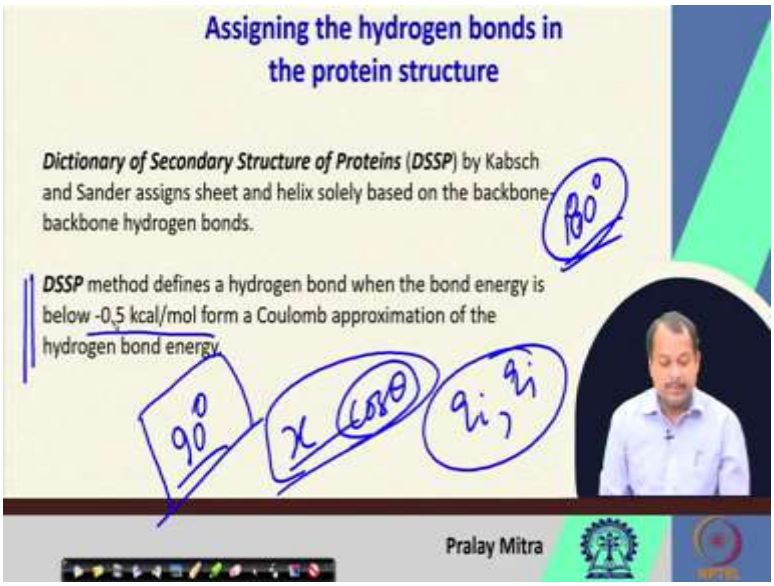
So, if between these two chains hydrogen bond formation will be there, then the hydrogen bond will be looking like this. Now, based upon the pattern, or the based upon the distance, or the offset at which they are running in parallel actually. So, there can be two situation, one you can see that the sheets are like this sorry, the hydrogen bonds are like this. Another the hydrogen bond is like this and this, this, and this.

Now, what is the basic difference you see? C, N, C, N, so on the left hand side the chain is same. Now, on the right hand side the situation is C is coming closer to N, C, C is coming closer, in this case it is not coming closer to N, C, there is some offset instead of this straight line. So, there is an offset, it is advancing, when it is advancing and at the same time, if this oxygens are here present, then you see it is forming the hydrogen bond like this way, this way, this way, this way.

Now the total number of hydrogen bonds 1, 2, 3, 4, 5, here also 1, 2, 3, 4, 5, here also 5, only thing is that whether they are parallel, or not parallel. Why not parallel? What is the difference? You follow here. So, it is N, now it is C, it is N, it is C, N, C, N, C, in this case N, N, N, N, all N. So, this is, this in this case it is running in parallel, in this case it is running in antiparallel, that way this sheet actually.

So, this sheet, actually varies in two different aspect, one is called as a parallel sheet, another antiparallel. Now, why it is parallel antiparallel, that you understand by looking at the they are running so N, C, N, C, and here it is N, N, N, N. Now, accordingly their hydrogen bonding pattern will also change. So, one is like this hydrogen pattern and another is like this, very good.

(Refer Slide Time: 23:41)



Now, let us come back to our actual problem, that is promise to discuss assigning the hydrogen bonds in the protein structure. So, there is one software tool, which is called as the DSSP, it is very popular one, the full form is the dictionary of secondary structure of proteins by Kabsch and Sander, it assigns sheet and helix solely based on the backbone-backbone hydrogen bonds.

Whatever we have discussed so far based upon that one, it performs the hydrogen bond assignment. And trust me their accuracy is very good in doing this one, although they are using only these two criteria. Now, DSSP method defines a hydrogen bond, when the bond energy is below minus 0.5 kcal per mole from a Coulomb approximation of the hydrogen bond energy.

So, please note down this second point, I mentioned during the discussion of the hydrogen bond formation, that what is the angle based upon that one, basically the total contribution due to that hydrogen bond will come. Now, if I assume that, if the angle is 180 degree, I mean the perfect, then the contribution by an hydrogen bond is say x, let us assume.

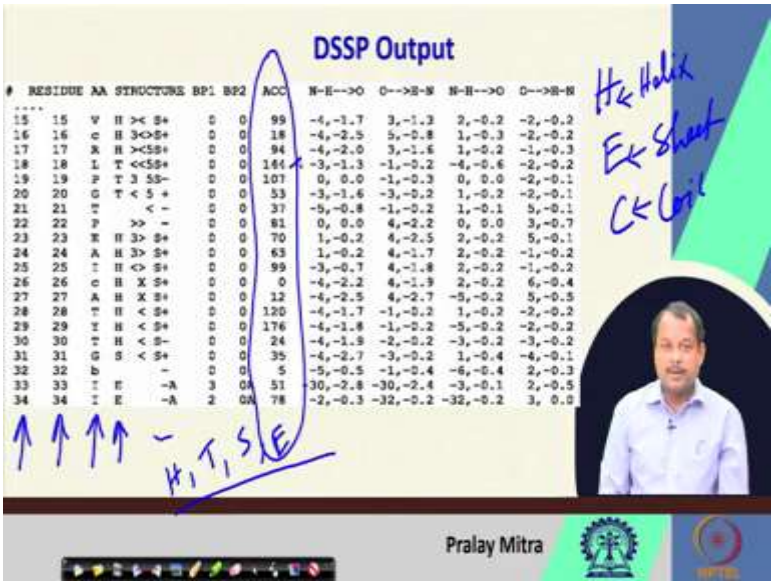
Then when it is not 180 degree, then x multiplied with some q i, q j etcetera, then cos theta it will come and accordingly, or proportionally it will reduce, I will allow them to reduce until I will reach to minus 0.5 kcal per mole. So, when they decide about this energy, then they have in your in their mind, that what is that total contribution of the hydrogen bond, I am assuming.

But I do not wish to restrict you with that particular value, what you can do that, you can calculate your own q i, q j, the partial charges plus delta and minus delta, you compute the value considering that there is a straight line then for 90 degree definitely, this if there is a cos theta. So, it is going to diminish, but apart from that one, so what is the variation and which where up to which variation you will allow, you decide it by yourself.

So, that is what I suggest you to do, that is why I am not restricting here, but the DSSP although its performance is really good, but based upon only two criteria specifically one that is the hydrogen bond formation and the main chain. But when they define a hydrogen bond, it is not only the geometric criteria, that we discussed following the Baker and (())(26:27).

But it is following the geometric criteria led by that and the fraction of the contribution based upon the angle they are considering, it is not only the angle, but also the distance between the two atoms will also be considered specifically. So, the energy contribution thresholded by that, they will consider.

(Refer Slide Time: 27:01)



Now, if I look at their, if I look at their output. So, output contains lot of information, so it starts with the index number, then following the atom number and at this point I would like to mention you should not rely completely on the say atom number, because atom number may vary.

So, atom number may not starts with 1, may starts with a negative number, atom number may be missing in between and also based upon the programming language you are using say, whether it is a pascal, or say java, or say python, or C, C plus, plus, whether you will start from 0, I mean your indexing, or non-zero that will vary.

So, better you have your own index number and that is what exactly they have done at the first column. It is not only the DSSP, most of the software tools use that one and that is why they have their own indexing. If you follow their own indexing, then you will have the unique entry at each row, otherwise that may not be guaranteed also.

Next is the atom number, it is retained for the housekeeping work and for the convenience for your use. Next is the amino acids. So, as per the protein structure file the 3 letter code has been converted to the single letter code. So, that you can see. Next is the secondary structure. Now, in this secondary structure, you can see that H, they have used say H, T S, E. So, within this small span, so they have used this.

So, actually they report 7 different secondary structure information, which includes a variety of say, turn variety of helix etcetera. But as we mentioned we will stick on to our H, E, C, which means helix sheet and coil. We will not move away from here. So, after now after that one there are some other information.

So, which is not of much interest right now, but one thing I would like to point out, they provide some accessibility information also, this accessibility is the related to the solvent accessibility, since we discussed solvent accessibility for a number of cases and also in future also I will be using that one for another problem. So, it might be good idea to note it down, but you can see that when I look at the accessibility.

So, it is providing some integer value, then the question is that how to use that one. So, better you can use some normalized value, or you can used this value directly. But if you use this value directly, then definitely your system needs to be say benchmarked according to their solvent

accessibility values. So, with this let us stop here, one more thing we will compare and discuss in the next lecture, that is the STRIDE. Thank you very much.